GRADUATE MANAGEMENT ADMISSION COUNCIL

Potential Impact of Item Parameter Drift Due to Practice and Curriculum Change on Item Calibration in Computerized Adaptive Testing

Kyung T. Han & Fanmin Guo

GMAC® Research Reports • RR-11-02 • January 1, 2011

Abstract

When test items are compromised over time for various reasons, such as security breaches, practice, and/or curriculum changes, item parameter drift (IPD) becomes a serious threat to test validity and fairness. Extensive research using simulated and real data has been conducted to investigate the impact of IPD on item parameter and proficiency estimates. Most of these simulation studies, however, have oversimplified IPD situations by assuming that the item parameters drift with all test takers. In reality, test items are exposed only to a small percentage of test takers, so IPD would occur only with those examinees. This study employed simulation studies to examine the impact of IPD on item calibration when IPD items were exposed only to a portion of test takers in a computerized adaptive testing (CAT) environment. Based on the simulation results, the short-term effect of IPD on item calibration turned out to be very limited and fairly inconsequential under the studied conditions.

Because of its distinctive advantages over conventional paper-and-pencil based tests (PBT)—more accurate score estimates with shorter test administration time—computerized adaptive testing (CAT) has become one of the most popular test approaches in the field of measurement. It is used for various purposes such as school and college admissions, statewide K–12 evaluation, and quality-of-life measures. Since the quality of CAT depends heavily on the quality of item banks, proper maintenance of item banks over time is critical (Bock, Muraki, & Pfeiffenberger, 1988). Maintaining an item bank properly, however, is a challenging task. Over time, test items in an item bank are often reused. Each reuse increases the likelihood that an item is improperly exposed and made available to test takers prior to testing day. Even if there were no direct threat to test security, changes in the interaction between a test item and a test taker still could occur over time for a variety of reasons. This change is known as *item parameter drift* (IPD) because the differential interaction between the item and test taker essentially results in item characteristics that differ from the initially calibrated item characteristics.

As noted, a security breach could be one cause of IPD, but only rarely; most CAT programs take test security issues very seriously and make every effort to protect test items from illegal public exposure (Lavelle, 2008). CAT programs also are very diligent at investigating and tracking items that test takers may have illegally disclosed (Rudner, 2010). As a result, potentially compromised items are immediately subject to disuse and discarded from the item bank. A historic event is another possible cause of IPD. For example, a national presidential election can raise the public's political awareness, and hence could increase test takers' familiarity with politically related content that might appear in a test item. Test items with content that is sensitive to historic events are relatively easy to identify, however, and usually are excluded during item pool construction. As a result, IPD due to historic events usually is inconsequential and of limited concern for CAT programs in practice.

A source of IPD that is much harder to control is the effects of a test taker's exam prep or practice and/or changes in curriculum. Practicing with related test problems is a legitimate learning technique and is often encouraged, but some test takers focus too much time and effort on test-taking strategies rather than on the skills and knowledge that the test will measure. As a result, some test items may become easier to test takers who practiced specific types of test items simply due to familiarity with the item and not necessarily because they improved their proficiency in the tested skill. This type of IPD can be a serious threat to test validity; such test takers' attempts are not uncommon especially when the test is high stake. Similarly, when students' classroom achievement is measured according to the curriculum, the test is initially developed based on the curriculum. Once test administration begins, however, lessons taught in the classroom tend to receive more weight in items that appear in the test, which helps students earn better test scores. Eventually, this could significantly change the actual class curriculum and as a result, test items containing content that was heavy weighted in the classroom could become easier for test takers to handle than they originally were.

Dealing with these latter instances of IPD is a challenge in the educational measurement field because, unlike security breaches or historic events, the effects of practice and curriculum change are not always obvious and vary widely across test takers. A number of IPD detection methods were developed (Donoghue & Isham, 1998) so that test developers could "flag" items suspicious of exhibiting IPD and exclude those items from the test administration. Most IPD detection methods were developed mainly for PBT, however, and are often ineffective or less powerful in CAT settings. For example, because of the item selection algorithm for CAT, item response data for each item almost always derive from a homogenous test taker group in terms of proficiency. When item responses only come from a homogenous group, items may not be recalibrated appropriately for the IPD detection procedures. In addition, the item response matrix usually is extremely sparse and the number of responses for each item is sometimes too small to conduct statistical analyses to flag IPD items. Therefore, the most effective way to handle IPD due to practice and curriculum change in CAT is to address it proactively before it becomes consequential. Thus, it is critical to understand when IPD first becomes consequential in CAT.

