Journal of Computerized Adaptive Testing

Volume 4 Number 1

August 2016

Effect of Imprecise Parameter Estimation on Ability Estimation in a Multistage Test in an Automatic Item Generation Context

Kimberly F. Colvin, Lisa A. Keller, and Frederic Robin

DOI 10.7333/1608-040101

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 4, Number 1, August 2016
DOI 10.7333/1608-040101
ISSN 2165-6592

Effect of Imprecise Parameter Estimation on Ability Estimation in a Multistage Test in an Automatic Item Generation Context

Kimberly F. Colvin
University at Albany, SUNY

Lisa A. Keller
University of Massachusetts, Amherst

Frederic Robin
Educational Testing Service

In adaptive testing, the availability of large banks of accurately calibrated items is critical for estimating examinees' ability and for effectively routing them through the test. The development and calibration of the item banks is thus a necessary step before adaptive testing can be implemented. Recent advances in automated item generation, in particular item cloning, might improve upon the traditional item bank development process. One approach to automated item cloning is to construct an item bank comprised of a small number of calibrated item parents. When the test is delivered, the items are then generated on the fly from the parent item that has the desired measurement parameters. The cloned item parameters, however, are inherited with some amount of error. This simulation study evaluated the accuracy with which examinees' abilities can be estimated when items cloned from a parent are used in multistage tests (MSTs) and the psychometric properties of the clones are assumed to be the same as those of the parent item. The behavior of the clones' item statistics in this study was modeled based on the results of Sinharay and Johnson's (2008) investigation into item clones that were administered in an experimental section of the Graduate Record Examination (GRE). The results of the current study indicated that the MST is relatively robust to considerable deviations between the clone's item statistics used for routing and scoring and the properties, or difficulty, of the clone as seen by the examinee.

Keywords: automatic item generation, multistage testing, adaptive testing, item clones, errors in item parameters

Computerized adaptive testing is used in achievement, licensure, and credentialing tests with increasing popularity. As tests are tailored to the examinees and administered with greater frequency, larger item banks are needed in order to maintain the security of the test and therefore its validity (Mills & Steffen, 2000; Stocking & Lewis, 2000). One strategy to address the need for more items is to automatically generate items (Bejar, 2010). This simulation study looked at the accuracy of examinees' ability estimates when automatically generated item clones were used in the context of a multistage adaptive test.

Multistage Testing (MST)

MST adapts the test to the examinee by sets of items, known as modules, rather than at the item level. It is a popular form of adaptive testing that retains most of the positive features of item-level adaptive testing and improves upon some of its less desirable characteristics (Luecht & Nungester, 1998). MST improves measurement across the ability scale and can usually do so with fewer items than in a fixed, linear test (Patsula, 1999).

In MST, an examinee responds to all items in a module and is then routed to the next module. Unlike item-level adaptive testing, in MST the test developers can inspect all possible test forms that examinees could encounter. From the examinee's point of view, MST can offer a reduction in test anxiety in that the examinee can move around within a module, skip a question, and return to it later (Hendrickson, 2007). Because the modules are assembled in advance, the content coverage of each module can be inspected, as opposed to item-level adaptive testing, where the particular set of items an examinee will see is not known until test administration. This pre-assembly allows test developers to assemble modules that are approximately equal in time to complete, overcoming another potential drawback of item-level adaptive testing (Bridgeman & Cline, 2004).

As in any test development, there are many decisions to be made, for example, test length. In MST, however, there are additional considerations, including number of stages, number of modules, and number of items per module (Yan, von Davier, & Lewis, 2014). These decisions, along with the range of item difficulty and overlap of difficulty of the modules, all affect the quality of measurement. For a complete discussion of these decisions and their implications, see Zenisky and Hambleton (2014).

Automatic Item Generation

As testing in general becomes more prevalent and computer-based tests of all types are more widely used, many items are needed to build the large item banks required to support these tests from an operational and security standpoint. This has led to increased interest in and development of processes by which items can be automatically generated and used in operational assessments (Drasgow, Luecht, & Bennett, 2006; Gierl & Haladyna, 2013). As the generation of these items continues to improve, it is even hoped that not all items would need to be pilot tested (Luecht, 2013).

Automatic item generation comes in several different forms. The first, and most straight-forward, is identifying features of items that can be easily manipulated. For example, the numbers in a mathematics problem-solving item can be changed to create a different problem. If the problem is intended for elementary school children, the replacement set of numbers might be only positive integers less than 10, whereas two-digit integers could be used to generate problems for middle-school students. In this context, the "new" items can be considered clones of the original items, as

they have changed only in surface characteristics. In addition to providing items automatically, it might also be possible to predict the item statistics of these cloned items by knowing the item statistics of the original items from which they are cloned, referred to as "parent" items.

Validity

Along with affording logistical benefits, automatic item generation can play a role in building a validity argument. Automatic item generation begins with analyzing the cognitive and content aspects of an item before deciding how it can be cloned or adapted to create new items that still measure the same content and tap into the same cognitive domain (Gierl & Haladyna, 2013; Luecht, 2013). Bejar (1993, 2013) addressed the concerns that modeling responses and scores ignores the response processes and the underlying construct the test is intended to measure.

The ability to predict what makes an item difficult is an important contribution to construct validation (Mislevy, Sheehan, & Wingersky, 1993). One possibility is that items can be automatically generated by developing item models in which item features are purposefully manipulated so that specific aspects of the construct are tested and the item difficulties can be predicted. Some cognitive psychological research relates the psychometric properties, including item difficulties, of automatically generated items with their parent items, or the item models from which they are derived (Embretson & Daniel, 2008; Enright, Morley, & Sheehan, 2002; Graf, Peterson, Steffen, & Lawless, 2005; Newstead, Bradon, Handley, Dennis, & Evans, 2006). The concept of item generation would allow more attention to be paid, not only to the content in terms of test specifications, but also to the underlying construct itself. If the research in developing item generative methods results in a deeper understanding of what processes an item elicits from the examinee, the potential exists for developing a test that truly measures the underlying construct of interest (Embretson, 1999).

