

Automated Generation and Scoring of Maze Reading Comprehension Assessments

Hatice Kubra Karakis¹, Walter Leite¹, Logan Scott¹, Xinyi Tai¹, and Akihito Kamata²

¹University of Florida ²Southern Methodist University

hkarakis@ufl.edu, Walter.Leite@coe.ufl.edu, loganscott1@ufl.edu,

xinyitai@ufl.edu, akamata@smu.edu

Abstract

This work proposes automated generation and psychometric scoring of Maze comprehension assessments using large language models (LLMs) and a multilevel item response theory (multilevel IRT) framework. Findings show reliable ability estimates across passages and items, offering scalable, curriculum-aligned formative assessment that reduces teacher workload and supports targeted reading instruction.

1 Introduction

Maze comprehension tests have a long history of use in elementary school education. It was popularized as part of the Curriculum-Based Measurement (CBM) (Deno, 1985) framework, which is a standardized testing process that has been widely used in schools to evaluate student achievement. CBM testing in reading can include a maze comprehension test and an oral reading fluency test. Maze is also part of the DIBELS assessment, which is widely used. Traditional paper-and-pencil implementations of the Maze tests, such as those in CBM administration and DIBELS, are standardized reading passages where every 7th word has been deleted and replaced by three choices. These choices are one word that is the correct one for the passage and two distractors.

There have been many studies providing validity evidence for different versions of standardized Maze tests. For example, Shin and McMaster (2019) conducted a meta-analysis showing that the correlation between the CBM Maze and state tests of reading comprehension was 0.6. However, the use of standardized passages are of limited use for formative assessment of reading comprehension, because they are challenging to produce. A teacher can take any passage from a reading curriculum and generate a Maze comprehension test manually, but it is time-consuming. An automated Maze generator provides an opportunity to reduce

teacher burden in creating these assessments. Existing Maze passage generators (Intervention Central) randomly draw distractors from a list of common English words or randomly sample words from the passage. This simple approach misses the opportunity to make Maze tests into formative assessments that simultaneously inform the teacher about the student's current level of reading comprehension and provide an opportunity for the student to practice specific decoding skills.

Large language models allow selection of distractors with fine-tuned characteristics that enable practice in addition of assessment. For example, distractors can be selected to form word chains with the correct answer, engaging students in a more complex reading activity. This approach has similarities with the concept of stealth assessment (Rahimi and Shute, 2024), because it combines practice with measurement. Furthermore, generative AI allows the automated generation of Maze tests that adapt to each lesson of a reading curriculum. For example, distractors can be selected containing the specific phonetic patterns a teacher is working on with the students during the week that the Maze is administered. Having a Maze that aligns with lessons of a reading curriculum would allow more precise evaluation of student reading comprehension, and would provide an opportunity to reinforce concepts taught in the classroom.

In the current study, we propose multiple methods to automatically create Maze tests that allow them to serve as opportunities for formative assessment and practice while reducing preparation burden on teachers. In addition, we propose a psychometric model based on the concept of explanatory item response theory to provide student ability estimates accounting for random variation in word difficulty and passage characteristics.

2 Related Literature

2.1 Automated Item Generation

Automated Item Generation (AIG) has become increasingly important, as assessments require large pools of high-quality items. Early AIG research focused on structured item models that specify content, cognitive demands, and constraints to guide systematic item creation (Gierl and Haldyna, 2012; Gierl et al., 2012). These rule-based approaches showed that item variants could be generated efficiently while maintaining the alignment of the construct. With the advances in computational methods, AIG has incorporated machine-learning and natural-language approaches, generating more linguistically diverse and contextually coherent items (Circi et al., 2023; Attali et al., 2022). Recent studies indicate that transformer-based models can generate reading passages, stems, and distractors that are more fluent and contextually appropriate, but quality varies based on the prompt and model (Sommer and Arendasy, 2025). For example, Bulut and Yildirim-Erbasli (2022) demonstrated that fine-tuned transformer models such as GPT-2 and T5 could automatically generate both reading passages and comprehension items for elementary students. However, the generated content still required human review before classroom use (Bulut and Yildirim-Erbasli, 2022).

