JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Summer 2019, Vol. 56, No. 2, pp. 437–451

Computerized Adaptive Testing in Early Education: Exploring the Impact of Item Position Effects on Ability Estimation

Anthony D. Albano
University of Nebraska–Lincoln
Liuhan Cai
Measured Progress, Inc.
Erin M. Lease and Scott R. McConnell
University of Minnesota

Studies have shown that item difficulty can vary significantly based on the context of an item within a test form. In particular, item position may be associated with practice and fatigue effects that influence item parameter estimation. The purpose of this research was to examine the relevance of item position specifically for assessments used in early education, an area of testing that has received relatively limited psychometric attention. In an initial study, multilevel item response models fit to data from an early literacy measure revealed statistically significant increases in difficulty for items appearing later in a 20-item form. The estimated linear change in logits for an increase of 1 in position was .024*, resulting in a predicted change of* .46 *logits for a shift from the beginning to the end of the form. A subsequent simulation study examined impacts of item position effects on person ability estimation within computerized adaptive testing. Implications and recommendations for practice are discussed.*

Computerized adaptive testing (CAT) programs rely on the assumption that item parameters are stable and generalizable across different administrations of an item and test. This assumption applies across test forms in the process of item precalibration, where parameters such as item difficulty and discrimination are estimated in one form and then constrained as fixed in one or more other forms. It also applies across test takers, who may receive the same item in different positions, depending on the personalized trajectory of the CAT. In both cases, items are administered under the pretense that changes in their context have a negligible impact on item performance.

Numerous studies have called into question this key assumption of precalibration and CAT, with strategies proposed for estimating and addressing what can be framed as instances of differential item functioning (DIF), where the grouping variable over which item performance differs is based not on person demographics but on features of the item, test, and administration conditions. For example, item parameter drift appears in the form of DIF that is attributed to the passing of time between administrations (Goldstein, 1983). The difficulty of an item may change from when it was originally calibrated, due to item content taking on different meanings over time. In certification testing, items may become more difficult as the knowledge and skills assessed within an item become less relevant to a profession and thus less familiar to test takers, requiring updated content specifications and recalibration (Babcock &

© 2019 by the National Council on Measurement in Education

Albano, 2012). In state achievement testing programs, differential boost is a form of DIF exhibited by item parameters that change for students receiving modified versions of items compared with students receiving unmodified versions of the same items. In this case, items may become less difficult when modifications are provided that make the item content more accessible to a particular student group (Wyse & Albano, 2015). Both of these examples of DIF can be problematic for precalibration and CAT, where it is assumed that item parameters are consistent across administrations, whether over time or for modified vs unmodified items.

This article explores DIF in the form of position effects resulting from changes in the context of a given item within its respective test form, that is, within the set of items appearing prior to and after it (Wainer & Kiely, 1987). Research in this area has focused on educational testing, where an item becomes more or less difficult as it appears toward the end of a test or, alternatively, where a person performs better or worse as they proceed through a test. Increases in performance suggest that practice or learning is taking place, whereas decreases indicate fatigue or disengagement.

For example, Albano (2013) modeled item position effects using data from pilot administrations of the quantitative and verbal reasoning sections of the Graduate Record Examination (GRE). Three quantitative item sets each contained 28 unique items administered in 13 different orderings, and three verbal item sets contained 30 unique items each administered in seven orderings (for additional details, see Davey & Lee, 2011). Results from explanatory item response theory (IRT) modeling (De Boeck & Wilson, 2004), specifically via the Rasch model, showed that performance decreased linearly toward the end of the test, with the majority of items becoming more difficult. Some items became significantly more difficult than the rest, whereas a small minority were estimated to be easier at the end of the test than at the beginning.

Debeer and Janssen (2013) modeled position effects for both items and persons using simulated data and data from two operational tests: a listening comprehension test, wherein two sets of items were presented to some test takers in reverse orders; and the 2006 PISA, where sets of items, called clusters, could appear in one of four positions based on a rotated block design (for details, see Organization for Economic Cooperation and Development, 2009). The simulation study served to demonstrate the basic functionality of an explanatory Rasch model, with the expected result of bias in item difficulty estimates for the traditional Rasch model that ignored simulated item position effects. In both of the operational data studies, overall performance was again found to be higher at the start of the test and lower at the end. However, the reverse was true for some examinees, suggesting a practice or learning effect, or a resilience to the negative influence of position in the test.

