

Using XGBoost to Construct a Vertically Aligned Skill Difficulty Scale

John Bielinski

Renaissance Learning

john.bielinski@renaissance.com

Abstract

We sought to construct a vertical difficulty scale for the skill library in Freckle, a widely used online mathematics practice system. The scale would span kindergarten through grade 8 and permit direct comparisons of skill difficulty across grade levels. We fitted an Extreme Gradient Boosting (XGBoost) prediction model to 3 million practice sessions covering over 1,800 mathematics skills. We applied the model as a counterfactual, fitting it to the full population of students to generate a population percent-correct score for each skill in adjacent grades. We linked adjacent grades using Thurstone's scaling method and validated the resulting scale against independently derived Rasch-based skill difficulties from Star Math. The two scales were strongly associated (Pearson $r = 0.80$), supporting the use of large-scale practice data to estimate comparable skill difficulties across grades.

1 Introduction

Following formal, teacher-led instruction, students are often given additional classroom time to practice newly learned skills. Online skill-practice systems are designed to support the transfer and fluent application of those skills. These systems often use adaptive algorithms to select skills aligned with each student's current proficiency. Algorithmically generated pathways can reduce teachers' workload while maximizing the time students spend practicing appropriately challenging skills.

Math skill practice pathways typically divide content into domains and employ a hierarchical structure that follows a logical instructional sequence. This approach can be very effective when paired closely with teacher-led instruction, where students are given the necessary guidance when struggling. However, for independent practice, moving all students through a fixed practice sequence can be problematic because it does not

account for content difficulty. Aside from basic computation skills, the hierarchical structure of mathematics is not obvious. Evidence suggests that skill difficulty is not tightly aligned to the instructional sequence (Robar and Bryan, 2021) and results in trip steps that pose a significant challenge for students and often require more direct teacher support to develop. Tasks that are too difficult increase cognitive load and lead to frustration and reduced engagement (Gupta and Zheng, 2020).

Another problem is misplacement. Most systems use a brief placement test to identify each student's starting point within the sequence. Misplacement is not uncommon. When a student is misplaced in the sequence, the student can get stuck in a repetitive sequence of content that is too difficult or too easy before the system can correct the mistake. During the time it takes the system to adjust, the student may have already disengaged.

Incorporating task difficulty enables the system to match content to each student's current proficiency. For independent practice, the objective is to build fluency and generalization, which are best supported when accuracy rates are between 85% and 100% (Burns et al., 2006). In a comprehensive online math practice system, many skills are often within this range or optimal practice zone (OPZ), and far more are not. Using the logistic relationship between student ability and task difficulty, all skills meeting the criteria can be immediately identified. This information can be incorporated into a task-selection optimization system to (1) identify tasks within each student's OPZ and (2) advance the student through the system with logic that promotes best practices such as interleaving, spaced repetition, and breadth of coverage. Additionally, task/skill difficulty can help domain experts identify skill dependencies, especially where such dependencies are not obvious. This paper presents the strategy we used to develop a vertically sequenced skill difficulty scale.

2 The Approach

We sought to construct a vertical scale for the library of skills from a popular online skill practice system known as Freckle. This vertical scale would span skills from kindergarten through grade 8. The principal challenge was to obtain skill performance scores from a representative sample of students in adjacent grades. Traditional methods for linking test scores require representative samples from two or more grades who complete the same test at a fixed time of the year (Kolen and Brennan, 2004). In the extant data, practice events for any given skill were distributed widely across the year, with some students practicing the skill at the beginning of the year and others in the middle or end of the year. Additionally, for any pair of skills, the ability level of students practicing them could vary by as much as ± 1 SD based on each student's recent Star Math screening test score, and sample sizes for each skill varied from about 200 to over 10,000. To overcome these limitations, we needed a model that was not encumbered by the strong assumptions of item response theory (IRT), was highly flexible, and could provide strong fit to the observed data. Such a model could be used as a counterfactual for inference to a representative population.