Wells, Subkoviak, and Serlin (2002) found that 20 percent of IPD items included in the test could have a significant impact on the score estimates. Han and Wells (2007) further investigated the impact of IPD items on the test score equating procedure and concluded that the test equating result could deteriorate significantly even with 10 percent of IPD items in the linking item set. Wollack, Sung, and Kang (2006) examined the compounding effect of IPD on the test score scale over time and found a large effect of IPD on the test score estimates depending on choice of linking method. The results of the aforementioned studies were based on nonadaptive tests, however, which limits the implications of those studies for CAT programs.

Unlike most nonadaptive tests using item response theory (IRT) models, item parameters of the operational (i.e., scoring) items in the item bank are fixed and are typically not recalibrated in CAT administrations. The item parameter values stored in the item bank are assumed to be accurate over time. The presence of IPD, however, could influence a student's score. Guo (2009) examined the effect of the compromised items on test scores in the case of the test security breach. Under his studied conditions, Guo found that the test scores of about 95 percent of test takers were unchanged even with test taker's prior knowledge of five test items. Guo's study, however, was specifically designed to evaluate the impact of items compromised by security breaches, and the possible impact of IPD in CAT due to practice and curriculum change remained unanswered.

The potential impact of IPD on CAT administration is twofold. First, IPD may directly influence score estimation for individuals. When students' score estimates are influenced by the IPD, it can, in turn, cause serious issues not only at the individual student level but at the CAT program level because score estimates are also used to calibrate the new pretest items. In other words, the IPD on the operational items might eventually pass on to the pretest items through the score estimates that were affected by IPD.

As a result, the pretest items that are calibrated based on score estimates influenced by the IPD may not be on the same scale as other preexisting items in the item pool. This study aimed to investigate such a possible impact of IPD on the item calibration of pretest items.

Method

Data

For the study, 1,000 quantitative items were drawn from the operational item bank of the Graduate Management Admission Test®, which is a CAT program for applicants seeking admission to postgraduate management education programs. From the 1,000 items, 100 items were randomly selected to form a pretest item set; the remaining 900 items were used as operational items in the item pool. The average item parameter values of the 1,000 items were 0.84, 0.55, and 0.20 for *a*-, *b*-, and *c*- parameters, respectively.

The proficiency levels (θ) for 50,000 test takers were sampled from $N(0.5, 1)$ to provide an approximate match of the *b*-parameter distribution using SimulCAT (Han, 2010).

IPD Modeling

In many previous studies (mainly in PBT settings), IPD was modeled in a way that changed the item difficulty parameter (*b*) and/or the discriminating parameter (*a*) from its original form, and applied the drifted item parameters to all test takers for each test occasion (Donoghue & Isham, 1998; Wells, Subkoviak, & Serlin, 2002; Wollack, Sung, & Kang, 2006; Han & Wells, 2007). Such an IPD model might plausibly simulate IPD due to, for example, historic events which equally affected all test takers. To model IPD that is due mainly to practice and curriculum change, however, it is important to reflect the reality in which the effects of practice and curriculum change differ across many test takers. Therefore, this study simulated IPD to exhibit only to a partial group of test takers (10%, 20%, 30%, 40%, and 50%) while there was no IPD for the remaining test takers. A zero IPD condition also was simulated to serve as a baseline.

Among the 900 operational items in the item pool, 20 percent (180 items) were randomly selected to serve as IPD items. The *b*-parameter values for those IPD items differed by –0.50 from the original value since the practice and curriculum change usually made particular items easier to test takers in practice. The –0.50 magnitude of IPD was selected to yield results that were comparable to previous studies, many of which simulated similar conditions (Donoghue & Isham, 1998; Wells, Subkoviak, & Serlin, 2002; Wollack, Sung, & Kang, 2006; Han & Wells, 2007). In fact, the IPD by ±0.50 with *b*-parameter at each individual item level is very important to study because it is considered an acceptable range of IPD considering that the standard error of estimation for *b*-parameter usually ranges between 0.30 and 0.50 even without IPD. An item with IPD of a magnitude larger than 0.50 may be detected easily by various IPD detection methods (Donoghue & Isham, 1998) and excluded from operational use, and so is usually inconsequential in practice. Items with IPD of 0.50 or less, however, often go undetected and end up being used in test operations. Conceivably, these items with minor IPD are the ones that might have a notable consequential impact on test programs, and thus are the ones this study attempted to simulate.

