frontiers in Education

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/feduc.2020.00133

Contextual Interference Effects in Early Assessment: Evaluating the Psychometric Benefits of Item Interleaving

Anthony D. Albano[1]*, *Scott R. McConnell*[2], *Erin M. Lease*[2] and *Liuhan Cai*[3]

[1] School of Education, University of California, Davis, Davis, CA, United States, [2] College of Education and Human Development, University of Minnesota, Twin Cities, Minneapolis, MN, United States, [3] Cognia, Dover, NH, United States

Research has shown that the context of practice tasks can have a significant impact on learning, with long-term retention and transfer improving when tasks of different types are mixed by interleaving (abcabcabc) compared with grouping together in blocks (aaabbbccc). This study examines the influence of context via interleaving from a psychometric perspective, using educational assessments designed for early childhood. An alphabet knowledge measure consisting of four types of tasks (finding, orienting, selecting, and naming letters) was administered in two forms, one with items blocked by task, and the other with items interleaved and rotating from one task to the next by item. The interleaving of tasks, and thereby the varying of item context, had a negligible impact on mean performance, but led to stronger internal consistency reliability as well as improved item discrimination. Implications for test design and student engagement in educational measurement are discussed.

Keywords: contextual interference effect, interleaving, classroom assessment, early childhood, item analysis, reliability

OPEN ACCESS

Edited by:
Yong Luo,
Educational Testing Service,
United States

Reviewed by:
Ting Wang,
University of Missouri, United States
Junhui Liu,
Educational Testing Service,
United States
Xin Luo,
Uber, United States

****Correspondence:***
Anthony D. Albano
adalbano@ucdavis.edu

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: *24 March 2020*
Accepted: *02 July 2020*
Published: *07 August 2020*

Citation:
Albano AD, McConnell SR, Lease EM and Cai L (2020) Contextual Interference Effects in Early Assessment: Evaluating the Psychometric Benefits of Item Interleaving. Front. Educ. 5:133. doi: 10.3389/feduc.2020.00133

1. INTRODUCTION

Assessment in early educational settings is becoming increasingly common as researchers and practitioners work to ensure children are prepared for entry into kindergarten and transition into early primary grades. Intervention systems typically rely on assessment to screen and monitor children's progress in key outcomes, such as early literacy (e.g., McConnell and Greenwood, 2013), with the goal of providing targeted instructional support to children who are most in need (Greenwood et al., 2011; McConnell, 2018).

The growing emphasis on assessment in early childhood development coincides with a recent focus in the educational measurement community on applications of classroom assessment, that is, assessment used to inform decisions made at the classroom level. Although educational measurement has an important role in helping determine policy and programmatic changes at higher levels of education, it is argued that measurement in the classroom has the potential for the greatest impact on teachers and students (e.g., Wilson, 2018). As assessment practices expand in scope and utility in early education programs, they can both leverage and contribute to innovations in measurement.

In this paper, we respond to the call for collaboration on classroom assessment issues with an investigation of psychometric questions confronted when transitioning a set of early educational assessments from linear to computerized-adaptive testing (CAT) administration. The main question is, how does performance change from one child to the next as tasks vary in their ordering and context, as would often occur in CAT? Previous research has approached the question of context in assessment from two perspectives that appear to have developed in isolation from one another. The first reviewed here is a cognitive psychological perspective. The second is a psychometric one. In this paper, we aim to combine the two by evaluating the psychometric consequences in early assessment of what are referred to as contextual interference effects.

Cognitive psychological research has shown that classroom assessment activities are critical to effective teaching and learning, as they facilitate retrieval practice and formative feedback processes (Kang et al., 2007; Karpicke and Grimaldi, 2012). The testing effect, or test-enhanced learning, occurs when the assessment process itself serves as a form of practice (Roediger et al., 2011). When students retrieve information from memory in response to assessment items or tasks, the memory of that information is assumed to be strengthened, and feedback then offers the opportunity for correction and remediation. The gains from assessment practice have been shown to surpass those of other forms of study (e.g., Rowland, 2014).

Learning through practice assessment depends in part on the context in which practice tasks are presented, that is, their arrangement and relationships with one another in a series. Studies have compared two main strategies for arranging tasks: blocking involves repetition of each task before moving on to the next (e.g., aaabbbccc), whereas interleaving involves rotating through different tasks so that none appears more than once in sequence (e.g., abcabcabc). Results confirm that long-term retention and transfer improve when practice consists of tasks of different types that are mixed via interleaving rather than grouped together in blocks. Improvements associated with interleaving have been documented in a variety of areas, including music (e.g., clarinet performance, Carter and Grahn, 2016), sports (e.g., batting in baseball, Hall et al., 1994), and education (e.g., problem solving in mathematics, Rohrer et al., 2015).

It is hypothesized that the disruptiveness and inefficiencies of transitioning between tasks in interleaving are actually what make it more effective. *Contextual interference effects* (Battig, 1966) may enable closer comparisons among tasks simultaneously in working memory, while also inducing more frequent forgetting and reconstruction of familiarity with each task (Magill and Hall, 1990). In the end, interleaving is thought to require more effortful cognitive processing, leading to weaker short-term but stronger long-term gains (Brady, 1998).

