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

Comparing fit indices for predicting difficulty of automatically generated items

Haoyang Yu [*†], Susan Embretson [†]
†Department of Educational Psychology, University of Kansas, Lawrence, 66045-3101, KS, USA
*Corresponding author. Email: Haoyang.yu@ku.edu

Abstract
Automatic item generation (AIG) aims to reduce development time and cost in large-scale assessments by generating structurally related item families. Operational use is especially benefited by AIG if the ability to predict item parameters—especially item difficulty—is strong enough to reduce reliance on extensive pretesting. Two approaches that can be used to evaluate prediction accuracy: (a) regression-based R^2, a limited-information index based on regressing calibrated Rasch difficulties on item features, and (b) Embretson's (1983, 1997) delta statistic (Δ^2), a full-information likelihood-based index comparing the Linear Logistic Test Model (LLTM) to the saturated Rasch model. This study uses a Monte Carlo simulation to compare how well Δ^2 and regression R^2 recover known levels of predictability under varied test lengths, sample sizes, and numbers of predictors. Across 90 design conditions with 1,000 replications each, Δ^2 consistently exhibited lower root-mean-square error (RMSE), bias, and mean absolute error (MAE), with the greatest advantage for short tests and small samples—conditions typical in AIG contexts. These results support the use of Δ^2 as a more accurate and robust index of predictability when evaluating explanatory item response theory models for item-feature specifications.

Keywords: LLTM, Delta Index, Item Difficulty Modelling

1. Introduction

Large-scale assessments require continual development of new, psychometrically sound items. Automatic item generation (AIG) has emerged as a scalable solution, enabling the creation of item families based on cognitive or structural features. Although AIG can reduce item production costs, further improvement would be obtained if these item features can reliably explain or predict item difficulty without extensive empirical pretesting.

In AIG and explanatory IRT applications, item predictors are typically operationalized in two broad ways. First, predictors may encode design or generating-model structure, such as indicators identifying items that originate from the same generating model or template family (e.g., Embretson & Kingston, 2018). Second, predictors may reflect content or complexity features hypothesized to drive difficulty—for example, linguistic variables for verbal items (e.g., word count, syntactic complexity, vocabulary level; Embretson & Daniel, 2010) or visual/figural complexity for nonverbal items (e.g., Embretson, 2023). These predictor specifications can be incorporated either as covariates in a limited-information regression framework or as item-feature effects within an LLTM estimated using full-information maximum likelihood.

Two analytic frameworks are commonly used to evaluate predictability. The first is regression-based predictability in which calibrated Rasch item difficulties are regressed on item features. The coefficient of

determination, R^2, is used as a summary index. Although simple and familiar, regression R^2 is a limited-information index that depends on the quality of the estimated item difficulties and may be unstable when the number of items is modest relative to the number of predictors.

The second approach involves explanatory item response theory (IRT) modeling, operationalized through the Linear Logistic Test Model (LLTM). LLTM models item difficulty as a linear function of item features within a Rasch framework and can be estimated using full-information maximum likelihood. Embretson's (1997) delta statistic, Δ^2, first proposed in 1983, compares the LLTM to the saturated Rasch model and quantifies the proportion of explainable deviance that is captured by the LLTM feature structure.

Although both regression R^2 and Δ^2 serve similar goals, their comparative performance under realistic AIG conditions has not been systematically evaluated. In particular, it is not clear how their accuracy depends on test length, sample size, or the number of predictors when the true level of predictability is known. This study addresses that gap using a Monte Carlo simulation designed to mirror typical AIG applications.

The purpose of this study is to compare the performance of a limited-information regression-based index and a full-information LLTM delta index in recovering specified levels of predictability in a simulation study. We examine how the indices behave as a function of the number of items, the number of examinees, the true R^2, and the number of item-feature predictors. Accuracy is evaluated using root-mean-square error (RMSE), bias, and mean absolute error (MAE) with respect to the specified true R^2.