Large language models (LLMs) have further advanced AIG by enabling learning from only a small number of examples, maintaining a consistent writing style, and flexibly adapting to new item formats (Brown et al., 2020). Research on LLM-generated educational materials shows that these models can produce coherent passages and reading comprehension items, though challenges such as semantic drift, inconsistent difficulty, and distractor quality remain concerns (Xiao et al., 2023; Tan et al., 2025). Linguistic analyses of LLM-generated texts indicate that readability, syntactic complexity, cohesion, and lexical diversity affect text difficulty, indicating that these features should be monitored when using LLMs for assessment (Palma and Soto, 2025; Zhang, 2025).

2.2 Maze Comprehension Tests

Maze reading assessments have long been used within CBM frameworks to efficiently and reliably monitor students' comprehension skills (Deno, 1985). Research shows Maze performance strongly correlates with standardized reading comprehen-

sion measures and sensitively indicates both level and growth (Shin et al., 2000; Shin and McMaster, 2019). Traditional Maze assessments use fixed passages, limiting content diversity and potentially introducing practice effects with repeated administrations. Recent efforts to automate Maze construction have used algorithmic distractor selection and text processing rules to expand item pools, but these often depend on surface-level linguistic features and do not fully capture natural variation in text complexity (Attali et al., 2022; Boyce et al., 2020; Conoyer et al., 2016). As a result, the linguistic characteristics of passages such as syntactic structure, cohesion, and semantic density may influence item difficulty in ways that are not systematically controlled. This issue becomes more pronounced with LLM-generated passages, which can introduce subtle but meaningful differences in readability and discourse structure. These observations raise a central psychometric question about whether student ability can be estimated independently of item difficulty and passage-level linguistic variation when passages differ across administrations.

2.3 Explanatory Item Response Theory

To address this issue, psychometric approaches that incorporate observable predictors into item difficulty provide a useful foundation. Early Explanatory Item Response Theory (EIRT) research showed that item difficulty can be modeled as a function of measurable features rather than as a fixed parameter (Zwinderman, 1991). Kamata (2001) extended this idea to a multilevel IRT formulation modeling item responses nested within passages and within students, which is very similar to the structure of Maze assessments. This framework is especially relevant when passages vary across administrations or are generated by large language models, because students may encounter different linguistic conditions. By modeling passage-level variability directly, multilevel IRT enables separation of student ability from text-driven differences in item difficulty. Using this framework, the current research investigates whether ability estimates derived from AI-generated Maze passages are independent of random item difficulty and passage-level linguistic characteristics. Recent extensions of this framework have examined nonlinear interactions between person and item covariates in EIRMs. Cho et al. (2024) demonstrated that assuming linearity in these relationships may obscure mean-

ingful variation, and proposed the use of smooth functions to model nonlinear effects in reading education contexts.

3 Methods

3.1 Sample

The dataset for this study was obtained from Storiza, a digital literacy platform that supports reading development by enabling children to craft and illustrate their own stories (Leite et al., 2025). The platform uses a large language model to generate individualized passages that align with the instructional goals of a phonics curriculum widely used in elementary schools in the United States (Lane and Contesse, 2022). On the platform, students select the settings for their story, including grade level, instructional topic, lesson, word count, and a brief idea. Then the platform creates a unique passage based on these selections.

The analytic sample included 129 students in Grades 1 through 3 who completed at least one Maze test. After excluding missing values, the dataset included 4,999 item-level responses across 461 automatically generated Maze tests. These data formed a three-level structure, with Maze items nested within stories and stories nested within students.

3.2 Automated Maze Generation

To test reading comprehension, the platform turns these stories into Maze tasks by replacing approximately every seventh word with a blank, excluding stop words and proper nouns. Each blank is accompanied by three choices that contain one correct answer and two distractors. Students complete the Maze by choosing one option for every blank throughout the passage.