In contrast to the cognitive psychological research, studies of context in educational measurement have been less concerned with benefits in terms of practice and learning, and more focused on the potential problems associated with changing item context, especially across multiple forms of an assessment (Leary and Dorans, 1985). In educational and psychological test design, the context of an item is determined by the features, content, and quantity of adjacent items within a test. Item context may initially be set based on a fixed ordering of items, but may vary in future administrations with the introduction of new items or the rearranging of existing ones.

Context effects in testing consist of changes in performance that are dependent on or influenced by the relationships between a given item and adjacent items in a test (Wainer and Kiely, 1987). When the context of an item changes, for example, when it appears at the start of one test administration and the end of another, test taker interactions with and responses to the item may change as a result. These changes can impact the psychometric properties of the item, such as its difficulty and discrimination, and can thereby lead to biased item parameter estimates (Leary and Dorans, 1985; Zwick, 1991), with implications for ability estimation and thus validity and fairness of score inferences.

Measurement research on this topic has focused mainly on context effects that result from shifts in item position, with the outcome of interest being variation in item difficulty. Study designs typically involve comparisons of two or more administrations of an item set, where some or all of the items change orders by administration (e.g., Davey and Lee, 2011). Variation in proportion correct or item response theory (IRT) item difficulty is then examined by administration (e.g., Mollenkopf, 1950). Findings have been mixed. In some cases, effects appear to be negligible (e.g., Li et al., 2012). In others, items have tended to become more difficult when administered at the end of a test compared with the beginning (e.g., Pomplun and Ritchie, 2004; Meyers et al., 2008; Albano, 2013; Debeer and Janssen, 2013; Albano et al., 2019). Studies have also identified items that decrease in difficulty by position (e.g., Kingston and Dorans, 1984).

Findings can be interpreted in a variety of ways. Decreases in performance, and corresponding increases in item difficulty, may result from participant fatigue and disengagement with longer or lower-stakes tests. Decreases may also be attributed to test speededness, where time constraints lead to lower focus, less careful responding, and guessing. On the other hand, increases in performance suggest that practice or learning is taking place over the course of the test. Here, it may be that challenging cognitive tasks or novel item types that are initially unfamiliar to test takers become easier as test takers warm up to them. Whatever their direction, non-negligible position effects pose a threat to assumptions of item parameter invariance over forms, and need to be evaluated in programs where item position may vary, for example, with CAT (Albano, 2013; Albano et al., 2019).

With prior studies focusing mainly on position, other aspects of item context have received limited attention, and their implications for test design are thus not well-understood. In CAT, for example, context may change for a given item based on transitions to new content (e.g., a different subtopic area) or different formats or layout (e.g., due to change in item type or cognitive task). In one administration, a given item may appear after others that assess similar content using a similar task format. In another administration, the same item may be preceded by items of differing content and task formats. According to the contextual interference research, changes, such as these may impact performance in meaningful ways.

This study stems from a larger project wherein we are exploring the use of CAT, compared with linear forms, with early educational assessments called Individual Growth and Development Indicators (IGDI; McConnell et al., 2015). IGDI are classroom measures designed to be brief and easily administered (typically requiring under 5 min and facilitated by a teacher), repeatable (so as to support progress monitoring), and predictive of socially meaningful longer-term developmental outcomes (according to principles of general outcome measurement; McConnell and Greenwood, 2013).

Like other general outcome measures (e.g., Fuchs and Deno, 1991), IGDI are used as indicators of achievement in relatively broad pre-academic or academic domains. Instead of being diagnostic for instructional planning, these measures are designed to identify children for whom additional intervention is needed and, once that intervention is in place, to monitor its effects on overall domain acquisition (Greenwood et al., 2011). Validity evidence supporting the use of IGDI in early intervention systems has been established (following Kane, 2013) through a formal measurement framework (Wilson, 2005) as well as models relating target constructs with criterion measures, such as ratings from teachers and future outcomes (e.g., Rodriguez, 2010; Bradfield et al., 2014; McConnell et al., 2015).

Current IGDI measures target four main areas of pre-reading development: oral language, phonological awareness, comprehension, and alphabet knowledge. Here, we examine alphabet knowledge (AK), the knowledge of names and sounds that are associated with printed letters (National Early Literacy Panel, 2008). AK is considered an essential component of the broader alphabetic principle and of models of early literacy (Whitehurst and Lonigan, 1998) and general reading competence (Scarborough, 1998). Instruction and measurement in AK typically focus on the total number of letter names and sounds known, as well as knowledge of letter writing, concepts of print, environmental print, and name familiarity (Justice et al., 2006; Piasta et al., 2010). AK measures may be administered as part of larger multi-domain assessments (e.g., Woodcock et al., 2001) or specifically targeting predecessors of and initial skills for early reading (e.g., Reid et al., 2001). Studies have demonstrated applications of IRT (e.g., Justice et al., 2006). However, research on CAT with AK, and with early childhood assessments more generally, is limited.

In this study, we extend the literature on classroom assessment development in a novel and practical way, by approaching the issue of item context in terms of contextual interference effects induced via interleaving. We collected and analyzed data with the main objective of exploring whether or not the target AK assessment could be transitioned from linear forms, where item context was fixed, to CAT, where context could vary. Consistency across forms in features, such as item difficulty, discrimination, and reliability would provide support for this transition.