2. Method

2. 1 Overview

We conducted a fully crossed Monte Carlo simulation. For each condition defined by the crossed factors, 1,000 independent data sets were generated under a Rasch model with item difficulties determined by a linear combination of simulated item features plus normally distributed error. For each data set, we fitted the following models:

- Rasch model to recover item difficulty estimates
- LLTM target model using the same features used to generate the data
- LLTM null model with a common difficulty parameter
- Multiple regression model of recovered Rasch difficulties on the features

The Rasch model and LLTM log likelihoods for fit were used to compute the delta index. The regression model provided R^2. These two indices were then compared to the true R^2value used to generate the data in that condition.

All IRT models were estimated using the **mirt** package in R with a Rasch item type and unidimensional latent trait. All simulations were implemented using parallelized loops over replications.

2. 2 Simulation Design

Four factors were manipulated in a fully crossed design:

- Test Length: Number of items per form, $I \in \{20, 25, 30, 40, 50\}$,
- Sample Size: Number of examinees. $J \in \{250, 500, 1000\}$,
- True Explained Variance: $R^2_{\text{true}} \in \{0.25, 0.36, 0.49\}$,
- Number Predictors: $k \in \{2, 4\}$.

This yields $5 \times 3 \times 3 \times 2 = 90$ distinct simulation conditions. For each condition, $N = 1000$ replications were generated and analyzed.

2. 3 Generation of Item Features and True Difficulties

For each condition and replication, an item feature (design) matrix Q of dimension $I \times k$ was simulated:

$$Q_{ik} \sim \mathcal{N}(0,1), \text{ independently for } i = 1, \dots, I, k = 1, \dots, K.$$

A vector of baseline feature weights $\eta^{(0)}$was specified for $k = 2$ and $k = 4$ and then scaled to achieve the desired true R^2. Specifically, for two predictors, baseline weights were $\eta^{(0)} = (0.40,\ 0.30\ )^{\top}$. For four predictors, baseline weights were $\eta^{(0)} = (0.37,\ 0.27,\ 0.18,\ 0.09)^{\top}$. For a given condition with target R^2_{true}, the weights were scaled as $\eta = \sqrt{\frac{R^2_{\text{true}}}{0.25}}\ \eta^{(0)}$.

Let $\sigma^2_{\epsilon} = 1 - R^2_{\text{true}}$. The true item difficulties were generated as

$$b_i^{\text{true}} = Q_i.\eta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\epsilon}),$$

with ϵ_i independent across items. This construction implies that the linear predictor $Q\eta$ explains R^2_{true} of the variance in item difficulty by design.

2. 4 Generation of Examinee Abilities and Responses

For each replication, latent abilities for Jexaminees were simulated from a standard normal distribution. Under the Rasch model, the probability of a correct response for person $j$on item i is

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

Binary responses were then generated via Bernoulli trials: $X_{ij} \sim \text{Bernoulli}[P(X_{ij} = 1 \mid \theta_j,\ b_i^{\text{true}})]$, yielding a $J \times I$ response matrix, X, for each replication.

2. 5 Model Fitting and Fit Indices

2.5.1 Rasch and LLTM Models

For each replication, a unidimensional Rasch model was fit to X using **mirt**, yielding item difficulty estimates $\hat{b}_i$, $i = 1, \dots, I$, and log-likelihood $\ln L_{\text{Rasch}}$. The Rasch model is treated as the saturated model in the delta index framework.

Two LLTM models were then fit:

1. Null LLTM (intercept-only): All items share a common difficulty parameter. This corresponds to a design matrix with a single constant column. The null LLTM provides log-likelihood $\ln L_{\text{null}}$.
2. Target LLTM (feature-based): Item difficulties are modeled as a linear combination of the k features in Q. For item $i = 1, \dots, I$, $b_i = \sum_{m=1}^{K} q_{im}\ \eta_m$, where q_{im} is the value of predictor m for item i (entry in the Q-matrix), η_m is the weight (difficulty contribution) of predictor m, and K is the number of predictors. The target LLTM provides log-likelihood $\ln L_{\text{target}}$.