CAT Simulation

SimulCAT (Han, 2010a), a computer simulation software for CAT administrations, was used for this study. The gradual maximum information ratio (GMIR) approach (Han, 2009) was used as the item selection criterion, which looked for an item x maximizing

$$I_x[\hat{\theta}_{m-1}]\frac{[M-m(1-I_x[\theta^*])]}{I_x[\theta^*]M}, \quad (1)$$

where was the Fisher item information at the θ estimate after $m-1$ item administrations, θ^* was a θ point where the Fisher information for item x peaked, and M was the test length, which was 30 in this study. To control item exposure, the relative exposure limit was set to 0.20, by which no more than 20 percent of test takers could see the same item during the entire test administration. The fade-away method (FAM), in which the GMIR criterion value was inversely weighted by the ratio between the updated item usage and the relative item exposure

limit, was also applied to improve item pool utilization (Han, 2009). The combined use of the GMIR and FAM often results in score estimates as accurate as those found with the maximized Fisher information method and with few items excessively used (Han, 2010b).

For θ estimation, the maximum likelihood estimation (MLE) was used whenever possible. When MLE estimation was not possible, due to, for example, all correct or all incorrect responses, the expected a priori (EAP) method was used to estimate θ. The θ estimates were truncated to range between –3 and 3.

When an IPD item was administered, the CAT system used the original item parameter values for item selection since the CAT system does not recognize a change in item parameter in the real world. The drifted item parameter values were used only for test takers' response simulation.

In addition to the 30 operational items, each test taker was administered 10 pretest items that were randomly selected from a total of 100 pretest items.

Item Calibration

The 100 pretest items were calibrated using PARAM-3PL (Rudner, 2005), computer software designed for three-parameter logistic (3PL) model estimation that uses marginal maximum likelihood estimation (MMLE) with the θ estimate provided for each test taker from operational CAT administrations.

Once the 100 pretest items were calibrated, they were added to the 900 existing operational items to form a new operational item pool with 1,000 items for later CAT administrations.

Evaluation of IPD Impact

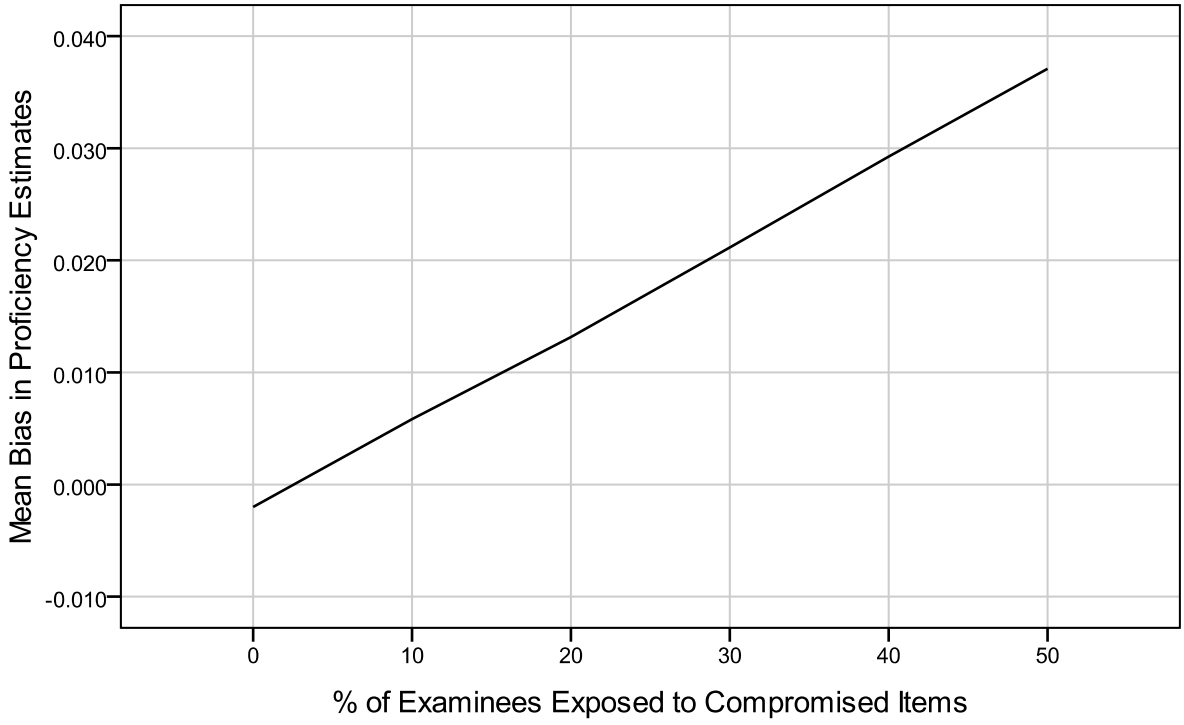
To evaluate the impact of IPD on CAT item calibration, its effect on score estimates was first investigated. The bias and mean absolute error (MAE) for $\hat{\theta}$ were computed. After the pretest items were checked, matched-pair t tests were performed on each of the a-, b-, and c-parameter estimates comparing them to the zero IPD condition. In addition, the mean absolute difference (MAD) between each studied condition and the true item response function was also computed. The MAD was measured at –2, –1, 0, 1, and 2 on the θ scale.

