Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Linear-on-the-Fly Testing (LOFT): Which Design to Choose?

S. Embretson [*1] A. J. Thierbach [2] and J. L. Walsh [3]
1 Department of Educational Psychology, University of Kansas, Lawrence, Kansas, USA
2 Air and Space Technology Sector, DCS Corporation, Dayton, Ohio
3 Air and Space Technology Sector, DCS Corporation, Dayton, Ohio
*Corresponding author. Email: sembret@ku.edu

Abstract

While advances in computerized adaptive testing (CAT) have addressed many of the drawbacks associated with traditional linear fixed-length high-stakes tests, the high costs of developing a large item bank often is a barrier for some organizations. An effective compromise between the limitations of traditional testing designs and the expense of CAT is the linear-on-the-fly test (LOFT) design. LOFT is often applied to a set of items developed for two or more parallel forms. LOFT permits dynamic, real-time selection of items or groups of items from the set, facilitating the creation of multiple test forms with specified test lengths. Trait estimates in LOFT are based on pre-calibrated item response theory parameters for the specific items selected for each examinee. There are various strategies for designing a LOFT. The most extreme approach involves random selection of items from the full set of items for each examinee, which can lead to extreme differences in item parameters. A more controlled design utilizes item bins along with a technique called "dipping." In this method, items are categorized into bins arranged by difficulty, and a computer algorithm randomly selects one item from each bin, resulting in more consistent and comparable test forms. Despite these innovations, there is a notable lack of literature on the impact of varying LOFT designs. The present study conducted a Monte Carlo simulation examining measurement precision and bias of several LOFT designs, as well as parallel forms. The results from the study have important implications for the selection of the LOFT design.

Keywords: linear-on-the-fly testing, LOFT, computerized adaptive testing, CAT, test forms

1. Introduction

Traditional fixed-length tests, particularly in high-stakes contexts, have several notable disadvantages. Because such tests typically rely on a limited number of alternative forms (often only two or three) item exposure is likely to occur. Increased item exposure can lead to pre-knowledge among some examinees, thereby threatening test validity and necessitating the development of entirely new test forms. Computerized adaptive testing (CAT) substantially reduces item exposure through individualized item selection for each examinee and offers a practical mechanism for addressing exposure when it does occur. Specifically, the removal of a small number of compromised items generally has minimal impact on the functioning of the overall item bank. As a result, CAT effectively mitigates one of the major limitations of traditional testing approaches. However, despite these advantages, the substantial costs associated with developing and maintaining the large item banks required for CAT often present a significant barrier for many organizations.

Linear-on-the-fly testing (LOFT) is an effective compromise between the limitations of traditional testing designs and the expenses of CAT, as it can be applied to the set of items developed for two or more parallel forms. LOFT involves random selection from the set, facilitating the creation of multiple test forms with specified test lengths. LOFT also can involve real-time selection of items or groups of items for each examinee using random procedures. Similar to CAT, trait estimates in LOFT are based on pre-calibrated item response theory (IRT) parameters for the specific items selected. Also, like CAT, the dynamic selection of items in LOFT

leads to an enormous number of varying tests. For example, a fully random LOFT selection of 5 items from a set of 20 items can lead to 1,860,480 (20x19x18x17x16) parallel forms of the tests!

Several strategies may be employed in the design of a LOFT. In the most extreme approach, items are selected entirely at random from the item bank for each examinee. Although this design can generate millions of distinct test forms and thereby enhance test security, it may also result in substantial variability in test difficulty across examinees and uneven coverage of content domains. To mitigate these issues, LOFT implementations often incorporate constraints on item selection. The primary objective of such strategies is to ensure adequate representation of item content and difficulty levels within each test form. One common approach involves grouping items into *bins,* that is, collections of items sharing specified characteristics, and randomly selecting a predetermined number of bins for each examinee. These bins can be constructed in various ways, such as by difficulty level, to reduce differences in overall test difficulty across examinees. Another bin-based approach further constrains item selection by randomly selecting one item from each bin as the test progresses. In this design, bins may be aligned with test positions, with each bin containing items of similar difficulty and content. This method promotes greater comparability across test forms while retaining the flexibility and security advantages of LOFT. Although LOFT has been available for a long time (see Becker & Bergstrom, 2013; Luecht, 2016), published studies comparing the impact of the various strategies on trait estimates could not be found. The current Monte Carlo simulation study evaluated the impact of various LOFT design strategies on trait estimates. These strategies were compared to equivalent forms and the item bank for the LOFT strategies were constructed from the equivalent form items. The study aimed to provide practical recommendations for practitioners while also expanding our understanding of LOFT design.