2.5.2 Delta Fit Index

The delta index Δ^2 was computed as the ratio of deviance reduction achieved by the target LLTM relative to the saturated Rasch model:

$$\Delta^2 = \frac{\ln L_{\text{null}} - \ln L_{\text{target}}}{\ln L_{\text{null}} - \ln L_{\text{Rasch}}}.$$

By construction, Δ^2 values range from 0 (target model no better than the null) to 1 (target model as good as the Rasch model). In this simulation, Δ^2 is intended to recover the known proportion of variance in the item difficulties explained by the features, namely the true R^2.

2.5.3 Regression-Based R^2

Limited-information regression models were fit by regressing the estimated Rasch item difficulties $\hat{b}_i$ on the feature matrix Q for each replication. The usual coefficient of determination R^2 from the linear model was extracted and treated as a competing fit index for the quality of the item-feature model. This regression R^2 should approximate the true R^2 to the extent that the Rasch difficulty estimates are accurate and the linear model holds.

2. 6 Accuracy Metrics

For each condition, and for each index type (Delta vs Regression), let Est_i denote the value of the index from replication $i \in \{1, \dots, N\}$, and let True denote the condition's target R^2_{true}. We summarize performance of each index using the following three metrics:

- Root-Mean-Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Est_i - True)^2}$$

- Bias

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{Est}_i - \mathrm{True})$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid \mathrm{Est}_i - \mathrm{True} \mid$$

metrics quantify overall accuracy (RMSE), systematic over- or underestimation (Bias), and typical absolute error magnitude (MAE).

3. Results

Results are summarized at the condition level by averaging RMSE, Bias, and MAE across the 1,000 replications for each combination of items, examinees, true R^2, and number of predictors, separately for Δ^2 and regression R^2. In Figures 1–4, Δ^2 estimates are denoted by triangle markers and regression R^2 estimates by circle markers; solid lines indicate k = 2 predictors and dashed lines indicate k = 4 predictors.

3. 1 Overall Accuracy

Across all conditions, the LLTM delta index (Δ^2; triangles) showed better recovery of true predictability than regression R^2 (circles). Figure 1 displays mean RMSE averaged across Test Length and Sample Size for each true R^2 value. RMSE decreased as true predictability increased for both indices, but Δ^2 remained lower (or at least comparable) across the full range of true R^2. The advantage of Δ^2 was most apparent at the lower and moderate predictability levels (true R^2 = .25 and .36), with the largest discrepancies occurring in the higher-complexity (k = 4; dashed) conditions.

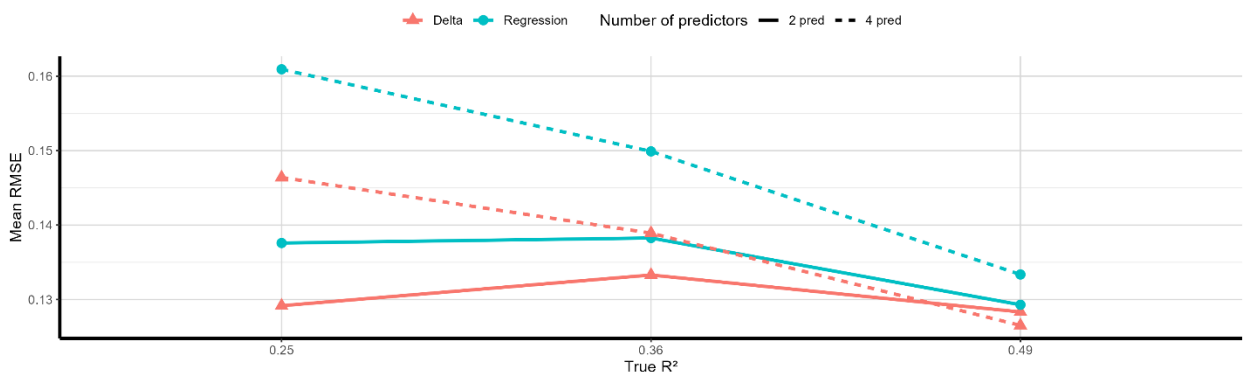

Figure 1. Overall mean RMSE across Test Length and Sample Size.