Ability Estimation

In linear (i.e., non-adaptive) testing, item calibration typically occurs after test administration. In an adaptive test, previously calibrated items are used to estimate the examinee's ability and route him or her through the test as it is delivered. However, if items are generated on the fly, the psychometric properties of those items are unknown. For such an item, one solution is to substitute the psychometric properties of that item's "parent" for those of the cloned item. The work of Sinharay and Johnson (2008) documented the variability of item difficulty within item families. Because of this variability, substituting parent information for the cloned items is not a perfect solution. Although there has been work that investigated the accuracy of predicting item statistics, the effect of using these inherited statistics on resulting ability estimates has not yet been studied.

In a study using mathematical models and cognitive theory, Embretson (1999) found that predicted item parameters correlated only in the .70s and .80s with calibrated item parameters, while the correlation was below .50 when the model was based on item features. As reported in Gorin and Embretson (2013), most studies found that item difficulty models explained between 30% and 60% of the variance in predicted item difficulty. Ability estimates contain two sources of error. First, as ability estimates are based on items and responses to them, the error in item parameter estimates (whether calibrated or predicted), is manifested in the ability estimate. Second, the error associated with the difference between true ability and estimated ability is still a factor.

To study the effect of the uncertainty of item parameters on ability estimates, Embretson (1999) conducted a simulation study incorporating different levels of uncertainty in the item parameters in a fixed-length test. She found that the uncertainty did lead to increased bias and decreased precision in the ability estimates, particularly at the extremes, because very easy or very difficult items tended to have clones whose predicted difficulties regressed to the mean. She concluded that by increasing test length by only a few items, specifically easy or difficult items, the adverse effects of using estimated item parameters could be overcome. In the context of MST, this could be accommodated by increasing the number of items in item modules that target the extremes of the ability scale.

The effect of imprecise parameter estimates on ability estimates has been studied in a variety of testing situations. For a fixed, linear test, Zhang, Xie, Song, & Lu (2011) developed asymptotic expansions of the maximum and weighted likelihood estimators of abilities that quantify the effect of the imprecision or bias of item parameters on the ability estimates. In an item-level adaptive test, systematic bias in person abilities has been observed even when the imprecision in the item parameters is unbiased (Doebler, 2012). In the case of a variable-length adaptive test, the effect of imprecision of the item parameters on ability estimates varied for different stopping rules. The largest effect on ability estimates occurred when stopping rules were dependent on the standard error of the examinee's ability estimate (Patton, Cheng, Yuan, & Diao, 2013). Taken together, these studies indicate that there is an impact of imprecise parameter estimates on examinee ability estimates but that the specific test format and nature of the imprecision can lead to very different effects on the ability estimates.

Purpose

The purpose of this study was to determine how accurately examinees' abilities can be estimated when automatic item generation is used in an MST. For adaptive testing procedures to route examinees through the test, the psychometric properties of each item must be known. Even though item clones are generated so that they will inherit the psychometric properties of the parent item, these items do not necessarily behave exactly as the parent item. Depending on the parent item and the changes made to it, the properties of the cloned item could be very similar or quite different from the parent item. This study examined the effect of using these inherited psychometric properties for automatically generated items to determine the accuracy of examinees' ability estimates. For testing programs for which assigning a score to each examinee is of utmost concern and the score has important consequences for the individual, the accuracy with which each examinee's true ability can be recovered is paramount. The study varied several MST conditions including the number of testing stages and the number of items per stage, as well as conditions related to automatically generated items, such as percentage of items automatically generated and the variability of the cloned items' inherited properties from their actual psychometric properties.

Method

Item Statistics

Conventionally, items that have been previously administered and calibrated are used in MST. In this study, items were simulated to represent two cases: (1) items that have been previously administered and (2) item clones that were generated on the fly at the time of administration. The

item statistics for the items came from an operational large-scale assessment. The descriptive statistics for each module are shown in Table 1, where *a* is item discrimination, *b* is item difficulty, and *c* is the pseudo-guessing parameter of the three-parameter logistic item response theory (IRT) model.

Table 1. Mean and Standard Deviation (SD) of Item Parameters in the 1-3 MST Design

	a		*b*		*c*	
Module	Mean	SD	Mean	SD	Mean	SD
Stage 1	1.011	0.15	0.042	1.13	0.144	0.08
Stage 2 Low	1.043	0.23	−2.066	0.60	0.185	0.08
Stage 2 Medium	1.177	0.27	0.068	0.68	0.148	0.08
Stage 3 High	1.090	0.14	2.050	0.53	0.144	0.12

For cloned items, there were two sets of item statistics. The first set were the item parameters inherited from the parent items (parent parameters); these were the parameters that the test developer uses when constructing the test and establishing cut scores for routing and scoring examinees. However, the clones might not behave as expected, so a second set of item parameters were necessary to reflect the function of the item in the operational test (clone parameters). The item clone parameters were used to simulate the examinees' responses, while the parent item statistics were used in the routing and scoring procedures of the MST, as these would be the only item parameters available to the test developer at the time of administration.

The objective of generating item clones is to create items with the same item statistics as the parent item. In this study, to simulate a clone, a previously administered item with known item statistics was considered the parent item and replaced with its clone. The inherited item statistics for this clone were the same as the parent's item statistics for all three parameters. It should be noted that the parent items were simply non-clone items from the baseline condition. For example, if item 7 was to be a clone in a new condition, the inherited statistics for that clone were inherited from item 7 in the baseline condition. This allowed for a straightforward comparison across conditions and with the baseline. Simply replacing an item with its clone, with the same parameters as the original parent item, would seemingly not change anything, except for the fact that the inherited item parameters might not accurately reflect how these new items actually function. While the cloned item was given the same item parameters as its parent, these item parameters were used only for the purpose of establishing the cut points used for routing, just as the calibrated item statistics were used for the previously administered, non-cloned items. These item parameters were not used for the purpose of estimating examinee ability.