We used four different strategies for automatic generation of the Maze comprehension test. In the phoneme-aligned strategy, distractor selection prioritized words with the same phonemes and number of syllables of the correct word. In the word-ladder strategy, distractor selection used words that were different from the correct word by one letter. In the lesson-aligned strategy, the words containing phonemes that were aligned with a specific lesson of the phonics curriculum were prioritized for selection. Therefore, the distance between words selected was more varied than the other strategies. Also, in the lesson-aligned strategy, distractors were selected from words containing the phonemes

targeted in the lesson. In the traditional strategy, distractors were selected from common English words. Examples of partial Maze tests showing each strategy are presented in the appendix.

All four strategies were implemented algorithmically within the platform. In the phoneme-aligned strategy, a large language model (LLM) was prompted to generate distractors that shared as many phonemes and syllables as possible with the correct word, matched its part of speech, and were not semantically related to it. In the word ladder strategy, the LLM was prompted to generate distractors differing from the correct word by exactly one letter substitution, keeping the same word length and part of speech; a lexicon-based neighbor search was used as a fallback when the LLM did not produce valid candidates. In the lesson-aligned strategy, target words were selected by matching a grapheme pattern from the student's assigned phonics lesson (e.g., *ff*, *ll*, *ss*, *zz* for lesson 42 of the phonics curriculum), with a spacing constraint of six to nine words between blanks enforced to maintain consistent assessment density. Distractors were also required to follow the same phonics pattern, so that all three choices in each blank shared the lesson grapheme, targeting phonics decoding rather than semantic inference. In the traditional strategy, distractors were selected algorithmically from a frequency-filtered pool of common English words matched by part of speech and similar word length, without LLM involvement.

Students engaged with the digital literacy platform by generating individualized passages and completing the corresponding Maze tasks. The system automatically converted each passage into a Maze using one of four Maze generation strategies, which were randomly selected each time a Maze was created. The lesson-aligned strategy was introduced later in the data collection period, resulting in a smaller number of administrations relative to the other three strategies. When students completed a Maze test, the platform recorded item responses, completion time, and the Maze generation strategy used for each item. After data collection, we merged item, story, and student-level information into a single analytic dataset and assigned *d*-ELP difficulty values to the correct target words. We also checked the data for missing values, limited variability, and outliers, and standardized continuous predictors to ensure comparability across scales prior to model estimation. Table 1 presents descriptive statistics for each Maze generation strategy.

3.3 Variables

The main outcome variable represented whether a student responded correctly to each Maze item. Since the response was dichotomous, we estimated all models using a logistic link function. At the item level, we examined several predictors reflecting features of the Maze task. Differently than traditional IRT models for dichotomous items (Embretson and Reise, 2000), which estimate item difficulty as one of the model parameters, we treated item difficulty as known and included it as a predictor in the model. More specifically, for each word selected as the correct option for a Maze item, we obtained a lexical difficulty from the Developmental English Lexicon Project (Compton et al., 2025). This public dataset provides IRT-based difficulty estimates for high-frequency English words. For target words not found in the d-ELP database, we estimated difficulty using phonemic similarity. Specifically, we retrieved the CMU Pronouncing Dictionary pronunciation for each missing word and computed the Levenshtein distance between its phoneme sequence and those of all words in the d-ELP database. We then averaged the d-ELP difficulty values of the phonemically closest words and assigned that average as the difficulty estimate for the missing word.

We also included three dummy-coded indicators of the Maze generation strategy used by the platform. Completion time was recorded once per story, so all items within the same story shared the same value, although it varied across stories and students. We standardized this variable prior to analysis to reduce scale differences across predictors and improve model stability and interpretability.

At the story level, we examined linguistic and textual characteristics of the AI-generated passages. We initially reviewed twelve metrics representing syntactic, lexical, semantic, and safety dimensions for redundancy and limited variability. After examining correlations and distributions, we retained seven features: Flesch Reading Ease, lexical diversity, syntactic complexity, coherence, semantic drift, text novelty, and factual hint. For completeness, we also included the item-level difficulty variable in the preliminary correlation analysis to ensure that it did not introduce multicollinearity with the story-level features. All story-level variables were standardized to ensure a comparable scale.