3. 2 Effects of Test Length and Sample Size

Figure 2 decomposes RMSE jointly by Test Length and Sample Size. The advantage of Δ^2 is most evident for shorter tests (20–30 items) and smaller samples (250 examinees), which reflect common AIG calibration regimes with limited data. In these lower-information conditions, Δ^2 produces lower RMSE than regression R^2 across true R^2 levels, indicating tighter recovery around the generating-model predictability.

As the number of items increases to 40 or 50 and sample size increases to 1000, RMSE values decrease for both indices and the gap between Δ^2 and regression R^2 narrows. Under these more informative designs, regression R^2 approaches Δ^2, but does not show a consistent advantage.

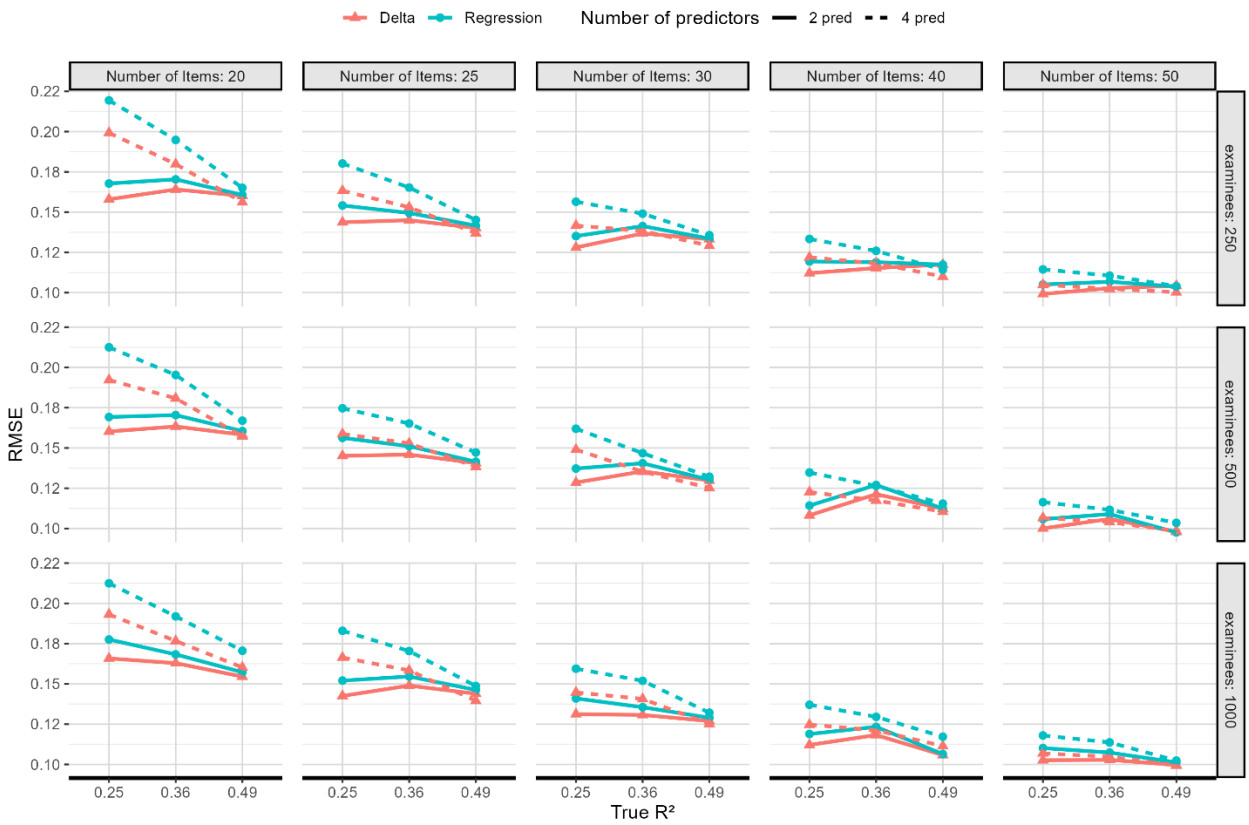

Figure 2. RMSE by Test Length and Sample Size.