2. Methods

A Monte Carlo simulation was conducted to compare true traits (θ) against trait estimates ($\hat{\theta}$ *[theta hat]*) obtained under various LOFT conditions. Each simulee was administered a 20-item test drawn from a pool of 60 items across four distinct LOFT conditions.

.

2. 1 Simulated data

True trait levels (θ) and item difficulties were defined to generate item response data. Random selection of 1,000 trait levels from a normal distribution, $\theta \sim N(0,1)$, defined true trait levels, while random selection of 60 items from a normal distribution, $\beta \sim N(0,1)$, defined true item difficulties. Item response probabilities were computed for each item using the Item Response Theory (IRT) Rasch model. Thus, the probability of a correct response for person $j$on item i was obtained as follows:

$$P(X_{ij} = 1 \mid \theta_j, \beta_i^{\text{true}}) = \frac{\exp\,(\theta_j - \beta_i^{\text{true}})}{1 + \exp\,(\theta_j - \beta_i^{\text{true}})} \qquad (1).$$

To highlight comparisons between the conditions, stochastic error was not included in generating item responses. Responses were scored as one (X_{ij} = 1) for items with probabilities greater than .50 and as zero (X_{ij} = 0) otherwise. For test forms to represent varying LOFT conditions, 20 item responses were selected from the pool of 60 items for each condition. Due to computational constraints for the full study, only 20 replications were conducted.

2. 2 Conditions

The five test forms of 20 items each were constructed for simulees according to the following five conditions: one non-LOFT condition, referred to as Linear Forms, and four LOFT conditions, referred to as Random Forms, Unbalanced Bins, Balanced Bins and Dipping Bins. The first two conditions represent extremes for comparison. In the Linear Forms condition, three traditional static forms were created and matched on item difficulty. These forms remained fixed and were randomly assigned to simulees within a replication. In contrast, the Random Forms condition involved randomly selecting 20 items from the full pool of 60 items for each simulee, resulting in a very large number of unique test forms.

For the other three conditions, items were grouped into bins. Two of these conditions involved random selection of bins for each simulee. That is, the Balanced Bins condition represents a LOFT strategy to assure comparability of item difficulties across examinees. In this condition, 12 bins (5 items each) were constructed. The bins were balanced on item difficulty. Four out of 12 bins were randomly selected for each simulee. This produced many different test forms (i.e., 12*11*10*9). The Unbalanced Bins condition also consisted of randomly selecting 4 bins from 12, but in this case each bin contained similar item difficulties. This produced the same number of test forms as Balanced Bins, but the scope of item difficulty could vary substantially between simulees. The final condition, Dipping Bins, consisted of 20 bins, each of which had 3 items of similar difficultly. One of the 3 items was randomly selected from each of the 20 bins for each simulee. This condition also led to many possible test forms (i.e., 320 forms), but with more control of item difficulty.

2.3 Estimation

Maximum likelihood estimates for trait levels ($\hat{\theta}$) and their standard errors of measurement (SE_{msmt}) were obtained using the R package **mirt** (see Chalmers, 2012). These estimates were obtained across 20 replications of all five study conditions with a sample sizes of 1,000.

2. 4 Comparisons

We compared means (M) and standard deviations (SD) for trait estimates ($\hat{\theta}$), as well as for SE_{msmt}, bias and root mean square errors (RMSE, according to the formula below),

$$\text{RMSE} = \sqrt{\sum_{j=1}^{J} \left(\hat{\theta}j - \theta j\right)^2} \qquad (2).$$

Reliabilities were also estimated using internal consistency coefficients and correlations of estimated thetas with true thetas. The conditions were compared using an analysis of variance (ANOVA), with Linear Forms serving as the reference (baseline) for contrasts.

2. Results

Table 1 displays a summary of the results across 20 replications for all five study conditions.