At the student level, we included a grade indica-

tor. To create a stable student-level measure, we identified the grade level each student selected most frequently across all their stories. This modal grade served to create dummy-coded Grade 2 and Grade 3 predictors, with Grade 1 serving as the reference category.

3.4 Analysis

We predicted item-level accuracy using a three-level multilevel logistic regression model (Kamata, 2001), with items nested within stories and stories nested within students. The model is expressed as:

$$\text{logit}(p_{ijk}) = \beta_{0jk} + \beta_1 D_{ijk} + \sum_{m=1}^M \beta_m S_{jk} + \sum_{c=1}^C \beta_c C_{ijk} \quad (1)$$

$$\beta_{0jk} = \gamma_{00k} + u_{0jk} \quad (2)$$

$$\gamma_{00k} = \delta_{000} + v_{00k} \quad (3)$$

where p_{ijk} denotes the probability that student k answers item i in story j correctly. Item-level difficulty is represented by D_{ijk} , story-level linguistic features by S_{jk} , and student- and assessment level control variables (e.g., grade level, Maze version, completion time) by C_{ijk} . The term β_{0jk} is the story-specific intercept for student k , and γ_{00k} represents the student-specific mean performance level. The story-level random effect u_{0jk} captures deviations of individual stories from each student's expected performance. The parameter δ_{000} denotes the overall mean log-odds of a correct response across all students, stories, and items, and the student-level random effect v_{00k} captures each student's deviation from this population mean and therefore represents the model-based ability estimate.

This approach addressed the non-independence of observations and partitioned variance at the story and student levels, similar to a testlet IRT model. Given the binary outcome variable, we used a logistic link function and included random intercepts for both stories and students to allow baseline accuracy to vary across groups.

Our model-fitting strategy followed a sequential model-building approach. Model 1 was an intercept-only model (i.e., unconditional model), which we used to estimate variance components and intra-class correlations (ICCs) (Snijders and Bosker, 2012). ICCs quantified the proportion of variance attributable to story and student levels, which justified the use of a multilevel framework

over a single-level logistic regression model. We then added sets of predictors in successive models to evaluate how each group of variables improved the explanation of item-level accuracy.

Model 2 included control variables such as three dummy-indicators of the four Maze generation strategies described earlier, completion time, and student grade level to account for differences in assessment format, test-taking speed, and differences across grade levels. Model 3 added seven linguistic features of the AI-generated stories to determine whether text characteristics explained additional variance beyond the control variables. Model 4 incorporated item-level difficulty, the only Level 1 substantive predictor, to assess its unique contribution after accounting for story and student factors.

We compared nested models using likelihood ratio tests to assess whether each block of predictors improved model fit. All analyses were conducted in R using the lme4 package (Bates et al., 2015), and random intercepts for stories and students were retained across all models.

In this multilevel logistic regression model, student ability is represented by the student-level random intercept. Specifically, the term v_{00k} measures each student's deviation from the overall mean log-odds of a correct response after accounting for item difficulty, story-level linguistic features, and control variables. Because v_{00k} reflects student-specific differences in performance across all items and stories, it functions as a model-based ability parameter similar to θ in traditional IRT models. These random intercepts were obtained directly from the fitted multilevel model and provide a student-level measure of reading ability that accounts for the nested structure of the data.

4 Results

As shown in Table 2, Model comparisons indicated that some predictors improved overall fit, while others did not. In the null model, substantial variance at both the story and student levels supported the use of multilevel framework: ICCs showed that 29.0% of the variance in student responses was attributable to differences between students, 24.2% to differences between stories within students, and 46.8% to item-level variation.