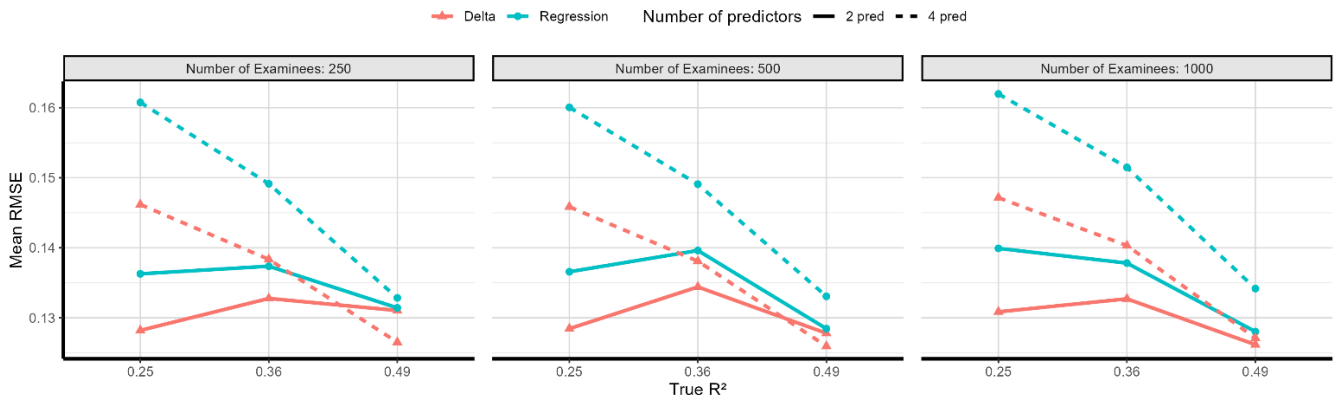

Figure 3. RMSE by Sample Size.

3. 3 Influence of the Number of Predictors

Figures 3 and 4 summarize how estimation accuracy changes with model complexity (k = 2 vs. k = 4). Figure 3 shows RMSE by sample size (averaged over test length), and Figure 4 shows RMSE as a function of test length (averaged over sample size). Across both summaries, increasing the number of predictors from two to four has a larger adverse effect on regression R^2 than on Δ^2, especially when information is limited. For short forms, the k = 4 regression results (dashed circles) show noticeably higher RMSE, consistent with increased sensitivity to noise in recovered Rasch difficulty estimates. In contrast, Δ^2 remains comparatively stable across k, indicating robustness of the full-information likelihood index to added predictors under these conditions.