Regardless of the efforts to generate operationally identical clones, the clone's true item difficulty might differ slightly or considerably from that of the parent item. To account for the potential difference between the observed and inherited item parameters for the cloned items, the second set of item statistics reflected the observed item parameters. This second set of item statistics were used to simulate examinees' responses to the cloned items and these item statistics represented the items' true difficulties. The results of Sinharay and Johnson (2008) were used as the basis for simulating the true item parameters for automatically generated items within an item family.

To generate the item statistics for the cloned items, the *c* parameter for each cloned item was inherited from its parent. The cloned item's difficulty and slope parameters were randomly

generated according to previous research on the behavior of cloned items (Sinharay & Johnson, 2008). Sinharay and Johnson's observed variances of the difficulty and the log-slope parameters were used to establish the deviations of the cloned item parameters from those of the parent items.

The difference between a parent item's difficulty and the range of its clones' true difficulties were categorized into three groups: (1) small—within 0.2, (2) moderate—greater than 0.2 and less than 0.4, and (3) large—greater than 0.4 and less than 0.6.

The variation of the log-slope parameter was relatively consistent over sets of item siblings and was not clearly associated with the variability of the difficulty parameter. For this reason, all item clones were simulated under one condition of variability of slope parameters. To replicate the behavior of the log-slope parameter from Sinharay and Johnson (2008), a random number from a uniform distribution from −0.33 to 0.33 was added to the logarithm of a parent item's slope parameter. The exponentiated result of this sum was the slope parameter of the item clone.

The percentage of item clones (four levels) and the magnitude of the clones' variation in item difficulty (three levels) were varied for 12 conditions, in addition to two conditions where the number of clones in each stage differed, for a total of 14 conditions. A small pilot study was conducted to select clone conditions that were most promising with respect to bias and consistency. Conditions were retained if the absolute bias was less than 0.10. The most erratic conditions involved clones with large variability; these conditions also had absolute bias greater than 0.10. Ultimately, seven clone conditions plus the baseline condition of non-cloned items were chosen for further investigation. These eight conditions, shown in Table 2, were replicated 100 times.

Table 2. Description of Study Conditions

Variability Between Clone and Parent Item Difficulty	Stage 1	Stage 2
Baseline: no item clones	No Clones	No Clones
One-third of items cloned with small variability	33 Small	33 Small
One-half of items cloned with small variability	50 Small	50 Small
All items cloned with small variability	100 Small	100 Small
One-half of items cloned with moderate variability	50 Moderate	50 Moderate
One-third of items cloned with large variability	33 Large	33 Large
One-half of Stage 1 items cloned with moderate variability	50 Moderate	No Clones
All Stage 1 items cloned with moderate variability	100 Moderate	No Clones

Examinees

For each replication of each condition, 1,000 examinees were simulated at each of the 61 locations evenly spaced along the θ ability scale from −3.0 to 3.0 in increments of 0.1. These θ values were considered the examinees' true ability. Although this was not a realistic θ distribution for most testing programs, it provided a reasonable number of examinees at the extremes of the distribution so that the accuracy of each test design could be evaluated along the θ scale. Since the item statistics were pre-determined, the examinees' responses did not affect the item statistics.

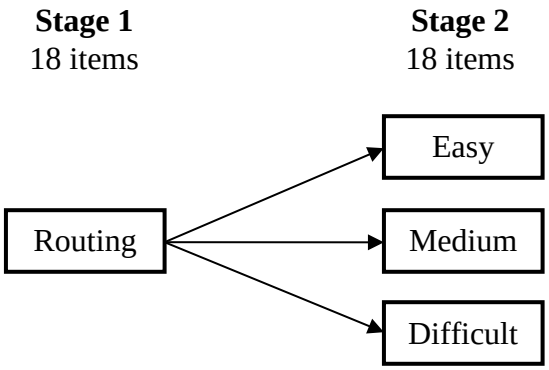
Examinees' responses were simulated using the three-parameter logistic model (Hambleton, Swaminathan, & Rogers, 1991; Birnbaum, 1968) with D = 1.7 and item statistics appropriate to the given test condition. The known item statistics were used for non-clones, whereas those of the

item clones were based on the item statistics inherited from the parent, with changes to the difficulty and slope as described above. In each simulation and in each condition, the probability of a correct response was calculated for every item encountered by each examinee using the observed parameters, that is, the parent parameters plus appropriate deviations to simulate the cloned items. A uniform random number between 0 and 1 was also generated for each. If the probability of a correct response was greater than the random number generated, the response was coded as correct; and if less than the random number, it was coded as incorrect. The simulated examinees' responses were then used for routing and scoring. It should be noted that the parameters of the clones were used to simulate examinee responses and their scores, while the parameters of the parent items were used for "administrative purposes," such as establishing cut scores for routing and scoring.

MST Design

As they are the most common in research and practice, two- and three-stage designs were used in this study (Yan et al., 2014) with different numbers of modules per stage. A total of four test designs were studied; but because the results and conclusions were almost identical across all designs, only the two-stage design with one routing test and three modules is described here. The 1-3 design is shown in Figure 1 and had a total of 36 items, with 18 items per stage.

Figure 1. Two-Stage Design with Difficulty Level of Each Module



Routing and Scoring

The cut points used for routing examinees to the appropriate Stage 2 module were located at θ = –1, and 1. These points were chosen so that each module in the second stage would roughly cover the same distance along the θ scale. The routing cut points on the θ scale were then converted to a corresponding number of correct items for an examinee with a θ level equal to the θ cut point. Suppose that the routing cut point on the θ scale was 1.0. The expected number of correct answers in the first stage module, or routing test, for an examinee with θ=1.0 was computed based on the item statistics for each item in Module A. For an item clone, the parent item statistics were used. As described above, the parent statistics were the only set of item statistics known to the test administrators at the time of administration. For an examinee with an ability of θ, the expected

number of correct responses to a set of items is the sum of the individual probabilities of correctly responding to each item at the given θ level:

$$E\left(\text{number correct} \mid \theta_j\right) = \sum_{i=1}^{n} P_i\left(\theta_j\right), \tag{1}$$

where θ_j is the ability level of interest, P_i is the probability of correctly responding to item *i*, and *n* is the number of total items. The probability of correctly responding to each item was based on the three-parameter logistic model and the parent item statistics.