We chose Extreme Gradient Boosting (XGBoost) to fit the data. The model was trained on 3+ million skill practice sessions representing over 1,800 distinct skills from grades K–8 from an entire school year. XGBoost represents a state-of-the-art predictive modeling tool for structured data. It combines individual decision trees in an ensemble to improve prediction by building trees sequentially, with each new tree correcting the errors of the previous ones. It seamlessly captures non-linearity and interactions, handles missing values automatically, includes regularization to prevent overfitting, and can readily accommodate time-series data (Chen and Guestrin, 2016).

Our premise was that a prediction model with strong fit could update student proficiency daily as students completed practice sessions and standardized assessments. The fitted model could then be used to generate skill-performance predictions for a representative population, including students who had not practiced a given skill.

2.1 Vertical Linking Strategy

We treated the data as a common-item design, with skills serving as shared test components across

groups. The population mean percent correct for each skill corresponded to an item's mean score, and the distribution of skill means represented the raw-score distribution. This analogy is appropriate because the distribution of mean skill scores parallels the distribution of total scores in the population. A skill-level percent-correct score can also be viewed as an aggregate of component responses. Although aggregation discards some information, the resulting score remains sufficient for estimating difficulty.

The XGBoost skill prediction scores are generated for each skill across the full sample of students in adjacent grades to provide representative sample scores on common items across two grades. We applied Thurstone's scaling method for two adjacent grade groups as described in Chapter 9 of Kolen and Brennan (2004) and propagated the linking down and up the chain to link all adjacent grade pairs.

In the first step, we trained an XGBoost model on several million Freckle skill practice sessions (details provided below). Next, we used the trained XGBoost model to predict skill performance for a full population of students across adjacent grades. For example, predictions were made for all 211 grade 3 skills for the total sample of 50k+ grade 3 students and the 74k+ grade 4 students. This procedure was repeated for all grade-level skill sets from K–8. The mean predicted score for each skill by grade served as the raw score points for the Thurstone scaling procedure. The scale was centered at grade 5 and propagated across grades using the overlap in adjacent grades.

3 The Data

The training set consisted of Freckle math skill practice sessions across a school year where students completed 10 items per skill per session. The final sample included about 400k students distributed across KG through grade 8 (see Figure 1). On average, each student completed about 8 skills across the school year.

3.1 Online Math Practice System

The Freckle skill practice system contains 1,819 math skills spanning grades KG through 8 (see Figure 1). It was originally designed to represent the standards in the Common Core Math Standards (National Governors Association Center for Best Practices and Council of Chief State School Of-

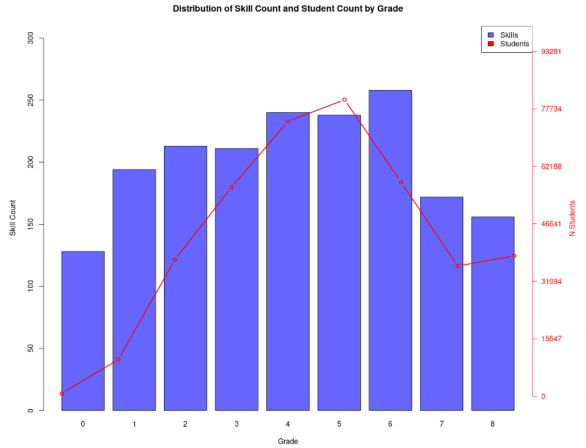


Figure 1: Distribution of skill count and student count by grade.

ficers, 2010). It includes number and operations, measurement and data, geometry, fractions, operations and algebraic thinking, ratios and proportions, number systems, and statistics. Item types include multiple-choice, matching/drag-drop, and fill-in-the-blank.

Students enter the system either through teacher-assigned targeted practice or through the adaptive pathway. Students are placed into the adaptive pathway through a built-in placement test. Movement through the adaptive pathway is governed by rules tied to student performance. Students progress through a fixed skill hierarchy after demonstrating mastery of each skill. Within each skill, items are selected at random and administered one at a time. When the student submits a response, the system indicates whether the answer was correct and advances the student to the next item. Performance is reported as the percent of items answered correctly.