The overarching research question is, to what extent do changes in item context, as captured by blocked and interleaved designs, produce differences in the psychometric properties of the assessment? Due to the lack of empirical psychometric research on this question, our expectations were primarily based on the contextual interference literature, which suggests that interleaving could lead to increased engagement and testing time, resulting from increased cognitive effort, but also higher item difficulty and lower performance overall, from a weakening of short-term practice effects. An increase in engagement may also lead to a decrease in measurement error, with positive results for associated statistical indices, such as discrimination and reliability.

TABLE 1 | Sample sizes and demographic counts (and percentages) by form.

	n	Female	Male	Non-white	White
Blocked	50	21 (42)	22 (44)	13 (26)	30 (60)
Interleaved	55	24 (44)	28 (51)	19 (35)	33 (60)
Combined	105	45 (43)	50 (48)	32 (30)	63 (60)

n is sample size. Percentages in parentheses do not sum to 100 because 10 children had missing demographic information.

2. MATERIALS AND METHODS

2.1. Data

Data for this study come from a federally funded project centering on the expansion of existing IGDI measures, developed with 4-years-old, to the assessment of language and early literacy development with 3-years-old, or children who are more than one academic year away from kindergarten enrollment. In a recent phase of the project, data were collected for item piloting, to inform the selection of item sets for scaling, scoring, and standard setting in later phases. Measures assessed children's oral language and phonological awareness, in addition to AK.

Participants were recruited from urban and suburban sites, including both public prekindergarten programs within local education agencies and private center-based child care sites. All children enrolled in the participating classrooms who met age criteria were invited to participate. Children were not excluded based on home language or disability status, unless teachers felt that either factor interfered with the child's ability to participate in the assessments. Sample sizes and demographics are presented in **Table 1** and discussed below.

2.2. Procedures

Assessments were administered by members of the research team, consisting of undergraduate and graduate students who were trained in administration protocols. Assessment items were preloaded on mobile tablets, which allowed for standardized administration and expedited scoring. Children were assessed in their classrooms or nearby in a hallway or adjacent room to minimize disruptions and distractions. The examiner viewed a mobile tablet that displayed the item prompt (e.g., "Point to the letter L"), the correct answer, and whether the child selected the correct response or not. The child's device was linked to the examiner's tablet via Bluetooth and displayed the corresponding item content. The examiner instructed the child to say or touch the correct response and the examiner verified accurate scoring on the device. Total administration time across measures was ∼15 min, including assessments in other domains

TABLE 2 | Form design with item and task by item position.

	Blocked		Interleaved	
Item position	**Item ID**	**Task**	**Item ID**	**Task**
1	i1	F	i2	F
2	i2	F	i8	O
3	i3	F	i14	S
4	i4	F	i16	N
5	i5	F	i3	F
6	i6	O	i9	O
7	i7	O	i11	S
8	i8	O	i17	N
9	i9	O	i5	F
10	i10	O	i7	O
11	i11	S	i15	S
12	i12	S	i20	N
13	i13	S	i4	F
14	i14	S	i6	O
15	i15	S	i13	S
16	i16	N	i19	N
17	i17	N	i1	F
18	i18	N	i10	O
19	i19	N	i12	S
20	i20	N	i18	N

Tasks are abbreviated as F, O, S, and N, representing letter finding, orienting, selecting, and naming, respectively. Item ID is a generic identifier for items with consistent meaning across forms.

of language and literacy development. AK was administered first and administration lasted ∼5 min per child.

The AK measure consisted of four tasks: letter find, letter orientation, letter selection, and letter naming. The first three of these tasks were selected-response or receptive tasks, whereas the fourth was a constructed-response or expressive task. In the letter find task, children were presented with an uppercase or lowercase letter and two distractors (e.g., a circle, ampersand, or arrow) and asked to point to the letter. In letter orientation, children were instructed to point to the letter facing the correct direction among two distractors that pictured the same letter but turned 90, 180, or 270°. The letter selection task prompted the child to point to the correct letter among two other letter options. Finally, in letter naming the child viewed a letter on their screen and was asked to name it.

Items were arranged by task into blocked and interleaved forms. **Table 2** compares the form designs. A generic item identifier (ID) is provided to clarify the shift in items from one form to the other. Tasks are abbreviated as F, O, S, and N, representing letter finding, orienting, selecting, and naming, respectively. Tasks are shown sequenced in groups for the blocked form, and in a rotation for interleaved, with item position being constant but the items themselves differing by form. For example, item i5, which assessed task F (finding), appears in position 5 on the blocked form and position 9 on the interleaved form.

Consented children were randomly assigned to complete one of the two forms. After removing incomplete data and outliers, described below, resulting sample sizes were 50 children for blocked and 55 for interleaved. **Table 1** shows the sample size by form, along with counts and percentages for gender and ethnicity.

2.3. Analysis

Prior to analysis, the data were prepared by checking for data entry errors and outliers. Two participants were removed, one due to a high number of invalid responses (17 out of 20, with the next highest being 6 of 20) and another due to an outlying total response time (over 9 min, where the distribution of total response time had a mean of 126 s with standard deviation 30 s).