Both routing and scoring were based on number correct, rather than on an examinee's estimated θ, to replicate what is done in practice with the GRE (Robin, Steffen, & Liang, 2014). Number-correct scoring is robust and sufficient in most cases (Luecht & Nungester, 1998) and is easier to explain to a non-technical audience. More importantly, examinees with an aberrant response pattern will have a more stable ability estimate with number-correct scoring than with estimates based only IRT (Robin et al., 2014). The examinee's final score on the θ scale was based on the total number of items correct across all stages. Using the procedure described earlier to determine the expected number correct for a given θ level, a conversion chart from number correct to ability on the θ scale was constructed. While examinees were simulated to have abilities from –3 to 3, the ability estimates ranged from –4 to 4. Any θ estimate that would have been less than –4 or greater than 4 was reported as –4 or 4, respectively. Each path through the MST had its own conversion chart because the complete set of items was unique for each path.

Target Test Information Functions

A target test information function (TIF) was constructed for each module in the MST based on the parent item statistics. The TIF for a given module reflected the desired information and therefore had maximum values around the cut points for routing for that particular stage. The TIF for the Stage 1 routing module ideally would have local maximum values near –1 and 1, the cut points for routing examinees into the appropriate module in Stage 2.

The target TIFs for the Stage 2 modules were designed with maximum information for a specific set of θs. For the easy module, the maximum information was desired for θs less than –1; for the medium module, between –1 and 1; and for the most difficult module, maximum information was sought for θs greater than 1. As real items are selected, the resulting TIF is evaluated; and a decision is made whether to swap out items in an attempt to obtain a more ideal TIF. At some point, the psychometric properties of the items make it impossible to achieve an ideal TIF.

Evaluation Criteria

One goal of this study was to determine the accuracy with which examinees' abilities were estimated under the conditions described above. Systematic error, random error, and an overall measure of error were all considered. To quantify systematic error, or bias, the average of the deviations between θ estimates and true θ was calculated:

$$\text{Bias} = \frac{\sum_{j=1}^{J}\left(\hat{\theta}_j - \theta_j\right)}{J}, \tag{2}$$

where $\hat{\theta}_j$ and θ_j are the estimated and true abilities for examinee *j*, respectively, and *J* is the total number of examinees. The calculation of bias allows deviations in opposite directions, in effect, to cancel out. However, the bias calculation indicates whether the deviations are greater in one direction than the other.

The standard deviation of the θ estimates, conditioned on true θ, was used as a measure of random error. The overall error was summarized by the RMSE,

$$RMSE = \sqrt{\frac{\Sigma\left(\hat{\theta}_j - \theta\right)^2}{J}}. \quad (3)$$

The bias and RMSE for each replication were averaged over the 100 replications. The medians of the standard errors for each replication were reported.

Additionally, the path through the test was considered. For a given examinee's true θ, there is an optimal path. The percentage of examinees for each test design who took their optimal path, as well as those who took the path adjacent to their optimal path, was computed.

Results

Variability Resulting From Item Clones

Across all conditions, the mean difficulties for each of the four modules were almost identical to the mean difficulties for the test with the parent items, or the baseline condition (see Table 3). This was reasonable, considering how the clones were simulated. A random number within the boundaries of the small, moderate, or large guidelines was added to the parent item's difficulty.

Table 3. Mean and SD of Item Difficulty in Each Stage Across Conditions

	Stage 1: Routing		Stage 2: Easy		Stage 2: Medium		Stage 2: Difficult	
Condition	Mean	SD	Mean	SD	Mean	SD	Mean	SD
No Clones, No Clones	0.042	0.000	−2.066	0.000	0.068	0.000	2.049	0.000
33 Small, 33 Small	0.043	0.016	−2.067	0.017	0.067	0.016	2.048	0.016
50 Small, 50 Small	0.044	0.019	−2.068	0.018	0.068	0.019	2.048	0.022
100 Small, 100 Small	0.044	0.028	−2.069	0.025	0.069	0.030	2.047	0.027
50 Mod., 50 Mod.	0.035	0.044	−2.074	0.053	0.055	0.058	2.053	0.048
33 Large, 33 Large	0.030	0.079	−2.074	0.069	0.058	0.073	2.048	0.074
50 Mod., No Clones	0.043	0.055	−2.066	0.000	0.068	0.000	2.049	0.000
100 Mod., No Clones	0.037	0.069	−2.066	0.000	0.068	0.000	2.049	0.000

The correlations between the slopes and difficulties for the parent items and clones for each stage within each condition are reported in Tables 4 and 5, respectively. Note that modules with no clones have no values reported.

In Figure 2 the resulting TIF for each of the three paths in the 1-3 design are presented, where Path 1 is for low performers, Path 2 for medium, and Path 3 for high performers. The path is the

Table 4. Median and Range of Correlations of Slope Parameters Between Parent Items and Clones

	Stage 1: Routing			Stage 2: Easy			Stage 2: Medium			Stage 2: Difficult		
Condition	Mdn	Lo	Hi	Mdn	Lo	Hi	Mdn	Lo	Hi	Mdn	Lo	Hi
33 Small, 33 Small	0.65	−0.75	0.99	0.79	−0.30	0.98	0.85	−0.26	0.99	0.54	−0.70	0.96
50 Small, 50 Small	0.61	0.07	0.95	0.78	−0.06	0.98	0.78	0.17	0.95	0.57	−0.41	0.91
100 Small, 100 Small	0.64	0.19	0.81	0.77	0.52	0.92	0.77	0.49	0.92	0.56	0.01	0.81
50 Mod., 50 Mod.	0.63	−0.64	0.90	0.76	0.36	0.95	0.79	0.34	0.95	0.59	−0.38	0.96
33 Large, 33 Large	0.63	−0.28	0.94	0.78	0.02	0.98	0.75	−0.96	0.98	0.54	−0.40	0.97
50 Mod., No Clones	0.63	−0.14	0.95									
100 Mod., No Clones	0.62	0.33	0.84									

complete path including all the items from Stage 1 and the items from one of the three modules in Stage 2.