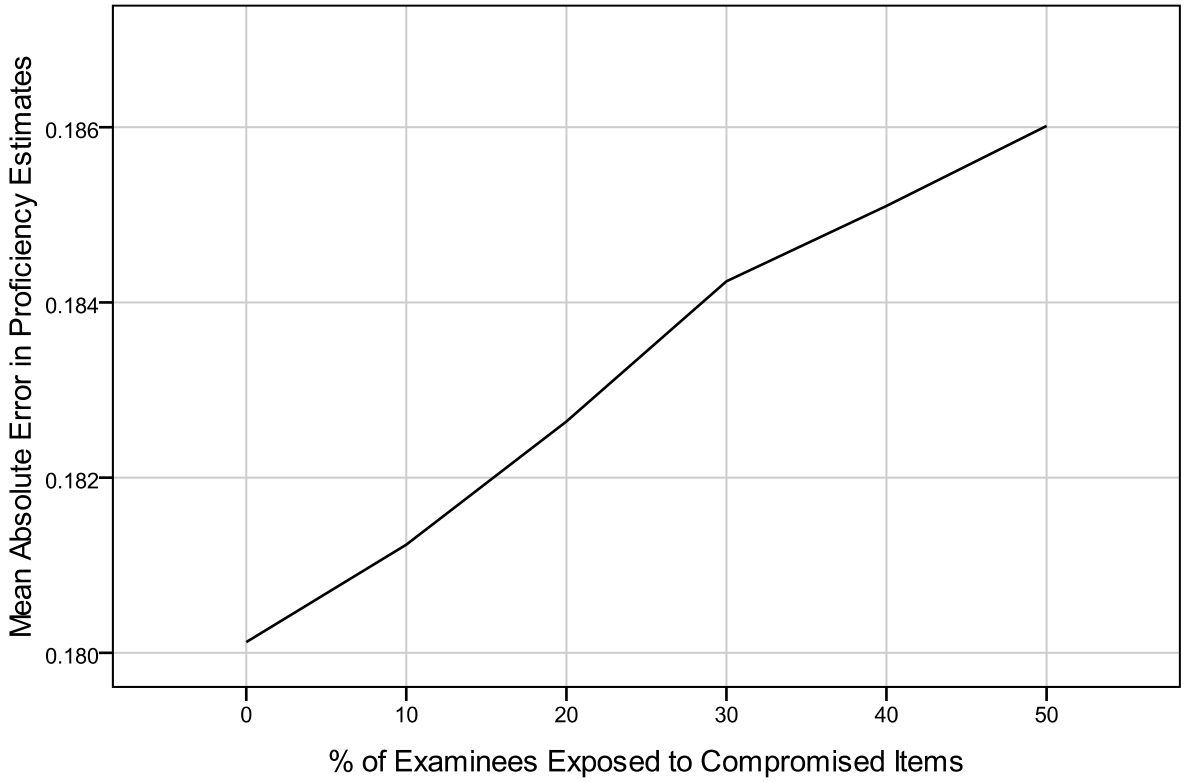
To evaluate the consequential impact of IPD, another round of CAT was administered with the same 50,000 test takers. For the second round CAT, both the 900 existing operational items and the 100 newly calibrated items were used as operational items. The score estimates from the second round of CAT were compared across the studied conditions.