Data analyses compared results across forms from three main perspectives, each one consisting of a set of analyses. The first set of analyses examined contextual interference effects in terms of participant engagement and attentiveness, which were measured indirectly using response time and long-string analysis (Johnson, 2005). Response time was simply the total duration of testing measured in seconds per test taker. Long-string analysis involved counting per test taker how many times they responded consecutively with the same choice. Shorter response times and higher counts for long-string analysis after controlling for the keyed correct choice suggest a test taker is responding inattentively (Wise and Kong, 2005; Meade and Craig, 2012). It was assumed that higher engagement and attentiveness resulting from contextual interference would be indicated by increased mean response time as well as lower mean long-string sequences for interleaving.

A chi-square test was first used to examine whether or not response choice (A, B, or C) on selected-response tasks was distributed independently of form. Statistical significance here would indicate a relationship between the two, where response choice differed overall when items were blocked vs. interleaved. Response time was then examined by item and across the test, with a t-test evaluating the mean response time difference by form. t-testing here and in subsequent mean comparisons assumed heterogeneity of variance with the Welch modification to degrees of freedom, and pooled variance was used to obtain standardized effect sizes.

The long-string analysis was applied only to the selected-response tasks. We first removed from each participant response string all choices that matched the key, that is, all correct responses. This step served to account for differences in keys by form. We then counted for the remaining responses the number of times a choice was repeated in sequence. For example, the string CCCBAA with key ABCABC would first reduce to CCBAA, and then produce counts of 2, 1, 2. We also found long-string counts without controlling for keyed responses. Having obtained these counts, we took the average by participant and then by form.

The next set of analyses included classical test theory (CTT) procedures for item analysis and internal consistency reliability analysis. Total scores were obtained for each student, with invalid responses coded as 0 and all items scored dichotomously. Performance was summarized by form and using combined data from both forms. Proportion correct (p-value) and corrected item-total correlations ($CITC$) were estimated for each item. Coefficient alpha was also obtained as an index of reliability.

The difference in reliability by form was tested for statistical significance using a bootstrap 95% confidence interval, obtained by sampling 1,000 times with replacement from each form data set, finding alpha by form and the difference, and then finding the quantile cutoffs corresponding to proportions of 0.025 and 0.975 in the resulting distribution of alpha differences. Similarly, bootstrap standard errors (*bse*) were estimated via the standard deviation of alpha estimates over samples by form. The difference in reliability by form was also tested using analytical standard errors (*ase*) and a corresponding 95% confidence interval for the alpha difference (see Duhachek and Iacobucci, 2004). In this case, the confidence interval is obtained as $\pm ase \times 1.96$.

The final set of analyses involved modeling within an item response theory (IRT) framework. Using a series of explanatory Rasch models, with item and person parameters treated as random effects (De Boeck, 2008), we examined differences in mean performance and item difficulty. A base model containing only item and person parameters was compared with models that included a main effect for form, estimating an overall mean difference as a fixed effect, and an interaction between form and item, estimating variability in item difficulty by form as another random item effect. In this way, form effects are examined as a type of differential item functioning (DIF), where blocked vs. interleaved enters the model as a categorical person grouping variable. DIF by gender (female/male) and ethnicity (non-white/white) was tested in the same way.

The base model can be expressed as

$$\eta_{pi} = \theta_p + \beta_i, \tag{1}$$

where η_{pi} is the log-odds of correct response for person p on item i, person ability is $\theta_p \sim N(0, \sigma^2_\theta)$, and item difficulty is $\beta_i \sim N(0, \sigma^2_\beta)$. Differential performance over groups is represented first by including an intercept δ_0 to capture mean performance in the reference group and δ_1 as the additive fixed effect for the focal group using indicator coding; and second by including an interaction between the grouping variable and random item effect. The interaction causes β_i to become β_{i0}, the item difficulty for the reference group, and β_{i1} is the additive effect on item difficulty for the focal group. The full model is then

$$\eta_{pi} = \delta_0 + \delta_1 group_p + \theta_p + \beta_{i0} + \beta_{i1} group_p, \tag{2}$$

with $\beta_i \sim MVN(0, \Sigma_\beta)$. Here, Σ_β contains $\sigma^2_{\beta_{i0}}$, $\sigma^2_{\beta_{i1}}$, and covariance $\sigma_{\beta_{i01}}$. DIF is indicated by the statistical significance of $\sigma^2_{\beta_{i1}}$.

Additional models also included fixed effect interactions between person grouping variables (gender by form and ethnicity by form) to test for differential performance by form (i.e., whether or not groups perform differently on blocked vs. interleaved). Models were estimated and compared using *lme4* (version 1.1-17; Bates et al., 2015) in R (version 3.4.4; R Core Team, 2018). Statistical significance was determined based on χ^2 difference tests from model fit comparisons.

3. RESULTS

3.1. Engagement and Attentiveness

The overall distribution of response choice was not found to differ by form ($\chi^2_2 = 0.63$, $p = 0.73$). For the long-string analysis, students taking the blocked form responded on average 1.49 times with the same choice in sequence. For the interleaved form, the mean was 1.55. Controlling for keyed responses, the adjusted means were 1.55 for blocked and 1.39 for interleaved. The difference of 0.16 was not statistically significant ($t_{87} = 1.42$, $p = 0.160$).

Table 3 contains means and standard deviations for total score distributions, as well as for response time (in seconds). Means were divided by 20 to represent average performance and response time at the item level. Mean total response time was found to differ by form, with blocked at 117.93 s and interleaved at 132.79 ($t_{95} = -2.67$, $p = 0.009$). The standard deviations for total response times were 22.73 for blocked and 33.65 for interleaved. The standardized effect size was $d = 0.51$, with pooled standard deviation 29 s.