Adding the control variables in Model 2 did not significantly improve model fit ($\chi^2 = 9.32$, $df = 6$, $p = .156$), indicating that Maze version,

completion time, and grade level explained little variance when considered together. In contrast, the addition of the seven story-level linguistic features in Model 3 resulted in a significant improvement over Model 2 ($\chi^2 = 15.09$, $df = 7$, $p = .035$). This finding suggests that characteristics of the AI-generated passages contributed meaningfully to item-level accuracy. The largest improvement was observed when item difficulty was introduced in Model 4 ($\chi^2 = 31.45$, $df = 1$, $p < .001$), which demonstrated that difficulty accounted for substantial additional variance beyond all previously included predictors. A targeted comparison between Model 2 and Model 4 further confirmed this pattern ($\chi^2 = 46.54$, $df = 8$, $p < .001$). This comparison showed that item difficulty provided unique explanatory value and that ability estimates were not independent of difficulty unless it was explicitly modeled.

Fixed-effect estimates from the full model showed that item difficulty was the strongest predictor of accuracy ($\gamma = -0.24$, $SE = 0.04$, $p < .001$; Table 3). Because lower values represent easier items, the negative coefficient indicates that accuracy decreased as items became more difficult. A one-unit increase in difficulty reduced the odds of a correct response by 22% ($OR = 0.78$), highlighting the central role of difficulty in Maze performance. Among control variables, only the indicator for the lesson-aligned strategy was significantly associated with accuracy ($\gamma = -0.95$, $SE = 0.45$, $p = .035$), indicating lower odds of correct responses compared to the traditional strategy. Grade level and completion time were not meaningfully related to item-level accuracy.

Of the story-level linguistic features, only Flesch Reading Ease was statistically significant ($\gamma = -0.68$, $SE = 0.23$, $p = .003$). Flesch Reading Ease is scaled from 0 to 100 with higher values indicating an easier passage to read. This negative association indicates that easier to read stories were associated with lower item response accuracy. This negative association may also reflect the fact that AI-generated passages can appear superficially simple while still containing distractors or semantic structures that increase difficulty. Coherence was not statistically significant but was close to the threshold ($p = .055$). The other linguistic features, including lexical diversity, syntactic complexity, semantic drift, text novelty, and factual hint, did not significantly predict accuracy after accounting for other variables.

Random-effect estimates indicated lower variance at both the student and story levels after including all predictors (student variance = 1.67; story variance = 1.57). The ICC from Model 4 indicated that 25.5% of the variance was attributable to differences between students, 24.1% to differences between stories within students, and 50.4% to item-level variation (see [Figure 1](#)). The persistence of variance at higher levels suggests that additional unmeasured factors may influence Maze performance even after accounting for difficulty, linguistic features, and control variables.

Supplemental validity analyses further supported the model-based ability estimates. The correlation between estimated ability (θ) and observed accuracy was strong ($r = 0.82$; [Figure 2](#)), indicating that the model captured meaningful individual differences in reading comprehension. Among students with perfect accuracy on a given story, ability estimates were positively correlated with the average difficulty of items encountered ($r = 0.39$; [Figure 3](#)), showing that higher-ability students succeeded on more challenging stories. These findings confirm that the multilevel model effectively captured variation in both item characteristics and student performance, and that estimated ability scores aligned with theoretical expectations for reading comprehension.

5 Discussion

The findings indicate that item difficulty is the main factor influencing Maze accuracy in passages generated by LLMs. Difficulty showed the strongest fixed effect and led to the greatest improvement in model fit, demonstrating that students' performance is highly sensitive to the lexical characteristics of target words. This result is consistent with Multilevel IRT, which predicts that item-level features have a substantial impact when passages differ across administrations.

Story-level linguistic features had a modest effect on accuracy, with only Flesch Reading Ease reaching statistical significance. Notably, higher readability scores were linked to lower accuracy, suggesting that seemingly simple AI-generated passages may still include distractors or discourse structures that increase cognitive load. Other features, such as coherence and semantic drift, did not significantly predict accuracy once difficulty was included, indicating that item-level variation may reduce their unique contribution.

Among the control variables, only the lesson-aligned strategy of the Maze generation strategy was linked to lower accuracy compared to the reference strategy (i.e., the traditional strategy). This finding suggests that distractor selection methods can significantly affect student performance, even when accounting for item difficulty and passage characteristics. Completion time and grade level were not meaningfully associated with accuracy, indicating that the observed effects were not due to differences in test-taking speed or grade level.