Results

The mean bias and absolute error in proficiency estimates for the first round of CAT are shown in Figure 1. Since the IPD directly influenced the proficiency estimates, the mean bias in proficiency estimate increased linearly as the percentage of test takers with the IPD effect increased. When no test takers were affected by IPD, the mean bias was slightly negative but very close to zero. When 50 percent of test takers were influenced by IPD (20% of IPD items in the item pool), the mean bias increased to approximately 0.038. For MAD, the more test takers with IPD effects, the larger the MAD. Considering the standard error of proficiency estimation in practice, which is usually about 0.3, IPD had a minimal effect on scores. In the worst-case scenario in this study (50% of test takers with IPD effects), the change in MAD due to the IPD was less than 0.006.

Figure 1. Mean Bias and Absolute Error in the Proficiency Estimates for First Round of CAT





While minor but significant bias in proficiency estimates due to the IPD was observed, the critical question was the impact of IPD on the item parameter estimation for the new pretest items being calibrated. The *a-, b-,* and *c-* parameter estimates for the 100 new items are reported in Figure 2. For *c*-parameter estimates, there was no obvious tendency found in bias due to the IPD. *b*-parameter estimates tended to be inflated as the percentage of test takers increased, but this tendency was not necessarily consistent across the studied conditions. In fact, as shown in Table 1, results from the matched-pair *t* tests indicated there was no significant difference among the IPD conditions. The *a*-parameter tended to be underestimated as the percentage of test takers with IPD effect increased, but this change was minimal. According to the significance test, the bias in *a*-parameter estimates differed significantly only when 50 percent of test takers exhibited the IPD. Overall, the new item parameter estimates were not significantly influenced by IPD under the studied conditions unless IPD affected 50 percent of test takers. Examining *a*-, *b*-, and *c*-parameters separately, however, often misses the true impact of IPD on each item (Han, Wells, & Hambleton, 2009), so it is important to compare the calibrated items in terms of item response function (IRF). The MAD in IRF between the true and estimated parameters across different IPD conditions was plotted in Figure 3. Unlike the results from the individual item parameter cases, the MAD in IRF monotonically increased as the percentage of test takers with the IPD effect increased. The *t* test results (Table 1) also showed that the change in MAD in IRF due to the IPD was significant when IPD affected 30 percent or more of test takers. To examine whether this IPD influence had a consequential impact on CAT item calibration, the second round of CAT was administered with the item pool including the newly (mis)calibrated 100 items.

Figure 2. Bias in Item Parameter Estimates for Pretest Items After First Round of CAT