3.2. Classical Test Theory

Mean total score performance (out of 20 items) was not found to differ by form, with means of 12.88 and 12.09 for blocked and interleaved, respectively ($t_{99} = 0.80$, $p = 0.43$). Coefficient alpha is also shown in **Table 3**, along with bootstrap and analytical standard errors. Alpha for both forms combined was 0.88. Alpha for the blocked form was 0.80 and for interleaved it was 0.91, with a difference of 0.11. Bootstrapping indicated that the difference was statistically significant ($\mathrm{CI}_{95\%} = [0.05, 0.21]$); however, analytical procedures did not ($\mathrm{CI}_{95\%} = [0.00, 0.22]$).

Figure 1 depicts the change in item difficulty (proportion correct, in the first plot) and discrimination (corrected-item total correlation, second plot) by form, with results from blocking on the x-axis and interleaving on the y-axis of each plot. Scatterplot points are represented by letters denoting the task for each item. A slight shift downward is evident in item difficulty, and a shift upward is evident for discrimination, for interleaved compared with blocked. p-values correlated at 0.81 across forms, and *CITC* correlated at 0.35.

3.3. Item Response Theory

Model comparison results are contained in **Table 4**. Results indicated that mean performance did not differ by form (row 2), or by form at the item level (row 3). Thus, the items were not found, overall, to differ in difficulty by form. Similarly, performance was not found to differ by gender (row 5), and interactions between gender and item (i.e., gender DIF, row 6) and gender and form (row 7) were not found to be statistically significant. Thus, there was no overall DIF by gender, or form effects by gender. There was a significant main effect for ethnicity (row 9), however, the ethnicity DIF model failed to converge (results not shown in **Table 4**) and the interaction between ethnicity and form was not found to improve model fit ($\chi^2 = 7.68$, $p = 0.0530$; although AIC decreased compared with the base model).

TABLE 3 | Descriptive statistics by form.

	Total score			Response time					
	M	***SD***	**M/20**	***M***	***SD***	**M/20**	**α**	***bse***	***ase***
Blocked	12.88	4.29	0.64	117.93	22.73	5.90	0.80	0.04	0.05
Interleaved	12.09	5.82	0.60	132.79	33.65	6.64	0.91	0.01	0.03
Combined	12.47	5.14	0.62	125.71	29.78	6.29	0.88	0.02	0.02

M and SD are the mean and standard deviation for the given variable across the test. M/20 is the mean divided by 20 to reflect averages at the item level. α is the reliability coefficient, and bse and ase are the bootstrap and analytical standard errors for alpha.

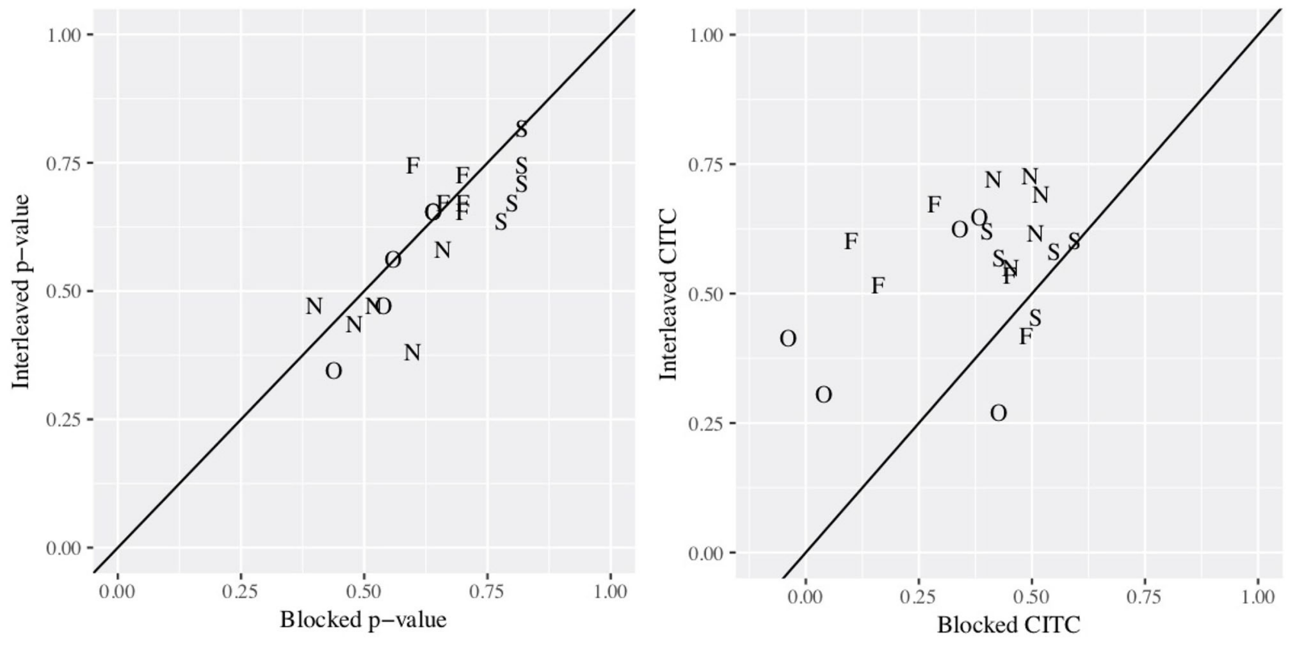

FIGURE 1 | Scatterplots comparing proportion correct (**left** plot) and discrimination **(right)** for each item by form (blocked on the *x*-axis, interleaved on *y*). Plotting characters represent the task for each item, abbreviated as F, O, S, and N (letter finding, orienting, selecting, and naming, respectively).