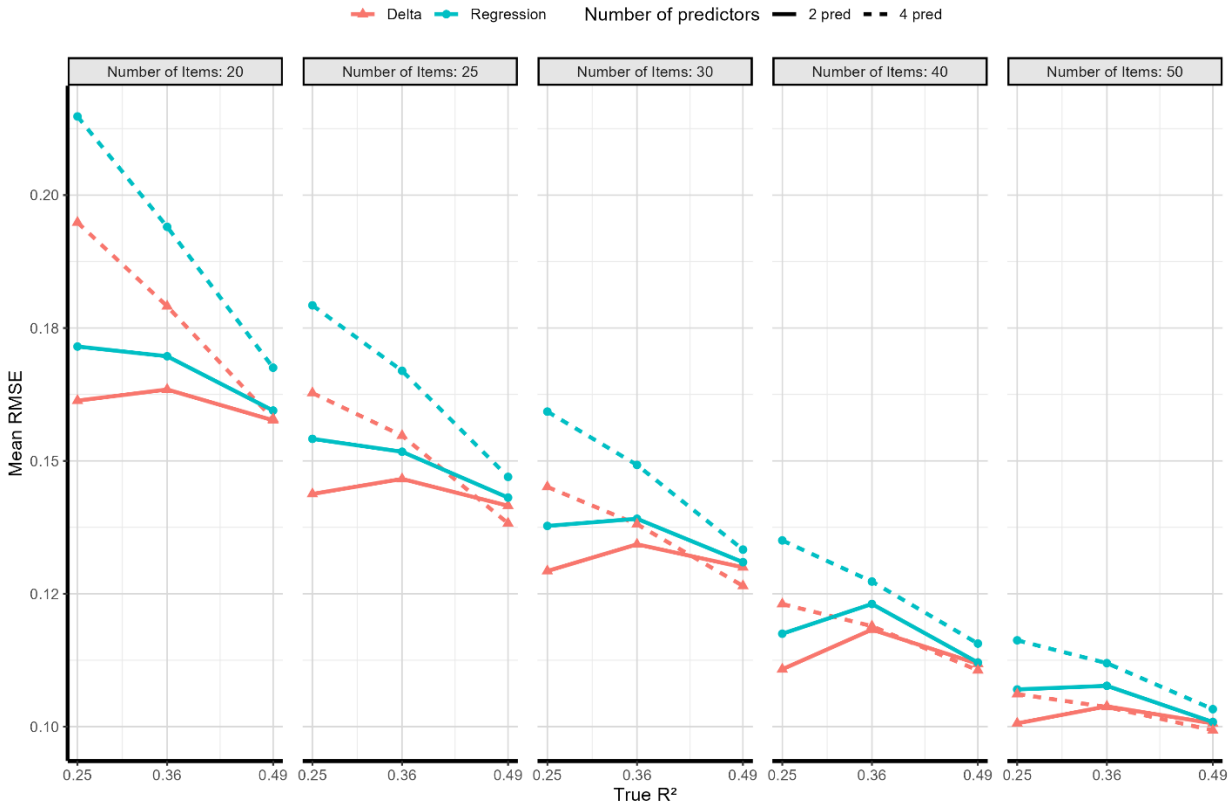

Figure 4. RMSE averaged over Sample Size as a function of Test Length.

3. 4 Bias Pattern

Bias analyses revealed systematic differences between the two indices. Figure 5 summarizes mean bias as a function of Test Length. Across all test lengths, regression R^2 exhibited larger positive bias, indicating a tendency to overestimate true predictability when items are few. The magnitude of this overestimation was largest for 20–25 item tests, whereas Δ^2 showed smaller bias. As Test Length increased to 40–50 items, bias decreased for both indices, but the advantage of Δ^2 persisted.