Although position effects have been documented in a variety of settings (e.g., Alexandrowicz & Matschinger, 2008; Debeer, Buchholz, Hartig, & Janssen, 2014), previous research in this area has focused primarily on detection, with limited attention given to practical significance, that is, to what extent these effects matter for test developers and test takers. An overarching question that remains unanswered is: How do position effects in precalibration impact person ability estimates in CAT? Furthermore, what magnitude of effects can safely be treated as negligible versus not, and what other factors should be considered?

Doebler (2012) outlines how error in item parameter estimation can carry forward to produce bias in person parameter estimation. Within the context of CAT, discrepancies between estimated and true item difficulty are summarized as coming from four main sources: (1) random error due to sampling of test takers, that is, standard error (SE), which tends to be higher in CAT because of higher item turnover and smaller sample sizes; (2) differences across person groups, that is, DIF; (3) testlet effects, a type of context effect that can arise with item sets that share a common stimulus; and (4) automatic item generation or cloning, where variability in item parameters may be ignored for items generated from the same template. In a series of CAT simulation studies, Doebler (2012) manipulated the first and fourth of these sources, with results demonstrating varying amounts of person parameter bias across different IRT models, estimators, test lengths, and item pool sizes. When item difficulty SE was simulated to be .25, mean bias in person ability exceeded .50 logits under some conditions.

Building on the simulation design from Doebler (2012), this article extends previous research on position effects by addressing the issue of impact in two connected studies. In Study 1, position effects are modeled with operational data from a linear, fixed length assessment used in early education, an area of measurement that has received relatively limited psychometric attention, and one that presents unique and interesting challenges such as more stringent limitations on testing time and test length. In Study 2, results from Study 1 serve as initial generating parameters within a simulation evaluating the influence of position effects on person ability estimation in CAT. Methods and results are presented by study, and then discussed together in terms of practical implications.

Study 1

Data

Data for Study 1 came from a larger project focusing on the development and validation of Individual Growth and Development Indicators (IGDIs; Bradfield et al., 2014) for measuring and improving instruction in early literacy with preschool children. IGDIs are brief, individually administered assessments targeting focused constructs such as phonological awareness, vocabulary, and alphabet knowledge. This study utilized data from an alphabet knowledge task, wherein students were asked to identify a verbally named letter from among three options, one of which was keyed correct. This task, along with others, is intended to be used in progress monitoring settings, where scores support normative and criterion-referenced score interpretations, often around predefined cut scores that help identify students in need of support.

Four test forms were created using 20 items selected to be representative of the alphabet knowledge item bank, with all items written and evaluated with feedback from teachers and content experts. Each of the four test forms contained the same 20 items but in different orders. In form 1, the items were presented sequentially from 1 to 20. Form 2 contained items 16 through 20 and then 1 through 15. Form 3 contained items 11 to 20 and then 1 to 10. Finally, form 4 contained items 6 to 20 and

Table 1
Item Orders by Form for Study 1

Part	Form 1	Form 2	Form 3	Form 4
1	1	16	11	6
	2	17	12	7
	3	18	13	8
	4	19	14	9
	5	20	15	10
2	6	1	16	11
	7	2	17	12
	8	3	18	13
	9	4	19	14
	10	5	20	15
3	11	6	1	16
	12	7	2	17
	13	8	3	18
	14	9	4	19
	15	10	5	20
4	16	11	6	1
	17	12	7	2
	18	13	8	3
	19	14	9	4
	20	15	10	5

then 1 to 5. Thus, each item appeared in each of four quarters or parts of the test, and within a part, the item order was always fixed. Table 1 summarizes the form design.

Students were randomly ordered and then assigned to forms sequentially, with sample sizes of 22, 25, 23, and 26 for forms 1 through 4, respectively. The test was administered individually in classroom settings using mobile tablet devices. Having previously completed similar tasks, students were familiar with the administration procedures.

Analysis

Forms were administered without time constraints. Students were not made aware of how much time had passed, and were not pressured by the examiner to hurry. Thus, it was assumed that any effects of speededness would be negligible. Still, prior to modeling position effects, speededness was examined using item and total completion times. The average time to complete the test was 124 seconds, after subtracting time for instructions. Total completion times ranged from 85 to 258 seconds. Response times by item were found to decrease slightly across the form. The correlation between item position (1 to 20) and response time (by item, ignoring item difficulty and student ability) was $-.12$ ($t = -7.75$, $df = 4334$, $p < .001$). Completion times aligned overall with expectations, given the brevity and simplicity of the task, and the observed tendency for students to lose interest toward the end of the test.