Even when all items were cloned, and the item difficulties of the clones were simulated to vary largely when compared with the parent items, the TIFs for the three paths do not differ greatly from the original.

Results of the Simulation Study

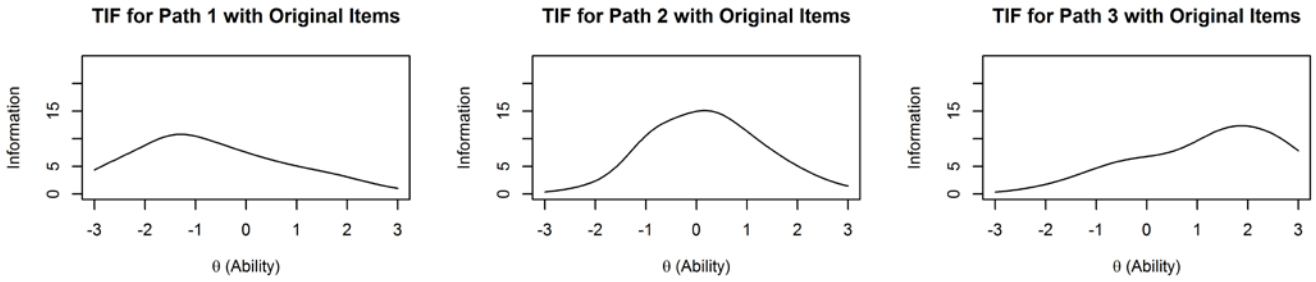
The systematic error, or bias, followed a consistent pattern across all conditions, as seen in Figure 3. The most striking result is that the bias is larger than in the other conditions for those examinees whose estimated number-correct scores are below average; whereas for stronger examinees, the cloned conditions exhibit more bias. Even though this difference is interesting, the magnitude of the difference in bias across all conditions is less than 0.04 of a standard deviation along the θ scale; in fact, the difference is smaller than 0.02 for most estimated number-correct scores.

Table 5. Median and Range of Correlations of Difficulty Parameters Between Parent Items and Clones

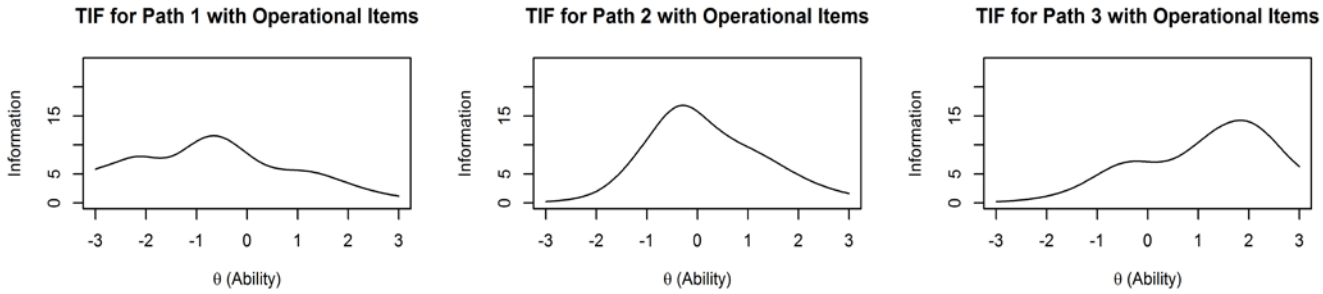
	Stage 1: Routing			Stage 2: Easy			Stage 2: Medium			Stage 2: Difficult		
Condition	Mdn	Lo	Hi	Mdn	Lo	Hi	Mdn	Lo	Hi	Mdn	Lo	Hi
33 Small, 33 Small	0.99	0.96	1.00	0.99	0.87	1.00	0.99	0.84	1.00	0.98	0.85	1.00
50 Small, 50 Small	0.99	0.99	1.00	0.98	0.94	1.00	0.99	0.97	1.00	0.98	0.92	1.00
100 Small, 100 Small	0.99	0.99	1.00	0.98	0.97	0.99	0.99	0.98	0.99	0.98	0.97	0.99
50 Mod., 50 Mod.	0.97	0.91	0.99	0.90	0.62	0.98	0.93	0.60	1.00	0.86	0.71	1.00
33 Large, 33 Large	0.93	0.64	1.00	0.77	−0.01	1.00	0.82	−0.11	1.00	0.79	0.26	1.00
50 Mod., No Clones	0.97	0.88	1.00									
-100 Mod., No Clones	0.97	0.96	0.98									

Figure 2. TIFs for Three Possible Complete Paths Through 1-3 Test Design for Three Conditions