TABLE 4 | Model fit results.

Model	**Df**	**AIC**	**BIC**	**logLik**	**Deviance**	**χ^2**	***df***	***p***
1. Form base	3	2363.20	2380.18	−1178.60	2357.20			
2. Form main	4	2365.02	2387.66	−1178.51	2357.02	0.18	1	0.6736
3. Form × item	6	2367.17	2401.13	−1177.59	2355.17	2.03	3	0.5667
4. Gender base	3	2142.15	2158.83	−1068.07	2136.15			
5. Gender main	4	2144.12	2166.36	−1068.06	2136.12	0.03	1	0.8660
6. Gender × item	6	2146.24	2179.60	−1067.12	2134.24	1.91	3	0.5917
7. Gender × form	6	2147.44	2180.80	−1067.72	2135.44	0.71	3	0.8719
8. Ethnic base	3	2142.15	2158.83	−1068.07	2136.15			
9. Ethnic main	4	2136.97	2159.21	−1064.48	2128.97	7.18	1	0.0074
10. Ethnic × form	6	2140.46	2173.82	−1064.23	2128.46	7.68	3	0.0530

Each χ^2 comparison was between the model for a given row, e.g., 3. form × item, and the corresponding base model, e.g., 1. form base. df and p are the degrees of freedom and p-value for the χ^2 difference test.

4. DISCUSSION

The purpose of this study was to examine the effects of context, via blocked and interleaved arrangements of tasks, on the psychometric properties of an early educational assessment. The study was motivated by the need to evaluate a key assumption of CAT, item parameter invariance over changes in item context, in preparation for transitioning to CAT from a fixed linear assessment design. Overall, our findings support this transition, although sometimes in unexpected ways.

We chose to evaluate context in terms of changes in the ordering of tasks, partly in response to a lack of previous psychometric research on this issue. Studies in educational and psychological measurement have cautioned that changes in item context may be problematic for CAT, to the extent that item calibration results are not invariant over such changes (e.g., Albano, 2013). However, these studies have predominately represented context in terms of item position. Guidance is limited with respect to conceptualizing and understanding item context effects based on arrangement by task. As a result, we turned to research in cognitive psychology, which suggested interleaving could lead to contextual interference effects that decrease performance while at the same time increasing cognitive effort.

Findings from our study indicate that increased cognitive effort may have led to increased engagement and attentiveness, as evidenced by an increase in response time for interleaving compared with blocking. The long-string analysis did not uncover a difference in response patterns by form. It should be noted that the AK assessments studied here, like other IGDI, are carefully designed to be engaging for children, in terms of their manageable length (20 items, completed in under 5 min) and format (presented electronically, with children responding verbally or via touch screen). Thus, it may be that there is limited room for improvement when it comes to increasing engagement and attentiveness.

Results from the CTT analyses provide a descriptive comparison of item difficulty and discrimination by form. Items appeared slightly more difficult and slightly more discriminating with interleaving. These results should be interpreted with caution, as changes in individual item *p*-values and *CITC* were not tested for statistical significance. Furthermore, the sample sizes of 50 and 55 per form likely contributed to instability in item statistics. That said, a descriptive comparison is still useful here, as would take place in practice with pilot studies of item performance. At one extreme, two letter orientation items with *CITC* below and near zero in the blocked form increased to *CITC* of 0.41 and 0.31 with interleaving. Again, changes, such as these were not tested for statistical significance, but they do constitute increases in discrimination beyond typical minimum thresholds for inclusion in operational testing (e.g., 0.20; McConnell et al., 2015). All items surpassed discrimination 0.20 with interleaving, but only 16 of 20 met this cutoff in the blocked form.

CTT results also supported inferential comparisons in mean performance and reliability. Overall, mean performance was not found to differ by form, in contrast to our expectation of decreased performance for interleaving. Reliability, estimated with coefficient alpha, was significantly higher in the interleaved form, when tested using bootstrap confidence intervals. However, the analytic standard errors and confidence intervals were slightly larger than the bootstrapped ones, and did not support the conclusion of a significant difference in reliability. Though ultimately inconclusive, results for reliability are promising, indicating a need for further study of interleaving and the specific mechanisms through which it may lead to improved measurement.

Finally, IRT analyses did not uncover differential performance by form, gender, or ethnicity, within or across forms. Model comparisons were used as omnibus tests of DIF, where significant results would indicate that item difficulties differed overall across groupings of participants by form, gender, or ethnicity. Differential form effects were also examined, including by gender and ethnicity. An effect here would mean that groups performed differently on average when items were blocked vs. interleaved. None of the IRT results indicated significant differences that would preclude the use of forms with interleaved tasks.