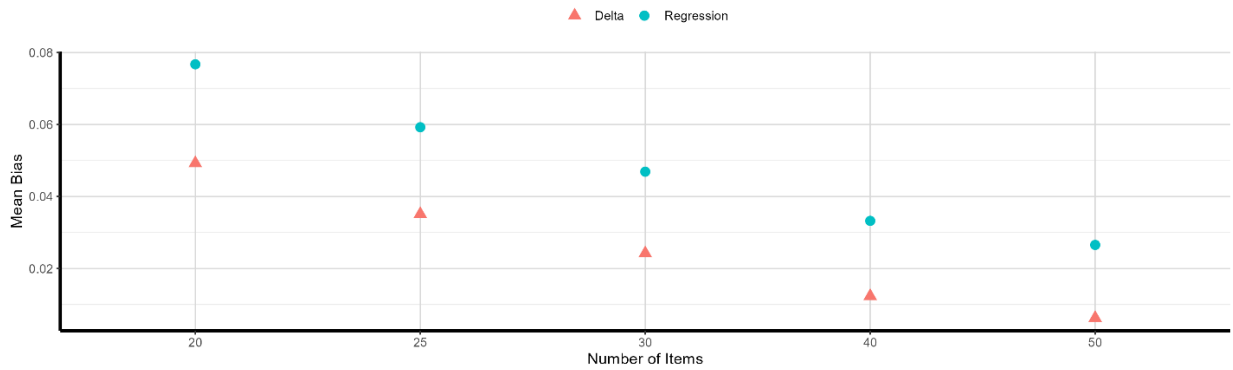

Figure 5. Average Bias by Test Length

Mean absolute error (MAE) results reinforce the same conclusion. Figure 6 shows that MAE decreases steadily as test length increases, with Δ^2 yielding lower MAE than regression R^2 at each test length. Figure 7 displays MAE aggregated by sample size; Δ^2 remains relatively stable across J = 250–1000, whereas regression R^2 is consistently higher and shows little improvement with increased sample size. Together, the bias and MAE results indicate that Δ^2 provides more accurate and less variable recovery of true R^2, particularly for short tests and noisier difficulty estimates.