Table 1. Descriptive Statistics for Trait Estimates and Reliability Across Replications

	Trait Estimates		SE_{msmt}		RMSE		Bias		Reliability	
Condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	r_{IC}	$r^2_{true,est}$
Linear Forms	.000	.885	.468	.028	.375	.282	.004	.469	.719	.779
Dipping Bins	.000	.883	.468	.028	.375	.286	.004	.471	.719	.777
Balanced Bins	.000	.853	.459	.026	.375	.286	.004	.471	.710	.778
Unbalanced Bins	.000	.899	.459	.034	.447	.338	.004	.559	.738	.690
Random Forms	.000	.734	.423	.024	.401	.308	.004	.505	.667	.763
Total	.000	.851	.455	.028	.395	.300	.004	.495	.711	.757

First, we conducted an ANOVA on the summary trait estimates ($\hat{\theta}$) in the replications to examine whether there were significant differences across conditions in terms of the M and SD. The results indicated that the estimated trait-level means did not differ significantly across conditions, $F(4, 95) = .82$, $p = .51$ in the 20

replications. In contrast, the SD's differed significantly across conditions, F(4, 95) = 112.06, p < .001. The Random Forms condition exhibited the lowest SD as compared to the other measurement error conditions.

Next, we conducted an ANOVA comparing the means and SD's of SE_{msmt} for trait estimates across conditions. Significant differences were found for conditions on both the means of SE_{msmt}, F(4, 95) = 249.44, p < .001, and the SD's of SE_{msmt}, F(4, 95) = 31.38, p < .001. As shown in Table 1, the means of SE_{msmt} were lowest in the Random Forms condition, which is consistent with the increased frequency of trait estimates toward the center of the distribution. The means of SE_{msmt} were somewhat higher for the Linear Forms and Dipping Bins conditions. Finally, the SD of SE_{msmt} was relatively higher for the Unbalanced Bins condition, reflecting the increased likelihood of selecting a large number of items with extreme difficulty levels.

We conducted an ANOVA on the RMSE, which reflects both systematic and random error. Significant differences across conditions were observed for both the *M* of RMSE, $F(4, 95) = 127.31$, $p < .001$, and the *SD* of RMSE, $F(4, 95) = 128.67$, $p < .001$. As shown in Table 1, the highest means and *SD* of RMSE were observed in the Unbalanced Bins condition, followed by the Random Forms condition, indicating lower accuracy in trait estimation.

Bias was also compared across conditions, although the means were equal, the SD's differed significantly, F(4, 95) = 150.37, p < .001. The SD's were larger for both Unbalanced Bins and Random Forms.

Additionally, we examined reliability. Internal consistency, r_{IC} ,differed significantly across conditions, $F(4, 95) = 65.35$, $p < .001$, with the Random Forms condition exhibiting substantially lower reliability. The correlations between the true trait level (θ) and the estimated trait level ($\hat{\theta}$), r^2_t , also differed significantly, $F(4, 95) = 163.51$, $p < .001$, with the Unbalanced Bins condition showing the lowest correlation.

To further investigate the RMSE and SE_{msmt} mean differences among the conditions, we conducted independent-samples t-tests comparing the Linear Forms condition against the other four LOFT conditions. Table 2 presents the results of the t-tests for RMSE and SE_{msmt}, and a total number of significant effects across 20 replications. RMSE did not differ significantly from Linear Forms for Dipping Bins and Balanced Bins conditions in any replication. However, RMSE differences were found for the other two conditions. Specifically, the Unbalanced Bins condition had significantly larger RMSE than Linear Forms in all 20 replications, while Random Forms condition had a significantly larger RMSE in 13 replications.

Regarding SE_{msmt}, significant results were obtained for most replications. Note that the t-values for SE_{msmt}, were converted into absolute values because the replications varied as to whether the average SE_{msmt} was higher or lower than Linear Forms. Clearly, the mean difference of .001 for Dipping Bins suggests that the differences in SE_{msmt} balance out across replications. For Balanced Bins and Unbalanced Bins, the SE_{msmt} mean was consistently higher than Linear Forms, but the differences are very small. Finally, for Random Forms, the SE_{msmt} mean was significantly lower than Linear Forms in all 20 replications. While this may seem inconsistent, it will be discussed below.