Position effects were modeled, as in previous research, within an IRT framework. Here, the Rasch model is presented in log-odds or logit form. The logit of correct response for person p to item i, η_{pi}, is modeled as a linear function of an intercept γ_0, and random effects for person ability u_{0p} and item difficulty u_{0i}:

$$\eta_{pi} = \gamma_0 + u_{0p} + u_{0i}, \tag{1}$$

where $u_{0p} \sim N(0, \sigma^2_{0p})$ and $u_{0i} \sim N(0, \sigma^2_{0i})$, and γ_0 is a fixed effect estimating overall mean performance across items, persons, and positions. Note that higher values here for u_{0i} indicate items with higher mean performance, i.e., easier items. Reversing the sign will produce the traditional directionality for IRT difficulty.

As a base model, Equation 1, and other formulations like it, can be extended to estimate the effects of other covariates and relationships with item and/or person performance (De Boeck & Wilson, 2004). Examples include the linear logistic test model (Fischer, 1973) and the hierarchical generalized linear model (Kamata, 2001). De Boeck (2008) discusses variations of the Rasch model with random person and item parameters, including formulations for examining DIF.

Covariates representing position within the test form or administration can be added to Equation 1 to estimate a main effect for position and interactions with items or persons. Here, $position_i$ is simply the order of administration of item i, coded as 0 through 19, and γ_1 is the estimated linear change in logit performance for an increase of 1 in item position:

$$\eta_{pi} = \gamma_0 + \gamma_1 position_i + u_{0p} + u_{0i}. \tag{2}$$

Shifting position to be 0 to 19, rather than 1 to 20, results in γ_0 now being the estimated mean performance across items appearing in position 1. Adding to Equation 2, an interaction term allows each item to have its own linear position effect slope, estimated here as a random residual from the main slope in γ_1:

$$\eta_{pi} = \gamma_0 + \gamma_1 position_i + u_{0p} + u_{0i} + u_{1i} position_i, \tag{3}$$

where $u_{1i} \sim N(0, \sigma^2_{1i})$ and a covariance between random effects for items is estimated as σ_{0i1i}.

Equations 1–3 are labeled as M0, M1, and M2. Each was fit to the alphabet knowledge data using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2018). Similar models were also fit with $position_i$ coded as 0 through 3, representing the part of the item within the test, as defined in Table 1. Statistical significance was determined using model fit comparisons, with a statistically significant χ^2 difference and decrease in the Akaike information criterion (AIC) indicating improvement in fit for a more complex model relative to a less complex alternative, that is, M1 compared with M0, and M2 compared with M1.

The random effects formulation of the Rasch model was chosen for theoretical and practical reasons. From a theoretical perspective, the specific items administered here were not the focus of study and individual item parameters were not of interest. Instead, the items were considered to be a sample representing a larger population, or universe (Brennan, 1992), of alphabet knowledge tasks. The focus was on the distributions of parameters for persons, items, and item position interactions. From a practical perspective, models with random persons and items involve fewer

Table 2
Model Fit Results

	df	AIC	BIC	logLik	Deviance	χ^2	df(χ^2)	$p(>\chi^2)$
M0	3	1,782	1,799	−888	1,776			
M1	4	1,780	1,802	−886	1,772	4.43	1	.035
M2	6	1,782	1,815	−885	1,770	1.88	2	.391

Note. AIC = Akaike information criterion. BIC = Bayesian information criterion.

parameters than their fixed effect counterparts, and thus they benefit from parsimony and improvement in model convergence. Furthermore, the means and variances provided by the random effects models served as generating parameters for data simulation in Study 2.

Results

Table 2 contains the model fit results used to compare the base model with the models additionally containing a main effect for position (M0 vs. M1) and interaction effects between item and position (M1 vs. M2). The df in the first column indicates the number of parameters estimated by each model, with 3 for M0 (intercept and variances for persons and items), 4 for M1 (with the addition of a position main effect), and 6 for M2 (with the further addition of a variance for the position effects and covariance with item). The AIC decreased for M1 compared with M0, and the χ^2 test was statistically significant ($p = .03$), indicating improved fit. Interaction effects in M2 were not found to improve fit over M1, with AIC increasing and the χ^2 test not statistically significant. Model comparisons with $position_i$ as 0, 1, 2, 3 similarly indicated M1 as fitting best. Thus, the remaining results focus on $position_i$ coded as 0 through 19.