6 Conclusion

This study provides preliminary validity evidence ([American Educational Research Association et al., 2014](#)) for the use of automated Maze generation and scoring in formative assessments of reading comprehension for early readers. More specifically, our finding that item difficulty is the main factor influencing performance on AI-generated Maze assessments is evidence of adequacy with respect to internal structure. The multilevel IRT framework effectively distinguishes student ability from item and passage characteristics, and the strong correlation between model-based ability estimates and observed accuracy is also internal structure evidence. Our results underscore the importance of explicitly modeling item difficulty. We also found that, once difficulty is accounted for, story-level linguistic features and most control variables have minimal impact, which is validity evidence with respect to content.

Limitations of the current study include lack of distractor-level information, because the digital literacy platform did not log the distractors selected. This information would allow examining how distractor properties affect item difficulty and ability estimates. Future research should include distractor-level metrics and additional linguistic features to further strengthen the validity evidence for the scores of the automated Maze assessments. Furthermore, future research should examine validity evidence with respect to response process, particularly regarding how students interact with the four strategies for Maze generation and whether the response process varies across grades. There could also be interactions between linguistic features of the passage and the Maze generation strategy. The results of the current study show that the automated Maze is a promising formative assessment for use in digital literacy platforms.

7 Tables and Figures

Table 1: Descriptive Statistics by Maze Generation Strategy

Strategy	N	Items		Prop. Correct	
		M	SD	M	SD
Phoneme-Aligned	162	11.22	7.48	.941	.235
Traditional	133	10.35	6.14	.951	.217
Word-Ladder	149	9.79	5.31	.923	.267
Lesson-Aligned	29	12.52	9.47	.873	.333

Table 2: Model comparison: Fit indices, variance components, and likelihood ratio tests.

	Model 1 (Null)	Model 2 (Controls)	Model 3 (Linguistic)	Model 4 (Full)
Variance				
Student var. (τ^2)	2.043	1.912	1.596	1.667
Story var. (σ^2)	1.703	1.630	1.528	1.571
ICCs				
Student (Level 3)	0.290	0.280	0.249	0.255
Story (Level 2)	0.242	0.240	0.238	0.241
Unexplained	0.468	0.482	0.513	0.504
Fit Indices				
AIC	1988.0	1990.7	1989.6	1960.2
BIC	2007.6	2049.4	2093.9	2071.0
-2 LL	1982.0	1972.7	1957.6	1926.2
LRT				
χ^2	—	9.32	15.09	31.45
df	—	6	7	1
p	—	0.156	.035*	<.001***

Note. Fit indices include AIC, BIC, and -2 log-likelihood. Sequential LRTs compare each model with the preceding model. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3: Fixed Effects Estimates from the Full Model (Model 4)

Predictor	γ	SE	z	p	OR
(Intercept)	3.438	0.393	8.751	<.001***	31.12
Control Variables					
Phoneme-aligned	-0.202	0.299	-0.673	0.501	0.82
Word ladder	-0.559	0.295	-1.897	0.058	0.57
Lesson-aligned	-0.954	0.451	-2.114	.035*	0.33
Grade 2 (vs. Grade 1)	0.012	0.389	0.030	0.976	1.10
Grade 3 (vs. Grade 1)	0.361	0.693	0.521	0.602	1.44
Comp. Time (scaled)	-0.049	0.120	-0.408	0.683	0.95
Item-Level Predictor					
Difficulty	-0.243	0.044	-5.537	<.001***	0.78
Linguistic Features					
Flesch (scaled)	-0.681	0.230	-2.968	.003**	0.51
Lex. Diversity (scaled)	0.072	0.170	0.427	0.670	1.07
Syntactic Comp. (scaled)	0.065	0.141	0.463	0.644	1.07
Coherence (scaled)	-0.194	0.101	-1.915	0.055	0.82
Semantic Drift (scaled)	0.137	0.198	0.691	0.489	1.15
Text Novelty (scaled)	-0.206	0.187	-1.103	0.270	0.81
Factual Hint (scaled)	-0.065	0.222	-0.294	0.779	0.94

Note. The traditional strategy for Maze generation served as reference; coefficients (γ) represent log-odds of a correct response; OR = odds ratio. * $p < .05$. ** $p < .01$. *** $p < .001$.