3.2 XGBoost Model Training

We configured the practice session data into a weekly time series, where each row represented a combination of student by skill by week of the school year. The number of rows per student per week equals the number of unique skills practiced in that week. The dependent variable was the percent-correct score on each Targeted Practice skill session in that week. Adaptive practice results were not included in the dependent variable because skill selection was already tied to prior placement test and practice scores, thus substantially biasing the sample and restricting the range of performance.

The predictors (a.k.a. features) are listed below

Feature	Description
<i>Properties of skills being predicted</i>	
Skill Difficulty	Mean skill percent correct score from all practice sessions across three years
SD of Skill Difficulty	—
Mean Test Score	Mean scaled score of the most recent interim math assessment for students used to compute Skill Difficulty
Prior Scaled Score	Most recent standardized test scaled score
<i>Aggregated prior skill practice</i>	
Mean Pct. Correct Same	On the same skill, prior 12 weeks
Mean Pct. Correct All	Across all skills practiced, prior 12 weeks
Mean Pct. Correct Difference	(Mean pct. correct, all skills, prior 12 weeks) – Skill Difficulty
<i>Other</i>	
Student Grade	—
Skill Grade	—
Student Grade – Skill Grade	—
Week of School Year	—

Table 1: XGBoost model features by category.

(see Table 1). They are clustered into categories by type.

Feature importance summary. The five most important features were Skill Difficulty, Mean Percent Correct Difference, Mean Percent Correct Same, Prior Scaled Score, and SD of Skill Difficulty. These features accounted for 90% of total feature importance.

These features allowed us to (a) achieve high predictive accuracy, (b) incorporate prior skill practice results to better represent current proficiency and specific skill proficiency, and (c) maintain interpretability. Prior to running XGBoost training, we employed a grid-search technique with 5-fold cross-validation to identify optimal hyperparameters. For training, the sample was divided randomly by student into an 80%/20% training/test data split. Results are based on the 20% test sample.

4 Results

4.1 XGBoost Model Fit

Overall model fit was evaluated using root mean square error ($RMSE = 16.85$), mean absolute error ($MAE = 12.34$), and $R^2 = 0.561$. To place these values in context, we estimated the proportion of true-score variance captured by the model. As a proxy for true-score variance, we used the average Pearson correlation between student performances

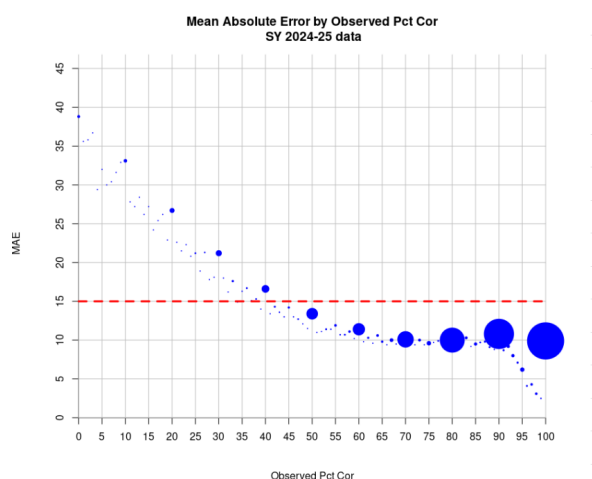


Figure 2: MAE as a function of the observed percent-correct score.

on skills practiced repeatedly within a two-week period. Relative to this proxy, the XGBoost model accounted for approximately 89% of the true-score variance ($0.56/0.63$). Because each practice session contained 10 items, the MAE corresponds to an average deviation of about 1.2 correctly answered items from the observed score.

Figure 2 presents overall model fit (MAE) as a function of the observed percent-correct score. MAE is stable around 10 for scores from 70% to 100% correct. This is the region with the highest concentration of data, as evidenced by the size of the points, which are proportional to the number of cases. MAE exceeds 15 for scores below 50% correct, where there is relatively little data to model.