In M1, logit person ability was estimated to vary on average by $\sigma_{0p} = 2.06$, and for item difficulty, $\sigma_{0i} = .64$. The mean SE of person ability estimates, obtained as the square root of the mean of conditional residual variances for each person, was .77, and the mean SE of item difficulty estimates, obtained similarly, was .28. The fixed effect intercept for mean performance overall at position 1 was 1.85. The main effect for position was then $-.024$ ($SE = .011$, $z = -2.153$, $p = .016$). This position slope corresponds to a change in logits of $-.12$ for a five-item increase in position, and a change of $-.46$ for an increase in position of 19. Recall that, because of the modeling framework used, decreases in logits indicate more difficult items.

Figure 1 depicts the change in item performance across position. In the first plot, change is shown in terms of raw proportion correct, with individual items as gray lines and the mean proportion correct at each position as the solid black line. With the administration design defined as in Table 1, a given item appeared in only four possible positions. This is evident in the item lines each spanning only four change points over 15 positions on the x-axis, for example, 1, 6, 11, and 16 for the item starting in the upper left. In the second plot, change is shown for estimated logit

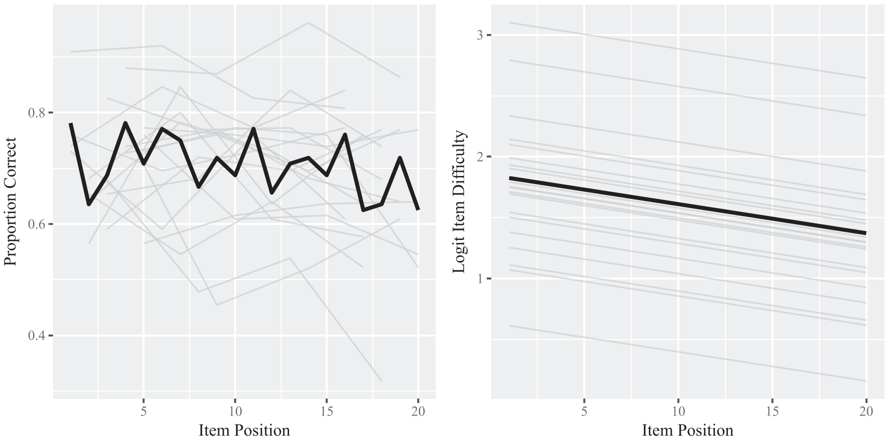

Figure 1. Changes across item position in item performance based on proportion correct, in the first plot, and logit item difficulty, in the second. Within each plot, gray lines represent items, and solid black lines represent means across items.

difficulty, with the main effect slope in the solid black line applied across items, which only differ in their intercepts.

Study 2

Simulated Data

Results from Study 1 aligned with findings from previous research on position effects. Items tended to be easier at the start of the test, with item difficulty increasing toward the end, suggesting a fatigue effect or decrease in motivation and engagement for test takers. The goal of Study 2 was to examine the impact of results such as these on a hypothetical subsequent CAT that might employ parameters like the ones from Study 1 as precalibrated values. Similar to Doebler (2012), CAT was simulated using different sets of generating parameters, where simulated test takers responded based on true item parameters, but the CAT algorithm made decisions based on estimates, as would take place operationally. The design of Study 2 was also informed by practical considerations associated with testing in early education and the use of brief, formative measures of preliteracy skills, such as alphabet knowledge.

Generating parameters were based on results from M1, Equation 2, in Study 1. The CAT item bank contained 200 items with true difficulties b_{true} generated randomly from a normal distribution with $\mu = -\gamma_0 = -1.85$ and $\sigma = \sigma_{0i} = 0.65$. SEs were generated for each item by sampling from a normal distribution with $\mu = 0$ and $\sigma = .28$. These were used to obtain initial item difficulty estimates as $b_{est} = b_{true} + SE$. Position effects were then generated by assigning each of the 200 items to a precalibration item position of 0 through 19, multiplied by the fixed slope $\gamma_1 = .024$. This term was used to obtain estimated item difficulties in the presence of a fixed position effect as $b_{pest} = b_{est} + .024 \times position$. Finally, true person ability was generated for 1,000 test takers from a normal distribution with $\mu = 0$ and

$\sigma = \sigma_{0p} = 2.06$, again based on results for M1 in Study 1. Note that signs were reversed for γ_0 and γ_1, as the CAT algorithm used here works in the traditional difficulty metric. Note also that this design situates true item difficulty at position 1, with position effects then always leading to increased difficulty. An alternative design, not explored here, could situate truth at the center of the test form, with items becoming easier when shifted earlier, and more difficult when shifted later in the form.