This study contributes to the cognitive psychology literature by integrating measurement considerations into the comparison of blocked and interleaved arrangements of tasks. The study also contributes to the measurement literature by demonstrating an experimental method for examining the psychometric consequences of item context effects. Overall, results support the use of assessment designs wherein context may vary from one item to the next based on task. Findings suggest that increases in engagement and cognitive effort, associated in previous research with contextual interference effects introduced through interleaving, can enhance the psychometric properties of an instrument through a reduction in measurement error.

With regard to transitioning to CAT, it should be noted that interleaving represents an assessment design wherein the interference from changing context is expected to be maximized. In CAT, without controls for item content, test takers may never encounter a form where items are completely interleaved. Thus, this study tests an extreme case. If interleaving is associated with contextual interference effects that lead to increases in item discrimination and reliability, as suggested here, CAT administration would likely not achieve as strong results without constraints on content exposure. Still, the findings for interleaving are informative, including for fixed linear tests where item arrangement is a concern.

This study is limited in three main ways. First, as noted above, small sample sizes may have resulted in underpowered comparisons, with non-significance due to statistical error rather than a lack of effect. This applies both to the comparison of reliabilities by form, as well as the IRT analyses that employed model fit comparisons (see McCoach and Black, 2008). Future research should address this potential precision issue by employing larger sample sizes. Second, with blocking and interleaving each represented by only one item arrangement condition, changes in item position were not accounted for, which may have impacted results. The study design supported an overall comparison of task arrangement, however, it did not support a more detailed focus on item arrangement within task. Future research should also seek to separate these effects via more complex conditions. Third, with a focus on early educational assessment, results may not generalize to other subject areas beyond early literacy, or other populations beyond preschool children. Assessment in other situations, such as with older students and different types of tests, may lead to different findings, especially if these other situations involve increased test length and testing time, as well as changes in effort, motivation, and engagement. What constitutes a change in item context will also depend on the structure and subject area of the test. Here, context was defined based on task, where task represented a particular item format for assessing AK. In other settings, context may involve a more traditional distinction between item types (e.g., selected-response and short-answer), or it may capture a contrast between items of different sub-scales (e.g., addition and subtraction). Thus, future studies should explore the impact of interleaving, and possibly other arrangements in item context, using tests and assessments from other settings.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Minnesota Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SM and EL managed the data collection. AA and LC performed the statistical analysis. AA wrote the first draft of the manuscript. All authors contributed to the conception and design of the study, contributed to the manuscript revision, and approved the submitted version.

FUNDING

This work was supported in part by the Institute of Education Sciences, US Department of Education, through Grant R305A160034, Expanding Individual Growth and Development Indicators of Language and Early Literacy for Universal Screening in Multi-Tiered Systems of Support with 3-Years-Old to the University of Minnesota. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

REFERENCES

Albano, A. D. (2013). Multilevel modeling of item position effects. *J. Educ. Meas.* 50, 408–426. doi: 10.1111/jedm.12026

Albano, A. D., Cai, L., Lease, E. M., and McConnell, S. R. (2019). Computerized adaptive testing in early education: exploring the impact of item position effects on ability estimation. *J. Educ. Meas.* 56, 437–451. doi: 10.1111/jedm.12215

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Battig, W. F. (1966). "Facilitation and interference," in *Acquisition of Skill*, ed E. A. Bilodeau (New York, NY: Academic Press), 215–244.

Bradfield, T. A., Besner, A. C., Wackerle-Hollman, A. K., Albano, A. D., Rodriguez, M. C., and McConnell, S. R. (2014). Redefining individual growth and development indicators: oral language. *Assess. Effect. Interv.* 39, 233–244. doi: 10.1177/1534508413496837

Brady, F. (1998). A theoretical and empirical review of the contextual interference effect and the learning of motor skills. *Quest* 50, 266–293. doi: 10.1080/00336297.1998.10484285

Carter, C. E., and Grahn, J. A. (2016). Optimizing music learning: exploring how blocked and interleaved practice schedules affect advanced performance. *Front. Psychol.* 7:1251. doi: 10.3389/fpsyg.2016.01251

Davey, T., and 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: ETS.

De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73, 533–559. doi: 10.1007/s11336-008-9092-x

Debeer, D., and Janssen, R. (2013). Modeling item-position effects within an irt framework. *J. Educ. Meas.* 50, 164–185. doi: 10.1111/jedm.12009

Duhachek, A., and Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *J. Appl. Psychol.* 89, 792–808. doi: 10.1037/0021-9010.89.5.792

Fuchs, L. S., and Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exception. Child.* 57, 488–500. doi: 10.1177/001440299105700603

Greenwood, C. R., Carta, J. J., and McConnell, S. R. (2011). Advances in measurement for universal screening and individual progress monitoring of young children. *J. Early Interv.* 33, 254–267. doi: 10.1177/1053815111428467

Hall, K. G., Domingues, D. A., and Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Percept. Motor Skills* 78, 835–841. doi: 10.2466/pms.1994.78.3.835

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Pers.* 39, 103–129. doi: 10.1016/j.jrp.2004.09.009

Justice, L. M., Bowles, R. P., and Skibbe, L. E. (2006). Measuring preschool attainment of print-concept knowledge: a study of typical and at-risk 3-to 5-year-old children using item response theory. *Lang. Speech Hearing Serv. Schools* 37, 224–235. doi: 10.1044/0161-1461(2006/024)