a. Original Items–No Cloned Items



b. 100% of Items Are Cloned With Large Variability



c. 100% of Items Are Cloned With Small Variability

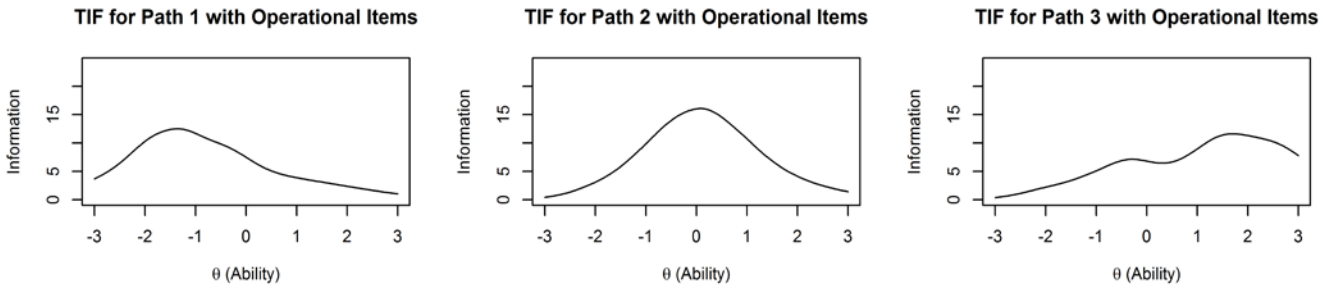
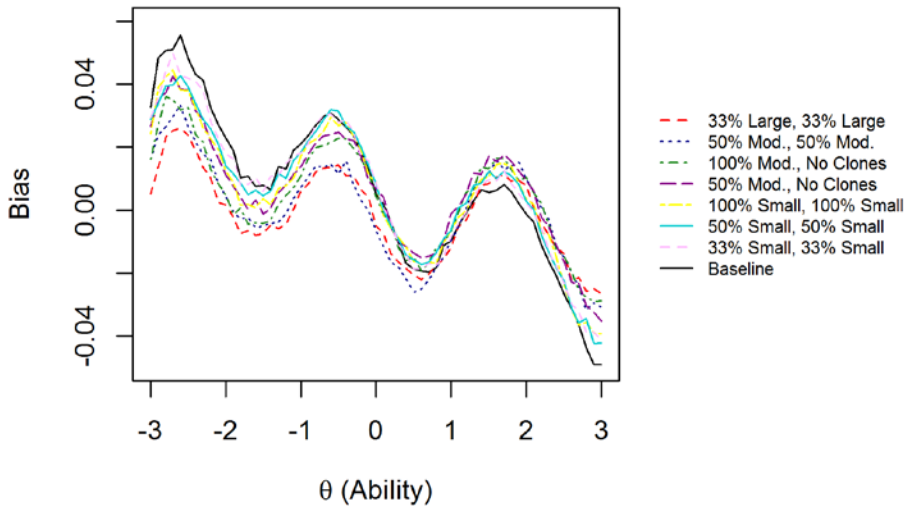


Figure 3. Bias Conditioned on θ for All Conditions



When evaluating random error and overall error using the standard deviation of observed θs and RMSE, respectively, the differences are so small that they are difficult to see when looking at the full scale, as shown in Figures 4 and 5.

Figure 4. Standard Deviation of θ Estimates Conditioned on θ

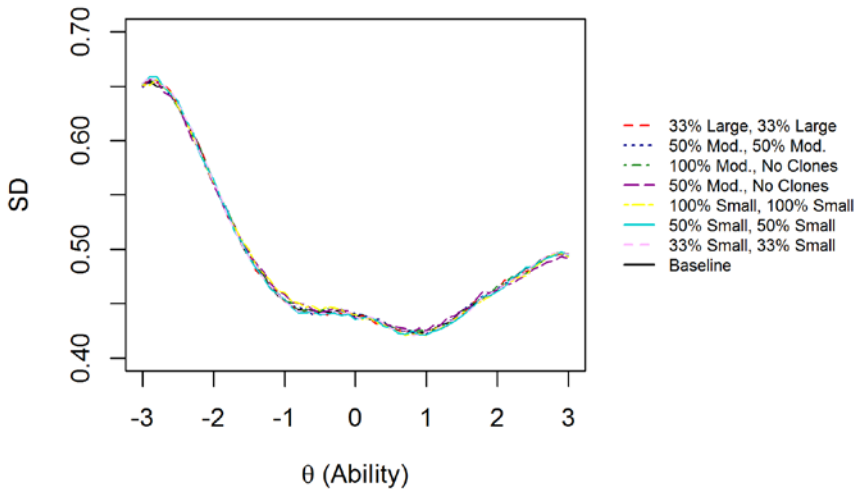
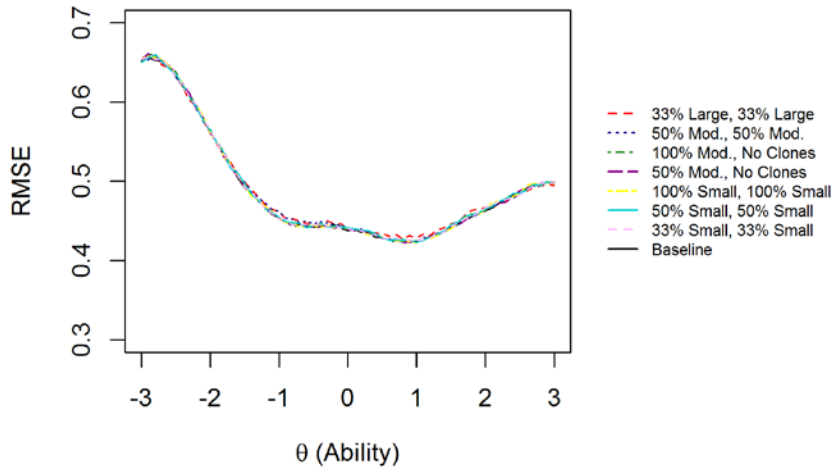
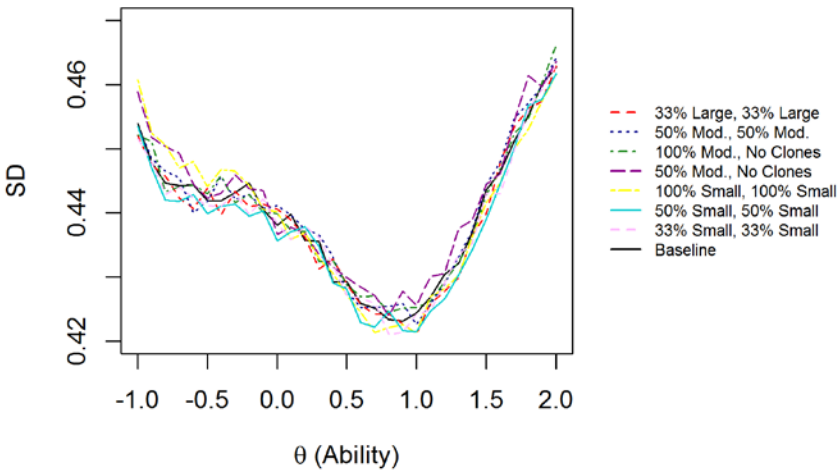


Figure 5. RMSE Conditioned on θ



To distinguish among the patterns of random error across the conditions, the patterns between true θs of −1.0 and 2.0 were examined, rather than looking across the entire scale (see Figure 6). There is no condition that systematically has more random error than the other conditions or is different from the baseline condition. In addition, for each expected number-correct score the variability in standard deviations is not more than 0.01 on the θ scale.