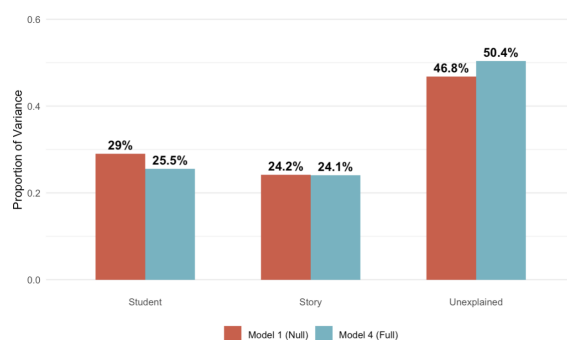


Figure 1: Proportion of Variance at Each Level in the Null and Full Models.

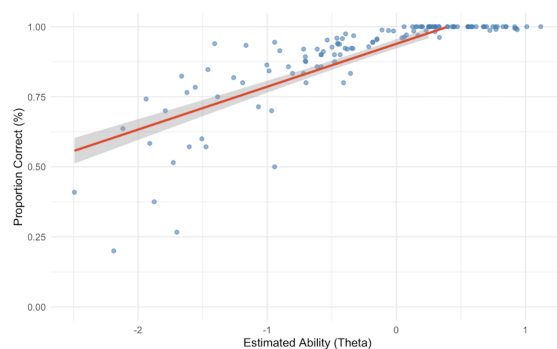


Figure 2: Relationship Between Model-Based Ability Estimates (θ) and Observed Accuracy.

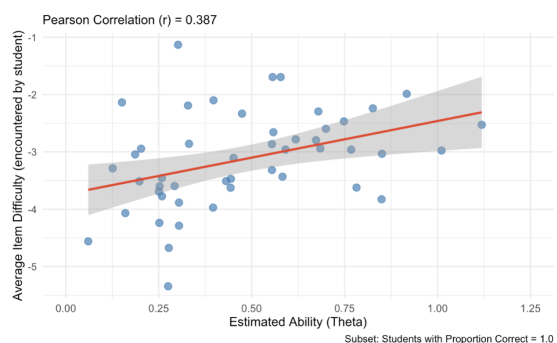


Figure 3: Relationship Between Item Difficulty and Ability Estimates for Students With Perfect Accuracy.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Y. Attali, A. Runge, G. T. LaFlair, K. Yancey, S. Goodwin, Y. Park, and A. A. Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903.