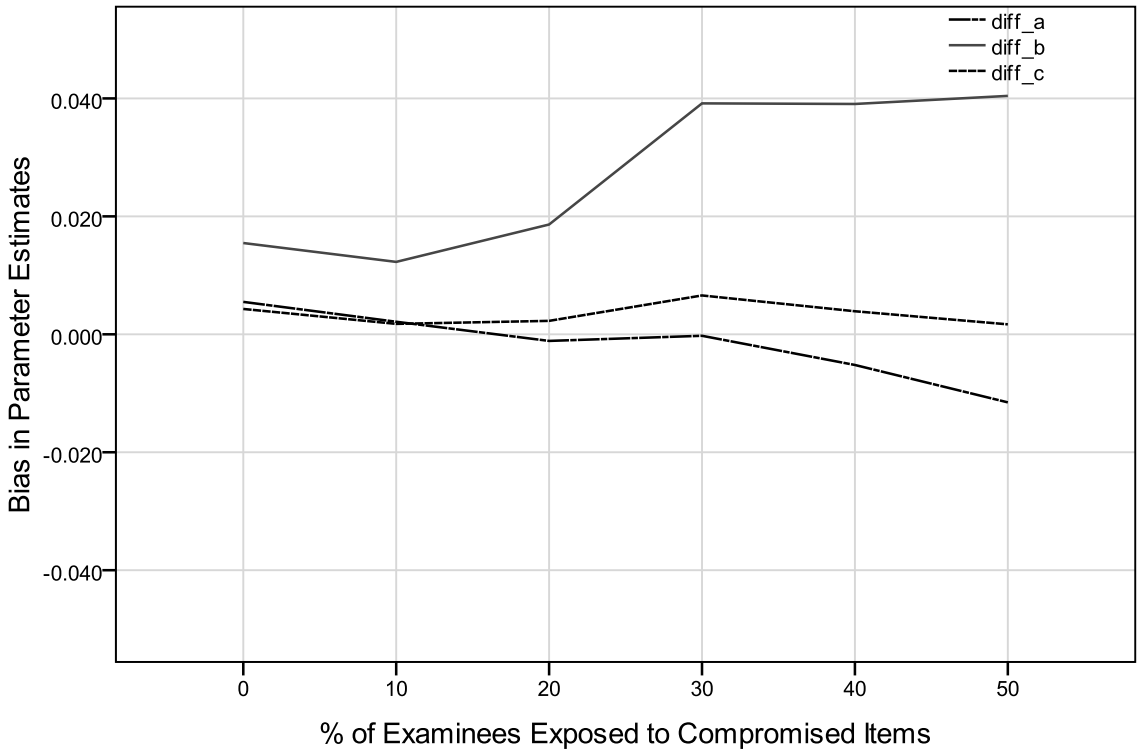


Figure 3. Mean Absolute Difference in Item Response Function Between the True and Estimated for Pretest Items After First Round of CAT

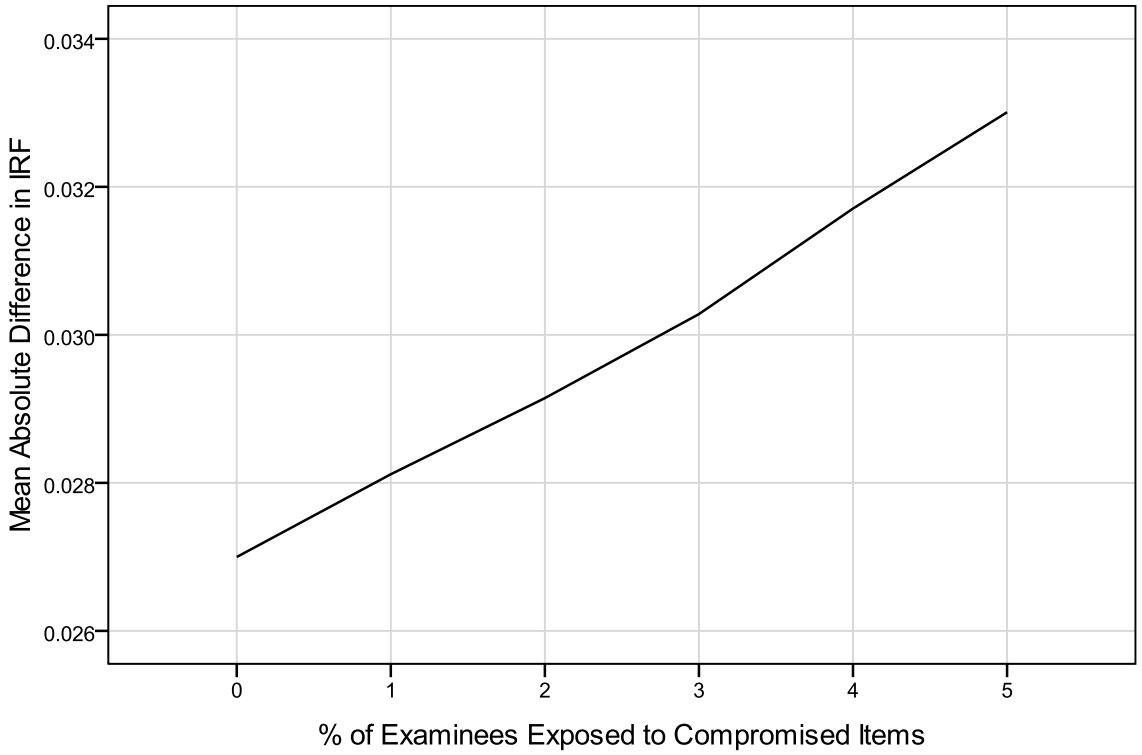


Table 1. Significance Test of Impact of IPD on Item Parameter Estimates[1]