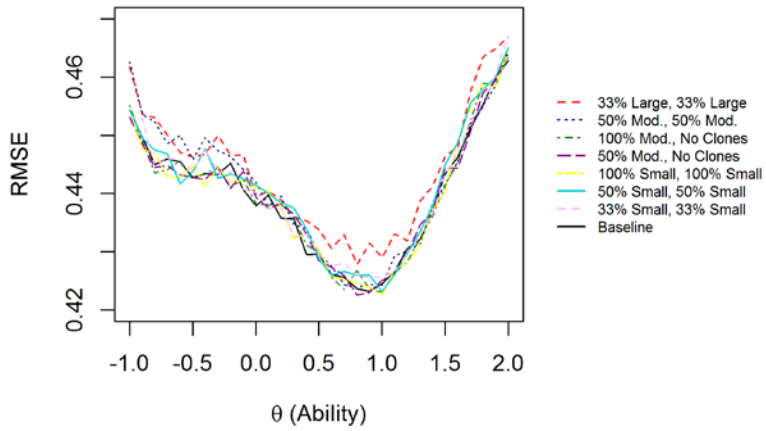
Figure 6. Standard Deviation of Observed θ Estimates Conditioned on θ, from θ = –1.0 to 2.0



The magnified RMSE results shown in Figure 7 indicate that the greatest RMSEs occurred in those conditions when one third of the items were clones with large variability and when half of

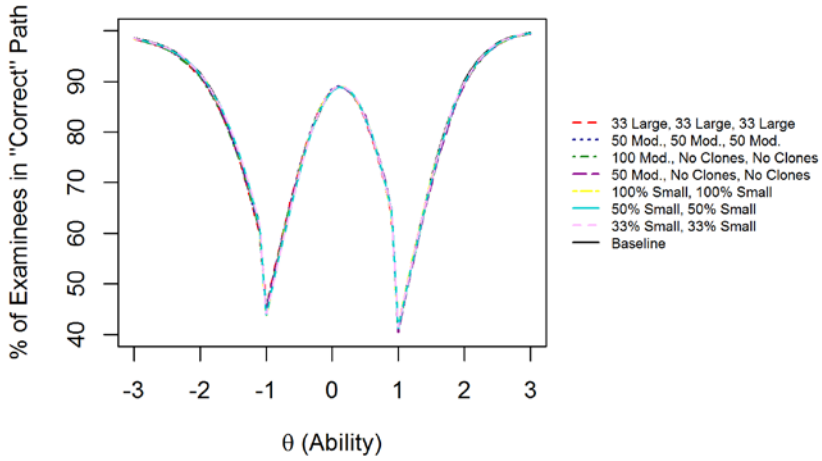
the items were clones with moderate variability. Once again, the magnitude of the differences between the cloned conditions and the baseline were small, less than 0.01 on the θ scale.

Figure 7. RMSE Conditioned on θ from θ = –1.0 to 2.0



Correct path. Given an examinees' true θ, there is one path that is the "true" path through the test. The percentage of examinees that followed their true path through the test is shown in Figure 8. Examinees with true θs closer to the cut-off values were more likely to be placed in a path adjacent to their true path.

Figure 8. Percentage of Examinees Following Their True Path Through the Test



Discussion

This study investigated the impact of using inherited item parameters for cloned items on the estimation of examinee ability. The key feature of the results is that the magnitude of any differences when the conditions were compared to the baseline, even in the most extreme

conditions, were only a few hundredths of a standard deviation, or less than one percent, over the range of the scale. Overall, the results were consistent with expectations that the more items that are cloned and the more the clones vary from their parent items, the more bias in the ability estimates, even though the bias was small in all cases.

Although there was considerable fluctuation in the precision of the examinee ability estimates along the θ scale, this was due to the distribution of the item difficulty parameters used. The item statistics for the parent items were taken from an operational testing program and were not ideal for creating a set of items such that the resulting TIF would be completely flat across the ability scale, as would be optimal when attempting to provide all examinees with estimates of their ability with equivalent errors of measurement. Had more extreme items (both easy and difficult), been available), the TIFs for the second stage modules would have been more consistent for all examinees. However, for the purposes of this study, all conditions including the baseline had the same issue with lack of precision for the lower ability levels. Even in the baseline condition with no cloned items, the bias changed along the θ scale. This fluctuation was consistent across the different clone conditions and does not affect the interpretation of these simulated results. Nevertheless, an operational testing program interested in accurately measuring the examinees all across the ability scale, not just at particular cut scores, for example, would want to address this fluctuation by selecting more appropriate sets of items, if possible.

Limitations

Because item clones were simulated based on empirical results where there was no evidence of systematic bias between the item statistics of a family of clones and the associated parent item, the result was that the presence of clones and their inherited item statistics, in effect, cancelled out. For example, item difficulties were randomly chosen to be over- or under-predictions of the true item difficulty. The extreme case where the difficulties of all clones were either over- or under-predicted was not explored. This is a potential limitation of the study in that clones could be created in such a way as to be systematically more difficult or easier in a given assessment.

The study also has a possible lack of generalizability because the properties of the cloned items were based on the results of one operational study (Sinharay & Johnson, 2008). Clones generated in different scenarios could have larger deviations from the parent items than were used here. The observed correlations of the item statistics between the parent and cloned items are stronger than were observed in other studies of automatic item generation. This study, however, used a specific example of automatic item generation, in particular, using item clones. In addition, the *c* parameter was not varied between the clone and the parent items.