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. [Maze made easy: Better and easier measurement of incremental processing difficulty](#). *Journal of Memory and Language*, 111:104082.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Okan Bulut and Seyma Nur Yildirim-Erbasli. 2022. [Automatic story and item generation for reading comprehension assessments with transformers](#). *International Journal of Assessment Tools in Education*, 9:72–87.
- Sun-Joo Cho, Amanda Goodwin, Matthew Naveiras, and Paul De Boeck. 2024. [Modeling nonlinear effects of person-by-item covariates in explanatory item response models](#). *Journal of Educational Measurement*.
- R. Circi, J. Hicks, and E. Sikali. 2023. Automatic item generation: Foundations and machine learning-based approaches for assessments. In *Frontiers in Education*, volume 8, page 858273. Frontiers Media SA.
- D. Compton, J. Rueckl, L. Steacy, N. Siegelman, A. Edwards, Y. Petscher, and C. Starke. 2025. The developmental english lexicon project (d-ELP): Continuous word reading difficulty estimates of the 9,961 most frequently printed english words for US children. Manuscript submitted for publication.
- Sarah J. Conoyer, Erica S. Lembke, John L. Hosp, Christine A. Espin, Michelle K. Hosp, and Apryl L. Poch. 2016. [Getting more from your maze: Examining differences in distractors](#). *Reading & Writing Quarterly*, 33(2):141–154.
- S. L. Deno. 1985. Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52:219–232.
- Susan E. Embretson and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.
- M. J. Gierl and T. M. Haladyna. 2012. *Automatic item generation: Theory and practice*. Routledge. Eds.
- M. J. Gierl, H. Lai, and S. R. Turner. 2012. Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8):757–765.
- A. Kamata. 2001. Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1):79–93.
- H. Lane and V. Contesse. 2022. *UFLI Foundations: An explicit and systematic phonics program*. Ventris Learning.
- W. L. Leite, S. Hammerschmidt-Snidarich, A. C. Huggins-Manley, H. K. Karakis, H. Lane, S. Rahimi, Q. Shen, X. Tai, and M. Schmidt. 2025. [Storiza: A platform to support children’s oral reading fluency development with generative AI](#). In *Proceedings of the 19th International Conference of the Learning Sciences – ICLS 2025*, pages 1574–1578. International Society of the Learning Sciences.
- D. Palma and C. Soto. 2025. Combining large language models with linguistic features for the readability complexity assessment of texts. In Tareq Ahram, Waldemar Karwowski, and Jay Kalra, editors, *Human Factors in Design, Engineering, and Computing*, volume 199 of *AHFE Open Access*. AHFE International, USA.
- Seyedahmad Rahimi and Valerie J. Shute. 2024. [Stealth assessment: a theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments](#). *Educational Technology Research and Development*, 72(5):2417–2441.
- J. Shin, S. L. Deno, and C. Espin. 2000. Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education*, 34(3):164–172.
- J. Shin and K. McMaster. 2019. [Relations between CBM \(oral reading and maze\) and reading comprehension on state achievement tests: A meta-analysis](#). *Journal of School Psychology*, 73:131–149.
- Tom A. B. Snijders and Roel J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2 edition. Sage Publications, Thousand Oaks, CA.
- M. Sommer and M. Arendasy. 2025. Automatic-and transformer-based automatic item generation: A critical review. *Journal of Intelligence*, 13(8):102.
- B. Tan, N. Armoush, E. Mazzullo, O. Bulut, and M. Gierl. 2025. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.
- C. Xiao, S. X. Xu, K. Zhang, Y. Wang, and L. Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 610–625.
- J. Zhang. 2025. Examining the robustness of large language models across language complexity. *arXiv preprint arXiv:2501.18738*.
- Aeilko H. Zwinderman. 1991. [A generalized Rasch model for manifest predictors](#). *Psychometrika*, 56(4):589–600.

A Appendix - Example of Maze tests for a lesson focused on graphemes oi and oy making the sound /oi/

A.1 Phoneme-aligned strategy

Once upon a time, there was a princess named Joy. Joy loved to (play / clay / pray) with her toys in the royal garden. One day, she (found / bound / sound) a shiny coin in the soil. She picked it up and (smiled / styled / filed) with joy. Joy decided to (show / glow / snow) the coin to her friend, a boy named Roy. Roy was (coy / toy / soy) but kind.

A.2 Word ladder strategy

Once upon a time, there was a princess named Joy. Joy loved to (play / clay / ploy) with her toys in the royal garden. One day, she (found / sound / round) a shiny coin in the soil. She picked it up and (smiled / tiled / piled) with joy. Joy decided to (show / slow / snow) the coin to her friend, a boy named Roy. Roy was (coy / boy / soy) but kind.

A.3 Lesson-aligned strategy

Once upon a time, there was a princess named (Joy / Boy / Toy). Joy loved to play with her (toys / boys / joys) in the royal garden. One day, she found a shiny (coin / join / loin) in the soil. She picked it up and smiled with (joy / soy / coy). Joy decided to show the (coin / loin / join) to her friend, a boy named Roy. Roy was (coy / soy / joy) but kind.

A.4 Traditional strategy: common English words

Once upon a time, there was a princess named Joy. Joy loved to (play / work / sing) with her toys in the royal garden. One day, she (found / made / lost) a shiny coin in the soil. She picked it up and (smiled / cried / ran) with joy. Joy decided to show the coin to her friend, a boy named Roy. Roy was (coy / mad / big) but kind.