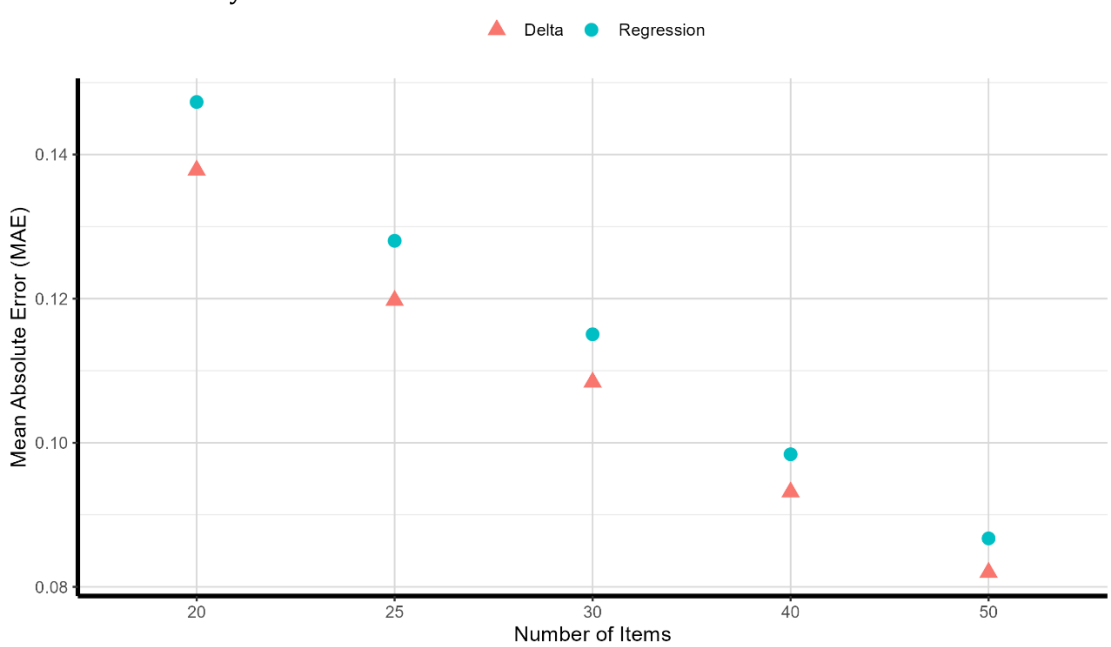

Figure 6. Mean Absolute Error by Test Length

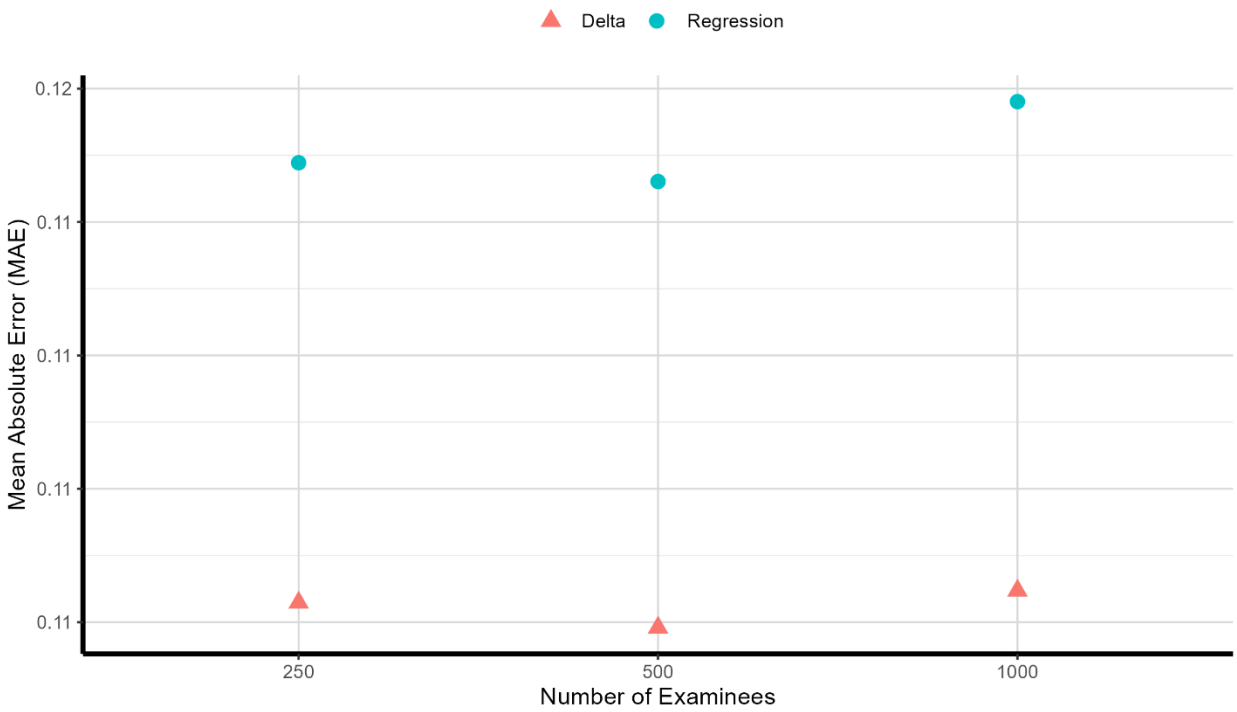

Figure 7. Mean Absolute Error by Number of Examinees

4. Discussion and Conclusion

This study compared a limited-information regression-based predictability index (regression R^2) with a full-information LLTM delta index (Δ^2) under a wide range of AIG-relevant conditions. Across nearly all conditions, Δ^2 provided more accurate and more stable recovery of the generating-model predictability than regression R^2, as evidenced by lower RMSE, smaller bias, and smaller MAE.

The results highlight the advantages of using full-information indices when evaluating explanatory IRT models for item-feature specifications, particularly in AIG contexts where calibration data may be limited and test forms may be short. Under these lower-information regimes—and especially when the number of predictors is larger relative to the number of items—regression R^2 was more variable and showed a stronger tendency to overestimate true predictability, whereas Δ^2 remained comparatively robust. As test length and sample size increased, the two indices converged, but regression R^2 did not show a consistent advantage over Δ^2.

For practitioners implementing AIG and explanatory IRT, these findings support using Δ^2 as a primary index of predictability when assessing item-feature specifications, with regression R^2 serving as a complementary summary rather than the sole criterion for evaluating model adequacy.

Acknowledgement

Competing Interests: None

References

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Embretson, S. E. (1983, June). An incremental fit index for the linear logistic latent trait model. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 305–321). Springer.

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

Embretson, S. E., & Kingston, N. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement, 55*(1), 112–131.

Embretson, S. E. (2023). Understanding examinees' item responses through cognitive modeling of response accuracy and response times. *Large-scale Assessments in Education, 11*(9).

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.