The item parameters in b_{est} and b_{pest} can be considered two separate item banks, where the latter only differs from the former due to position effect bias. CAT was conducted twice per true person ability value, once using b_{est}, a precalibrated item bank that only suffered from SE in the item difficulty estimates, and a second time using b_{pest}, a bank where item difficulty was additionally impacted by position.

CAT was also examined in a second simulation using the same item banks and estimators as described above, but with true person ability fixed for 1,000 persons at each of five theta values: −2, −1, 0, 1, 2. This allowed for an evaluation of bias at points along the theta scale, in contrast to the overall bias provided by the normal distribution of true theta in the initial CAT simulation. To differentiate between them, the simulations are labeled A for normally distributed true theta, and B for fixed true theta.

Analysis

CAT was conducted using functions from the catR package (Magis & Raîche, 2012), with default starting theta of 0 and item selection via maximal Fisher information. Test length was fixed at 20 items. In the CAT, person parameter estimates were obtained for all test takers using both b_{est} and then b_{pest}, with resulting values labeled θ_{est} and θ_{pest}. Four estimators were also examined: maximum likelihood (ML), weighted maximum likelihood (WML), expected a posteriori (EAP) with a normal prior, and maximum a posteriori (MAP) with a normal prior. Thus, the CAT was run eight times per person to obtain estimated theta under eight conditions per simulation (A and B).

Accuracy was evaluated first based on bias in person ability estimates. Bias was measured as the mean difference between estimated and true theta by condition:

$$bias = \frac{\sum_{p=1}^{N} \hat{\theta}_p - \theta_p}{N}, \tag{4}$$

where $\hat{\theta}_p$ represents the final estimated theta for a given person and condition and θ_p is the corresponding true ability. In simulation A, SE was summarized by condition by taking the root mean square of the SE values provided by the CAT engine for final ability estimates. Descriptive statistics and correlations between true and estimated theta were also examined for simulation A.

In simulation B, with 1,000 estimates of each fixed true ability θ, the root mean square error (RMSE) was obtained by condition as

$$RMSE = \sqrt{\frac{\sum_{p=1}^{N}(\hat{\theta}_p - \theta_p)^2}{N}}, \tag{5}$$

where p now represents a replication at a given theta and $N = 1{,}000$. Bias was obtained as in simulation A. The SE was found for simulation B as the standard deviation over ability estimates by condition. Note that this process for finding SE differs from simulation A, where it was based on the SE for each final ability estimate reported by the CAT engine.

Recall that Study 2 aims to represent a scenario in which position has an impact on item precalibration, resulting in biased precalibration estimates, but this impact is not accounted for in a subsequent CAT. Study 2 did not involve any estimation of position effects. Instead, all modeling within the CAT employed in simulations A and B was based on Equation 1, where the consequences of changing position are ignored, as may be done in practice. Test takers were simulated to respond to items according to the true, unbiased item parameters. However, items in b_{pest} were generated with hypothetical precalibration values that were biased based on position. Person abilities in θ_{pest} then came from a CAT algorithm that performed its item selection and estimated theta using these biased precalibration values.

It was expected that position bias in precalibration values would lead to increased bias in CAT person ability estimates θ_{pest}, compared with θ_{est}. However, with the differences in b_{pest} and b_{est} being systematic, SE was not expected to vary from θ_{pest} to θ_{est}, and RMSE, as an aggregate of bias and SE, was not expected to differ noticeably beyond changes due to bias. In the absence of differences, SE and RMSE would still provide context for evaluating the magnitude of bias introduced.

Results

Table 3 contains descriptive statistics for true and estimated theta by condition for simulation A, along with mean bias, SE, and correlations with true theta. For θ_{true}, distributional properties matched expectations based on the generating parameters, with the mean close to zero, and standard deviation (*SD*) just above 2. For θ_{est}, the mean for ML was .19, and the remaining means were closer to 0. *SD* for ML and WML were at or above 2, whereas EAP and MAP were closer to 1.5, a consequence of Bayesian shrinkage. For θ_{pest}, the mean for ML was again highest at .40, with remaining values all higher than their θ_{est} counterparts. *SD* for θ_{pest} roughly matched values for θ_{est} by estimator.