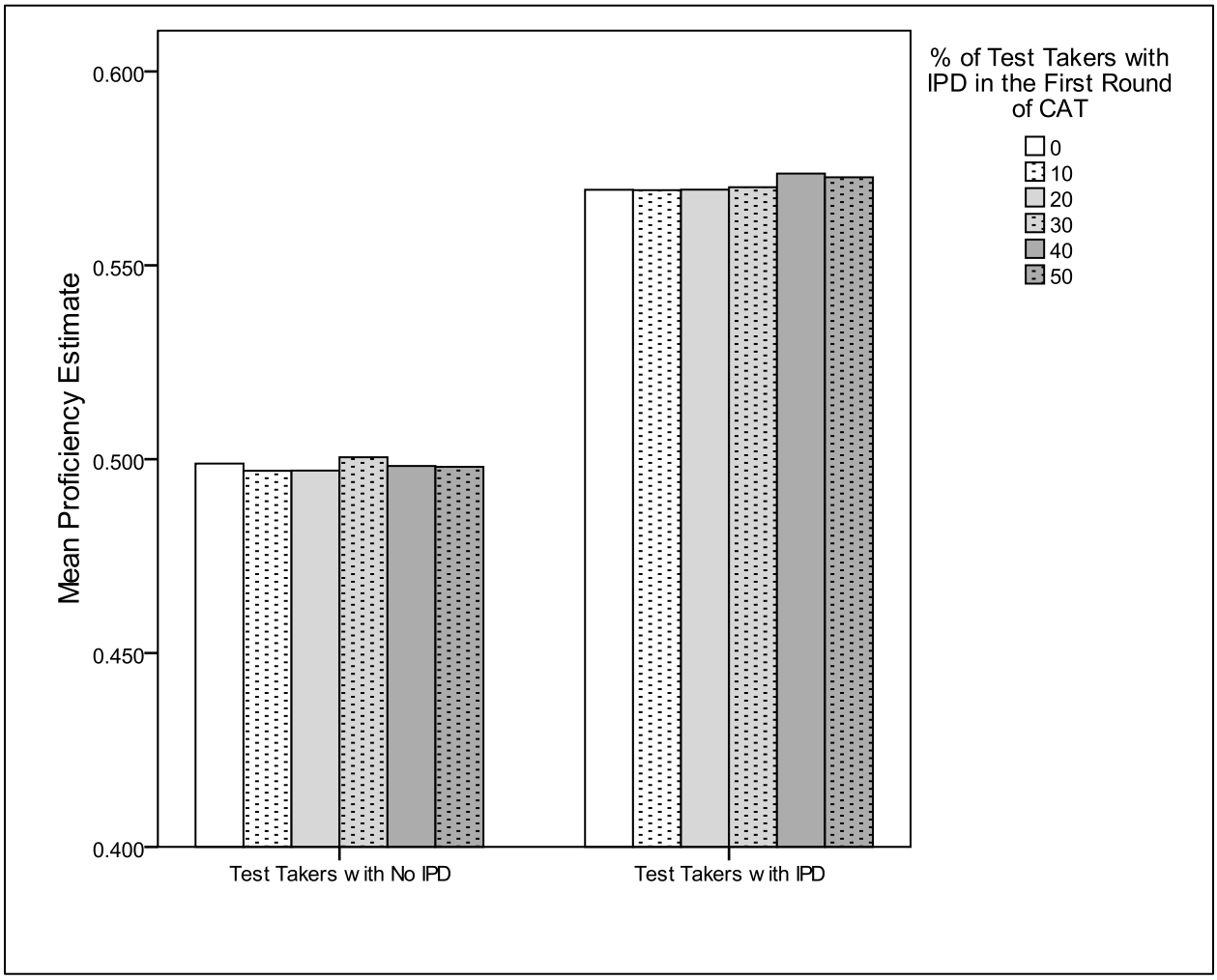
Item Parameter	% of Test Takers with IPD	Mean	Std. Deviation	*t*	*df*	*Sig.* (2-tailed)
a	10	.003	.033	1.030	99.000	.305
	20	.007	.045	1.482	99.000	.142
	30	.006	.059	.974	99.000	.332
	40	.011	.070	1.526	99.000	.130
	50	.017	.078	2.186	99.000	.031
b	10	.003	.078	.413	99.000	.681
	20	-.003	.137	-.229	99.000	.819
	30	-.024	.165	-1.439	99.000	.153
	40	-.024	.177	-1.336	99.000	.185
	50	-.025	.188	-1.331	99.000	.186
c	10	.003	.024	1.038	99.000	.302
	20	.002	.045	.447	99.000	.656
	30	-.002	.057	-.400	99.000	.690
	40	.000	.061	.064	99.000	.949
	50	.003	.063	.413	99.000	.680