Future Research

This study was based on the assumption that the goal of the test was accurately measuring examinees along the ability scale. Nevertheless, it would be interesting to consider the implication of automatic item generation in an MST in which the test developers were only interested in categorizing examinees, such as basic, proficient, and advanced, or even simply as passing or failing as in a credentialing exam. For example, if the desire for greater measurement precision is located at only a few points along the ability scale, then one could consider using clones only if the item difficulty is a certain distance away from the location of the cut points.

This study used number-correct scoring rather than IRT-based scoring for routing examinees and providing the final ability estimates. Because number-correct scoring is less computationally intensive, it is reasonable to use it for routing examinees through the test; but the final scoring of the examinees could be done using IRT-based scoring. Even though this could offer a slightly more accurate scoring of examinees, the magnitude of the improvement would need to be studied. Additionally, an investigation into the relationship of test length and use of clones would demonstrate how many additional items would be needed to maintain a specified level of accuracy. A testing program might accept a moderate increase in test length if it could avoid pilot testing items.

If scores do not need to be reported immediately, it would not be necessary to use the hypothesized item statistics for the cloned items when scoring the examinees. The hypothesized item statistics, or item statistics of the parent items in this study, would still be used for routing. If time allowed, however, the automatically generated items that had not been calibrated or pre-tested could be calibrated after test administration as in a linear test administration, which conducts calibration and equating after administration.

The most intriguing future research in this area is, as might be expected, the more difficult: Developing accurate and useful item models for improved automatic item generation. The more that a testing program can generate items that behave as expected and appear different to the examinees, the more accurate the examinees' ability estimates will be with the added benefit of a reduction in item exposure.

Conclusions

The efficacy with which items cloned on the fly, with no time for calibration or pre-testing, can be used in an MST hinges upon the ability to accurately model item statistics for the cloned items. The simulation in this study mimicked the degree of variation between a cloned item and its parent item found by Sinharay and Johnson (2008) in quantitative Graduate Record Examination (GRE) items in an operational setting. The current study used the item statistics of the clones to simulate examinees' responses but used the item statistics of the parent items for routing decisions and scoring. This afforded the ability to determine how accurately an examinee's ability level can be recovered when the items administered to the examinee have unknown item statistics and the testing program can only use the item statistics of each item clone's parent item. In this simulation study, the degree to which each item clone differed from its parent item was varied and was considered to have a small, moderate, or large difference. The percentage of items that were simulated to be item clones was also varied.

This simulation study, in a sense, presented the worst-case scenario of employing automatic item generation in an MST. All cloned items were assumed to have the same item statistics as the parent items; examinee scoring was based on the parent items' statistics rather than delaying the scoring until the new, cloned items had been calibrated; and item clones were distributed evenly along the item difficulty scale, rather than inserting clones strategically to replace items along the ability scale where ability estimates are more accurate. Yet even with this, the accuracy observed in this simulation could still allow some testing programs to incorporate automatic item generation and maintain the integrity of their test. On a 200- to 800-point scale, the bias was at worst four points larger than that of the baseline. Furthermore, in this study the calibrated parameters of the non-cloned items were assumed to be without error, which is not realistic; thus the impact of cloning would be even smaller when compared to the baseline condition. The potential for auto-

matic item generation use is great, with applications ranging from achievement and credentialing exams to summative and formative assessments in K-12.

References

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R.J., Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (1st ed., pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum.

Bejar, I.I. (2010). Recent development and prospects in item generation. In S.E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 201-226). Washington, DC: American Psychological Association. *CrossRef*

Bejar, I.I. (2013). Item generation: Implications for a validity argument. In M.J. Gierl & T.M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 105-122). New York: Routledge.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41*(2), 137-148. *CrossRef*

Doebler, A. (2012). The problem of bias in person parameter estimation in adaptive testing. *Applied Psychological Measurement, 36*(4), 255-270. *CrossRef*

Drasgow, F., Luecht, R.M., & Bennett, R. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

Embretson, S.E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433. *CrossRef*

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50*(3), 328-344.

Enright, M.K., Morley, M., & Sheehan, K.M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49-74. *CrossRef*

Gierl, M.J., & Haladyna, T.M. (2013). *Automatic item generation: Theory and practice*. New York: Routledge.

Gorin, J.S., & Embretson, S.E. (2013). Using cognitive psychology to generate items and predict item characteristics. In M.J. Gierl & T.M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 136-156). New York: Routledge. *CrossRef*

Graf, E.A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models*. (ETS Research Report RR-05-25). Princeton, NJ: Educational Testing Service.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26(2),* 44-52. *CrossRef*

Luecht, R.M. (2013). Automatic item generation for computerized adaptive testing. In M.J. Gierl & T.M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59-76). New York: Routledge.

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249. *CrossRef*

Mills, C.N., & Steffen, M. (2000). The GRE computerized adaptive test: Operational issues. In W.J. van der Linden & C.A.W. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston, MA: Kluwer. *CrossRef*

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55-78. *CrossRef*

Newstead, S.E., Bradon, P., Handley, S.J., Dennis, I., & Evans, J.S.B.T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning, 12*(1), 62-90.

Patsula, L.N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing*. (Unpublished doctoral dissertation), University of Massachusetts at Amherst.

Patton, J.M., Cheng, Y., Yuan, K., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*(1), 24-40. *CrossRef*

Robin, F., Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE revised general test. In D. Yan, A.A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 325-341). Boca Raton, FL: Taylor & Francis Group.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing*, *8*(3), 209-236. *CrossRef*

Stocking, M.L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston, MA: Kluwer. *CrossRef*

Yan, D., von Davier, A.A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: Taylor & Francis Group.

Zenisky, A., & Hambleton, R.K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A.A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications*. Boca Raton, FL: Taylor & Francis Group.

Zhang, J., Xie, M., Song, X., & Lu, T. (2011). Investigating the impact of uncertainty about item parameters on ability estimation. *Psychometrika*, *76*(1), 97-118. *CrossRef*

Author Address

Kimberly F. Colvin, University at Albany, SUNY, 1400 Washington Ave., EDU 214, Albany, New York 12203, U.S.A. Email: kcolvin@albany.edu.