Mean bias was smaller by condition for θ_{est} than for θ_{pest}, as expected. Comparing differences from θ_{pest} and θ_{est} within estimator, additional mean bias introduced by position effects ranged from .22 for ML and MAP to .26 for WML. Overall, ML produced the most biased results, with means of .24 for θ_{est} and .46 for θ_{pest}. The smallest mean bias of .00 (after rounding) was found for θ_{est} and MAP. The smallest mean bias under θ_{pest} was .22 for MAP.

Table 3

Simulation A Results by Condition

		Mean	*SD*	Min	Max	Bias		*r*
θ_{true}	Estimator	−.05	2.13	−7.14	6.89	.00	SE	1.00
θ_{est}	ML	.19	2.23	−4.00	4.00	.24	.92	.94
	WML	.02	2.03	−4.00	3.28	.07	.73	.94
	EAP	−.03	1.49	−3.60	1.92	.03	.49	.94
	MAP	−.06	1.46	−4.00	1.83	.00	.48	.93
θ_{pest}	ML	.40	2.20	−4.00	4.00	.46	.84	.94
	WML	.28	2.08	−4.00	3.56	.33	.74	.94
	EAP	.22	1.48	−3.59	2.09	.27	.48	.93
	MAP	.16	1.47	−3.98	2.01	.22	.48	.93

Note. Mean, *SD*, Min, and Max are for the true and estimated theta distributions in each row, Bias and SE are the mean bias and standard error by condition, and *r* is the correlation with true theta.

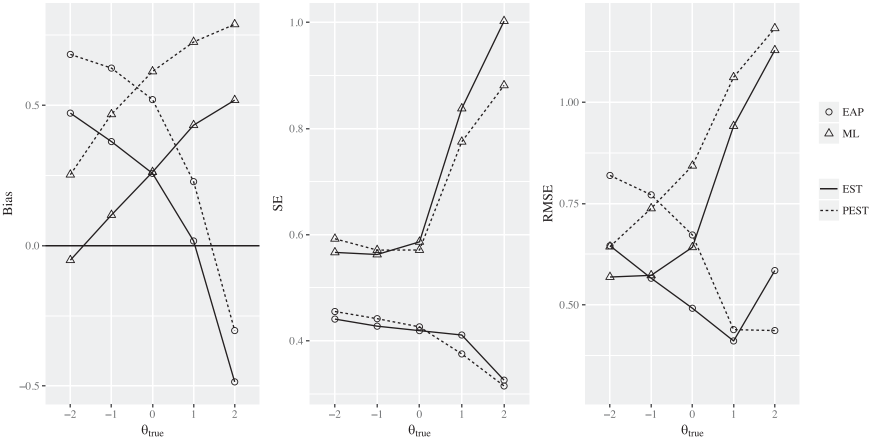

Figure 2. Error by condition for simulation B, with bias in the first plot, SE in the second, and RMSE in the third.

SE was relatively stable from θ_{est} to θ_{pest}, with the exception being for ML, which produced a mean SE of .92 in θ_{est} and .84 in θ_{pest}. SE were larger for ML and WML than for the Bayesian estimators in both theta conditions. Correlations with true theta ranged from .93 to .94, differing only slightly by condition.

Figure 2 shows the mean bias, SE, and RMSE by condition for simulation B, across fixed true theta values. In the interest of space, results for WML and MAP under θ_{est} and θ_{pest} are not shown as they were similar to results for ML and EAP. Overall, theta tended to be overestimated, as was the case in simulation A, both with and without position effects. For EAP, bias started positive at true theta −2, and decreased to negative bias by true theta 2. On the other hand, ML started with lower bias at −2 and increased for higher thetas. For all θ_{true}, bias was always more positive for θ_{pest} than θ_{est}, and was also larger in magnitude except at true theta 2 for EAP.

The largest difference between θ_{pest} and θ_{est} by condition was .36 for ML at true theta -1 and 0, and the smallest difference was .18 for EAP at true theta 2.

From Figure 2, the choice of estimator mattered the least at true theta 0, where the smallest discrepancies between mean bias were observed within θ_{est} and θ_{pest}. Theta 0 corresponds with the center of the person ability distribution. The differences for estimators then grow in the tails of true theta, with ML being least biased at true theta -2 and most biased by 2. The largest mean bias in magnitude was .786 for θ_{pest} and ML at true theta 2, whereas the smallest in magnitude was .02 for θ_{est} EAP at theta 1.