Kane, M. T. (2013). Validation as a pragmatic scientific activity. *J. Educ. Meas.* 50, 115–122. doi: 10.1111/jedm.12007

Kang, S. H. K., McDermott, K. B., and Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.* 19, 528–558. doi: 10.1080/09541440601056620

Karpicke, J. D., and Grimaldi, P. J. (2012). Retrieval-based learning: a perspective for enhancing meaningful learning. *Educ. Psychol. Rev.* 24, 401–418. doi: 10.1007/s10648-012-9202-2

Kingston, N. M., and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Appl. Psychol. Meas.* 8, 147–154. doi: 10.1177/014662168400800202

Leary, L. F., and Dorans, N. J. (1985). Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Rev. Educ. Res.* 55, 387–413. doi: 10.3102/00346543055003387

Li, F., Cohen, A., and Shen, L. (2012). Investigating the effect of item position in computer-based tests. *J. Educ. Meas.* 49, 362–379. doi: 10.1111/j.1745-3984.2012.00181.x

Magill, R. A., and Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Hum. Mov. Sci.* 9, 241–289. doi: 10.1016/0167-9457(90)90005-X

McCoach, D. B., and Black, A. C. (2008). "Evaluation of model fit and adequacy," in *Multilevel Modeling of Educational Data*, eds A. A. O'Connell and D. B. McCoach (Charlotte, NC: Information Age Publishing, Inc.), 245–272.

McConnell, S. R. (2018). "The path forward for multi-tiered systems of support in early education," in *Multi-Tiered Systems of Support for Young Children: A Guide to Response to Intervention in Early Education*, eds J. J. Carta and R. Miller (Baltitmore, MD: Paul H Brookes Publishing), 253–268.

McConnell, S. R., and Greenwood, C. R. (2013). "General outcome measures in early childhood and the individual growth and development indicators," in *Handbook of Response to Intervention in Early Childhood*, eds V. Buysse and E. Peisner-Feinberg (Baltimore, MD: Paul H. Brookes Publishing), 143–154.

McConnell, S. R., Wackerle-Hollman, A. K., Roloff, T. A., and Rodriguez, M. C. (2015). Designing a measurement framework for response to intervention in early childhood programs. *J. Early Interv.* 36, 263–280. doi: 10.1177/1053815115578559

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 1–19. doi: 10.1037/a0028085

Meyers, J. L., Miller, G. E., and Way, W. D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Appl. Meas. Educ.* 22, 38–60. doi: 10.1080/08957340802558342

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika* 15, 291–315. doi: 10.1007/BF02289044

National Early Literacy, P.anel, (2008). *Developing Early Literacy: Report of the National Early Literacy Panel–A Scientific Synthesis of Early Literacy Development and Implications for Intervention.* Jessup, MD: National Institute for Literacy.

Piasta, S. B., Purpura, D. J., and Wagner, R. K. (2010). Fostering alphabet knowledge development: a comparison of two instructional approaches. *Read. Writ.* 23, 607–626. doi: 10.1007/s11145-009-9174-x

Pomplun, M., and Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *J. Educ. Comput. Res.* 30, 243–254. doi: 10.2190/Y4FU-45V7-74UN-HW4T

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Reid, D. K., Hresko, W. P., and Hammill, D. D. (2001). *Test of Early Reading Ability.* Austin, TX: Pro-ed.

Rodriguez, M. C. (2010). *Building a Validity Framework for Second-Generation IGDIs.* Technical report, Center for Response to Intervention in Early Childhood, University of Minnesota, Minneapolis, MN, United States.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., and McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *J. Exp. Psychol. Appl.* 17, 382–395. doi: 10.1037/a0026252

Rohrer, D., Dedrick, R. F., and Stershic, S. (2015). Interleaved practice improves mathematics learning. *J. Educ. Psychol.* 107, 1–9. doi: 10.1037/edu0000001

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140, 1–32. doi: 10.1037/a0037559

Scarborough, H. S. (1998). "Early identification of children at risk for reading disabilities: phonological awareness and other promising predictors," in *Specific Reading Disability: A View of the Spectrum*, eds B. K. Shapiro, P. J. Accardo, and A. J. Capute (Timonium, MD: York Press), 77–121.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Meas.* 24, 185–201. doi: 10.1111/j.1745-3984.1987.tb00274.x

Whitehurst, G. J., and Lonigan, C. J. (1998). Child development and emergent literacy. *Child Dev.* 69, 848–872. doi: 10.1111/j.1467-8624.1998.tb06247.x

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2018). Making measurement important for education: the crucial role of classroom assessment. *Educ. Meas. Issues Pract.* 37, 5–20. doi: 10.1111/emip.12188

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Woodcock, R. W., McGrew, K. S., Mather, N., and Schrank, F. (2001). *Woodcock-Johnson R III NU Tests of Achievement.* Itasca, IL: Riverside.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educ. Meas. Issues Pract.* 10, 10–16. doi: 10.1111/j.1745-3992.1991.tb00198.x

Conflict of Interest: SM and colleagues developed the assessment tools and related resources known as Individual Growth & Development Indicators. This intellectual property has been licensed to Renaissance Learning for commercial development and sale. SM and colleagues, along with the University of Minnesota, have equity and/or royalty interests in Renaissance Learning. These relationships are reviewed and managed by the University of Minnesota in accordance with its conflict of interest policies.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.