Table 1. Significance Test of Impact of IPD on Item Parameter Estimates[1]						
MAD(IRF)	10	-.001	.008	-1.434	99.000	.155
	20	-.002	.012	-1.825	99.000	.071
	30	-.003	.014	-2.359	99.000	.020
	40	-.005	.017	-2.759	99.000	.007
	50	-.006	.019	-3.104	99.000	.002

[1] All significance tests were compared to the zero IPD condition at alpha of 0.05.

The second round of CAT was administered with the same 50,000 test takers under two different IPD conditions (0% or 100% test takers with the IPD effect). When no test takers exhibited the IPD effect in the second round of CAT, the mean proficiency estimates across different IPD conditions for the first round CAT showed no significant differences and were very close to the true mean score (0.50). When test takers in the second round of CAT exhibited the IPD effects again, their scores were overestimated by about 0.17, on average (Figure 4), but there were, again, no significant differences among the different IPD conditions in the first round CAT.

Figure 4. Mean Proficiency Estimates for Second Round of CAT



Discussion

The simulation results showed that IPD had a significant impact on item calibration under some of the studied conditions, but its effect was not consequentially meaningful. In other words, even if 20 percent of items in the item pool showed moderate drift (δ = –0.50) with 50 percent of test takers, its effect on item parameter estimation for the pretest items would be so small that the second CAT administration that included the newly calibrated items would not be influenced significantly in terms of score estimation. This is encouraging news for CAT programs because the IPD due to test practice and/or curriculum change, where IPD effects vary across test takers, is often hard to handle in practice.

Although the short-term impact of IPD on CAT administrations due to practice and curriculum change may be limited according to this study, the effect of such IPD can be cumulative and can become consequential at some point over the long term. This is especially true for IPD resulting from practice and curriculum change, which one could reasonably assume would influence increasingly more test takers over time. Moreover, unlike the studied condition in the second round of CAT, where the newly calibrated items (100 items) represented only 10 percent of the total item pool (1,000 items), the impact of IPD on the item calibration may rapidly become consequential as more pretest items are added to the item pool as operational items. Therefore, it is suggested that follow-up studies should look at the long-term impact of IPD on CAT programs due to practice and curriculum change.

References

Bock, R., Muraki., E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations, *Applied Measurement in Education, 17*(3), 265–300.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*(1), 33–51.

Guo, F. (2009, February). *Quantifying impact of compromised items in CAT*. Paper presented at the 2009 National Council on Measurement in Education Meeting, San Diego, CA.

Han, K. T. (2009). *A gradual maximum information ratio approach to item selection in computerized adaptive testing.* Research Reports 09–07, McLean, VA: Graduate Management Admission Council.

Han, K. T. (2010a). SimulCAT: Simulation software for computerized adaptive testing [computer program]. Retrieved March 20, 2010, from http://www.hantest.net/

Han, K. T. (2010b, May). *Comparison of non-Fisher-information item selection criteria in fixed-length CAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Han, K. T., & Wells, C. S. (2007, April). *Impact of differential item functioning (DIF) on test equating and proficiency estimates.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Han, K. T., Wells, C. S., & Hambleton, R. K. (2009, July). *Impact of item parameter drift on pseudo-guessing parameter estimates and test equating.* Paper presented at the annual meeting of the Psychometric Society, Cambridge, UK.

Lavelle, L. (2008, June 23). Shutting down a GMAT cheat sheet. *BusinessWeek*. Available online: http://www.businessweek.com/bschools/content/jun2008/bs20080623_153722.htm

Rudner, L. M. (2005). *PARAM-3PL Calibration Software for the 3 Parameter Logistic IRT Model (freeware).* Available: http://edres.org/irt/param/

Rudner, L. M. (2010). *ItemFind Software: Release 3.19*, Reston, VA: Graduate Management Admission Council.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77–87.

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.