Trends for SE, as shown in the second plot in Figure 2, confirm that position effects had minimal influence on random error in the estimation of θ_{est} versus θ_{pest}. SE did differ noticeably by estimator, with larger values for ML, especially above true theta 0, and smaller values for EAP. Findings for SE matched those from simulation A. In the third plot in Figure 2, RMSE tended to be larger for θ_{pest} than θ_{est}. Exceptions were at true theta 1, where error for EAP in the presence of position effects was lower than for ML both with and without position effects, and at theta 2, where EAP with position effects produced the smallest error overall.

Discussion

This article explored item position effects using explanatory Rasch models, with the simple goal of clarifying impact on CAT person parameter estimation. In a real data study, item and person parameters were estimated as random effects, that is, as model residuals, and a fixed intercept captured average performance across students and items. Subsequent models then included a main effect for item position, and a random effect for position interacting with item. Models were fit with position coded both as the ordering of the item relative to adjacent items (0 to 19), and the ordering of the part or block for an item within the form (0 to 3). Models with the main effect for position were found to fit best, and were used to determine generating parameters in a simulation study.

Results from Study 1 indicate a linear decrease in the performance from the start to the end of the test form. Even with an extremely brief test, without time constraints and completed by most students within a few minutes, it appears that fatigue and disengagement influenced performance on items with later positions. This finding aligns with the findings of previous research on position effects in longer tests and with older populations (e.g., Albano, 2013; Alexandrowicz & Matschinger, 2008; Debeer & Janssen, 2013). One difference here, compared with other studies, is in the simplicity of the best-fitting model, where only a fixed linear slope for position was found to be statistically significant. Larger sample sizes would likely aid in the detection of significant interactions between position and item, and the exploration of nonlinear effects.

Study 2 explored the impact of a fixed linear position effect, hypothetically from a precalibration setting, on subsequent simulated CAT administrations. Accuracy in estimating person ability was measured via mean bias, over a random normal distribution of true ability in simulation A, and a series of fixed true ability in simulation B. Mean bias was compared for item banks containing versus not containing position effects, and then across four estimators. Overall, results confirm that a position effect in precalibration can introduce nonnegligible bias in CAT. In this case, a linear position effect of .024 on a 20-item test was associated with as much as .46 logits of mean bias in CAT person ability estimates. Mean bias introduced by position effects, beyond bias in the item bank without position effects, ranged between .22 and .26 across estimators in simulation A, and ranged between .18 and .36 across estimators and true theta values in simulation B. These values describe the expected increases in person parameter bias when item position effects are ignored in CAT.

While simulation A provided an overall summary of error resulting from position effects for a population of CAT test takers, simulation B provided more detailed results in terms of systematic and random error at specific locations across the theta scale. Results from simulation B indicated that bias for each estimator depended on true theta. At lower true theta values, the maximum likelihood estimators (ML and WML) produced smaller bias but larger SE than the Bayesian estimators (EAP and MAP), which conversely produced larger bias but smaller SE. These differences align with expectations, based on the objectives of the estimators themselves (see de Ayala, 2009). As true theta increased, the maximum likelihood estimators produced larger bias and SE, while the Bayesian estimators decreased in both, even trending to negative bias. These changes may have resulted from the reduced item coverage at the top of the scale, with the item bank centered at -1.85 with *SD* .65 and true person ability having a mean of 0 and *SD* 2.06.

Strong positive correlations between true and estimated ability indicate that the relative ordering of test takers changes minimally across conditions, both with and without position effects. This suggests that position bias could be adjusted for with a linear transformation of the theta scale, as noted by Doebler (2012). Adjustments may be less useful when person or item parameters are not distributed normally, or when position has more than a fixed linear effect. This highlights a limitation of the studies conducted here, that is, the simplicity of the operational test and subsequent CAT scenario. Testing programs may employ longer tests administered to larger sample sizes and with more complex CAT engines, which could produce different results for item and person parameter estimation. The conditions explored here are thus most relevant to formative and early educational assessments, where testing time and sample sizes may be limited.

On a related note, the present studies were also limited by the simplicity of the models used, that is, Rasch models with a fixed effect for position that was applied across items. Research has shown that position effects can be nonlinear, and can vary by item. Additional complexity also arises when incorporating item discrimination parameters. Impact for CAT will likely depend on these factors. The Rasch model was chosen here because the early childhood assessment system that motivated this research was designed using a Rasch framework (Wilson, 2005). Items had been developed and tested according to this paradigm. Operational constraints also led to a restricted sample size that would not support a comparison of IRT models with discrimination parameters.