Table 2. Contrasts of Equivalent Forms Means with Conditions Across Replications.

Comparison: Linear Forms Vs.	Mean Difference	t-value	Probability	Replications p<.05
Root Mean Square Error				
Dipping Bins	.000	.023	.531	0
Balanced Bins	-.000	-.027	.592	0
Unbalanced Bins	-.072	-5.612	.000	20
Random Forms	-.026	-2.330	.139	13
SE Measurement				
Dipping Bins*	.001	5.843	.026	18
Balanced Bins	.009	13.602	.046	19
Unbalanced Bins	.009	9.967	.040	19
Random Forms	.045	68.688	.000	20

*t-values varied from positive to negative

To visually examine differences among conditions, plots of estimated $\hat{\theta}$ versus true θ were generated. In all plots, points represent simulated cases across 20 replications with 1,000 simulees each, and solid lines indicate curvilinear regressions fitted to the data (see Figure 1). In the Linear Forms condition, some discrepancies were observed at the extremes of the trait distribution, with both very high and very low true θ values tending to be underestimated. In contrast, the Random Forms condition exhibited substantially larger discrepancies at extreme trait levels, with relatively few estimated $\hat{\theta}$ values exceeding 2.00 or falling below -2.00. The Unbalanced Bins condition produced some extreme estimates; however, $\hat{\theta}$ were more broadly dispersed around the true θ, resulting in lower correlations between estimated and true trait values.