It should also be noted that the CAT simulations here did not account for any fatigue or learning that might take place during the CAT itself. The origin of the position effects was limited to the hypothetical prior administration, and excluded the simulated one. It is expected that performance decreases in precalibration would also occur under similar conditions in CAT, with the impact of position effects then potentially being larger overall. Future research should explore compounding issues like these.

Given the practical focus of this article, magnitude is a key consideration in interpreting results. Indexed against the Study 1 person ability *SD* of 2.06, which became the population ability *SD* in Study 2, increases in mean bias of .18 to .36 logits constitute effect sizes of .09 to .17. Thus, the relative impact for mean bias in such a variable distribution of person parameters could be considered small. This finding, along with the strong correlations between estimated and true theta, suggests that bias from fixed linear position effects may not matter in CAT when score uses involve normative comparisons and rank ordering of test takers. Still, more complex models and test designs may lead to different results, especially if position effects are found to differ across people, whether because of changes in test administration conditions such as differential test lengths, or differences by demographic group or trait/ability level. Furthermore, even small amounts of bias can have meaningful consequences in criterion-referenced or mastery decisions, for example, when CAT is used to determine student proficiency and grades.

It is recommended that testing programs estimate and account for, as needed, the effects of changing item position across test administrations. Study 1 demonstrated a simple design that supports estimation, where item position is varied over forms that are randomly administered to test takers. In this linear test design, the number of forms determines the number of positions over which an item may appear. Fewer forms will lead to fewer options for position variation and thus a coarser picture of change in the performance by position, whereas more forms will provide more detail. Three forms are recommended as a minimum, with more being optimal, depending on test length.

With a linear test, item orders can also be scrambled, that is, randomly determined for each test taker. In this case, items can be expected to appear across all positions for a large sample of test takers. Alternatively, within an operational CAT, position effects can be obtained by administering pilot items with constraints that ensure they are seen across a range of positions, whether a fixed subset or all possible ones. As with precalibration in CAT, position effects should be estimated after sufficient sample sizes have been obtained.

Although this article focused on impact in CAT, these recommendations apply similarly to nonadaptive linking and equating designs employing multiple forms, especially when precalibration can lead to large differences in ordering for anchor items. In the event that significant shifts in item difficulty are apparent, test design should be reevaluated with attention given to item ordering and test length. In lower stakes testing programs, characteristic decreases in performance could be addressed by strategies to improve motivation and engagement. Finally, the calibration model, or the CAT engine itself, could also be augmented to account for position effects statistically.

Acknowledgments

This work was supported in part by grant R305A160034 from the Institute of Education Sciences, U.S. Department of Education.

References

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*, 408–426.

Alexandrowicz, R., & Matschinger, H. (2008). Estimation of item location effects by means of the generalized logistic regression model: A simulation study and an application. *Psychology Science Quarterly*, *50*(1), 64–74.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*, 565–580.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bradfield, T. A., Besner, A. C., Wackerle-Hollman, A. K., Albano, A. D., Rodriguez, M. C., & McConnell, S. R. (2014). Redefining individual growth and development indicators: Oral language. *Assessment for Effective Intervention*, *39*, 233–244.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test* (ETS Research Rep. No. rr-11-26). Princeton, NJ: Educational Testing Service.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, *39*, 502–523.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164–185.

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

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 369–377.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1–31.

Organization for Economic Cooperation and Development. (2009). *PISA 2006*. Technical Report. Paris, France: Author.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wyse, A. E., & Albano, A. D. (2015). Considering the use of general and modified assessment items in computerized adaptive testing. *Applied Measurement in Education*, *28*(2), 156–167.

Authors

ANTHONY D. ALBANO is Associate Professor of Educational Psychology at the University of Nebraska–Lincoln, 114 Teachers College Hall, Lincoln, NE 68588; albano@unl.edu. His primary research interests include equating, item analysis, assessment literacy, and classroom assessment.

LIUHAN CAI is Associate Psychometrician/Research Scientist at Measured Progress, Inc., 100 Education Way, Dover, NH 03820; cai.liuhan@measuredprogress.org. Her primary research interests include item parameter drift, equating, and multistage adaptive testing.

ERIN M. LEASE is Research Coordinator in Educational Psychology at the University of Minnesota, 56 E River Road, Minneapolis MN 55455; elease@umn.edu. Her primary research interests focus on increasing the use and expansion of assessments and interventions in early childhood.

SCOTT R. MCCONNELL is Professor of Educational Psychology at the University of Minnesota, 251 Education Sciences Building, Minneapolis MN 55455. His primary research interests focus on assessment and intervention for language and early literacy development among preschool children.