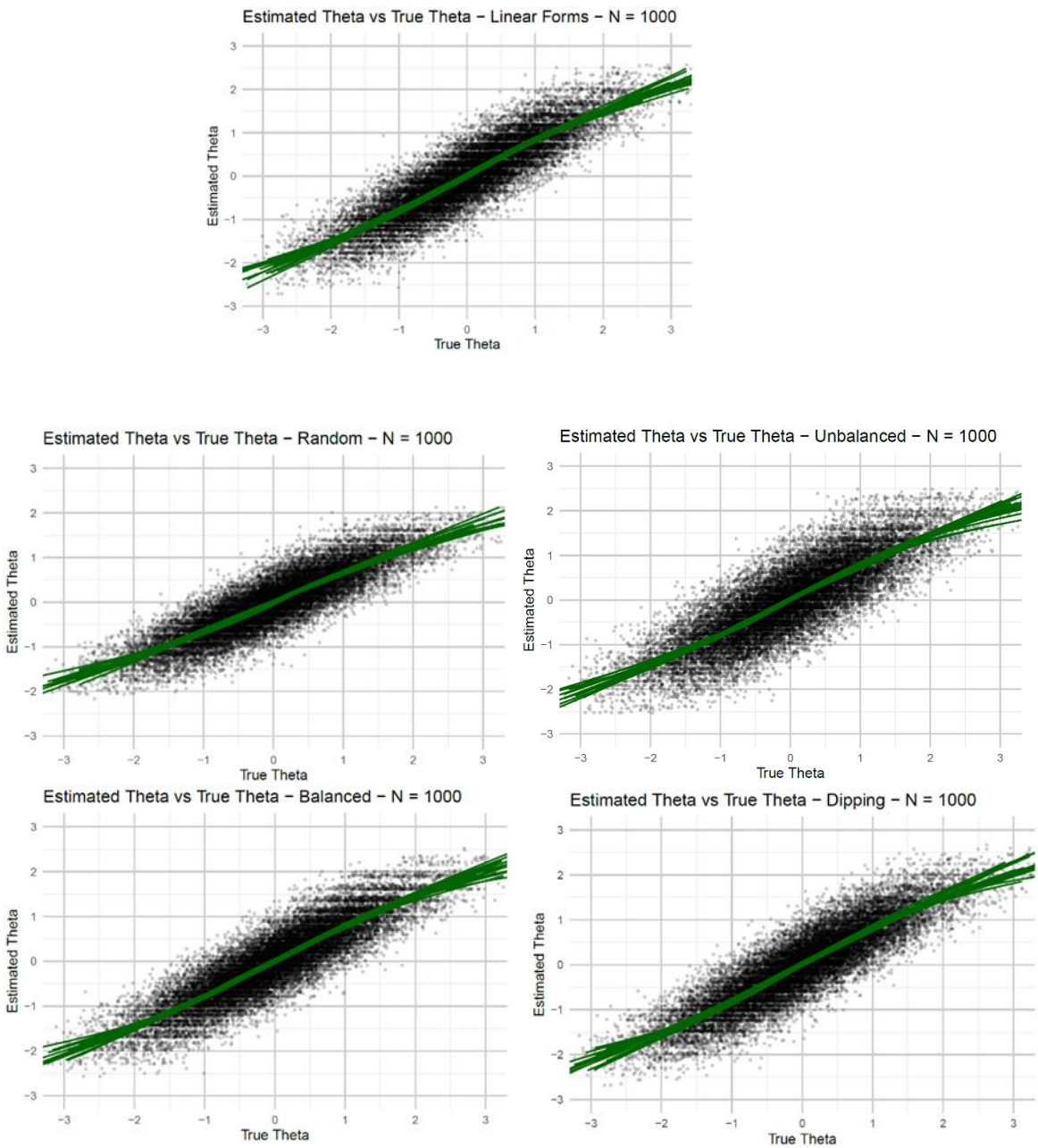

Figure 1. Estimated vs. True Theta

4. Conclusion

The conclusions derived from the Monte Carlo simulation are twofold. First, for the Dipping Bins and Balanced Bins designs, estimation accuracy (as indicated by RMSE) did not differ significantly from that of Linear Forms across replications. Moreover, the similarity of trait means, standard deviations, standard errors of measurement (SE_{msmt}), and reliability estimates supports the viability of these designs as alternatives to Linear Forms.

Second, trait estimates obtained from the Random Forms and Unbalanced Bins designs differed from those produced by Linear Forms in several important respects. Specifically, Random Forms yielded higher RMSE values and lower internal consistency reliability than Equivalent Forms. Although Random Forms exhibited lower SE_{msmt} values, which might superficially suggest improved measurement precision, this effect appears to be attributable to a reduced standard deviation of the trait estimates. This reduction resulted in a greater concentration of estimates near the center of the distribution, where SE_{msmt} is inherently lower. Taken together, the higher RMSE, lower reliability, and reduced representation of extreme trait values indicate comparatively poorer measurement performance for the Random Forms design. Trait estimates from the Unbalanced Bins design, by contrast, exhibited SE_{msmt} values comparable to those of Linear Forms, but demonstrated higher RMSE and weaker correlations with the true θ values.

These findings suggest that when practitioners are choosing between Linear Forms and either Random Forms or Unbalanced Bins LOFT designs, Linear Forms should be preferred. In contrast, when the decision is between Linear Forms and the Balanced or Dipping Bins LOFT designs, pragmatic considerations should guide selection. Such considerations include whether item content or taxonomic structure must be carefully balanced across bins, in which case the Dipping Bins design may be most appropriate. If content and taxonomy balancing are less critical and efficiency or resource constraints are prioritized, the Balanced Bins design may be preferable, as it effectively consists of multiple mini parallel forms matched on difficulty. Overall, this study supports both Dipping Bins and Balanced Bins as viable alternatives to Linear Forms.

Acknowledgement

We gratefully acknowledge the support and sponsorship of our research by the Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX), and the Air Force Research Laboratory (AFRL) 711 Human Performance Wing. Their continued partnership and investment in advancing scientific knowledge have been instrumental to the success of this work. We sincerely thank our AFPC/DSYX, particularly Dr. John Trent, and AFRL customers, particularly Dr. Thomas Carretta, for their guidance, collaboration, and commitment to innovation.

The views and opinions presented herein are those of the author(s) and do not necessarily represent the views of DoD or its Components. Appearance of, or reference to, any companies, commercial products or services does not constitute DoD endorsement of those companies, products or services

Funding Statement This research received no additional external support to AFPC and AFRL.

Competing Interests None.

References

Becker, K. A. & Bergstrom, B. A. (2013) "Test Administration Models", *Practical Assessment, Research, and Evaluation. 18*(1)., 179-205.

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

Luecht, R. M. (2016). Computer-based test delivery models, data, and operational implementation issues. *Technology and testing: Improving educational and psychological measurement*, 179-206.