4.2 Thurstone Scaling

We applied the trained XGBoost model to predict student performance on all skills within a grade, for all students from that grade and the next higher grade. The mean predicted percent correct by grade for each skill served as a raw score in the distribution of scores by grade. For example, in grade 3 there were 211 skills, each representing a raw score. We examined the scatter plots by skill set across adjacent grades and found them to be linear in all cases. Thus, it was reasonable to apply the normalized z-score transformation to the score distribution. The within-grade normalized z-score transformation was based on a clipped range of skills with a mean percent correct between 30% and 90%, in line with the recommendation of Guliksen cited in Kolen and Brennan (2004), p. 385. Next, the Thurstone regression was fitted to the

common skill set across adjacent grades. The scale was centered on grade 5, then propagated upward through grade 8 and downward through kindergarten. In the final step, all scores were placed onto the anchor-grade scale. Further smoothing of the grade-to-grade growth was used to better align the scale to national norms.

To validate the scale, we compared it with skill-difficulty estimates derived from the Star Math item bank. Each Star Math skill difficulty is based on all Rasch-calibrated items associated with that skill. We smoothed these estimates by regressing skill difficulty on the point in the grade when each skill is typically introduced. A fitted curve defined the learning progression (LP), and projecting each observed difficulty onto the curve yielded its LP difficulty. We then used a mean-sigma transformation to place the new Freckle vertical-scale difficulties on the Star Math LP difficulty scale.

Figure 3 shows a strong, linear relationship between the XGBoost-derived skill difficulties and the Star Math LP difficulties (Pearson $r = 0.80$). A small number of points fall well outside the line of best fit. Because these differences exceed expected sampling variation, they likely reflect genuine differences in how corresponding skills are represented in the two products. Some Freckle skills appear substantially easier than their Star Math counterparts, while others appear more difficult. The pattern may also reflect differences in estimation methods: XGBoost modeling for Freckle versus IRT calibration for Star Math. In addition, the data were collected in different settings. Freckle data came from low-stakes, self-paced practice sessions with performance feedback, whereas Star Math calibration data came from moderate-stakes interim assessments administered under standardized conditions. Differences in content and item format may also contribute.

We also examined the relationship between the vertical Freckle skill difficulty scale and the mean percent-correct scores for each skill, using all practice data from three school years. Figure 4 shows this relationship for skill sets in kindergarten and grades 2, 5, and 8. As expected, the relationship is negative, because the observed skill difficulty scale functions as an easiness scale, with higher values indicating easier skills. Within-grade correlations are strong, ranging from 0.85 in kindergarten to 0.94 in grade 8. The upward shift in points by grade also reflects the increasing difficulty of skills at higher grade levels. This shows that the new Freckle skill

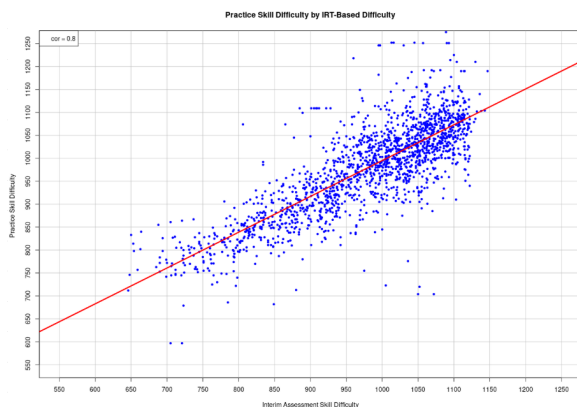


Figure 3: Freckle skill difficulty vs. Star Math difficulty.

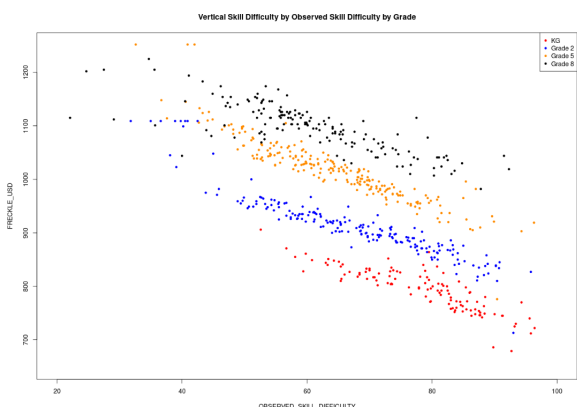


Figure 4: Relationship between the new Freckle skill difficulty scale and the population mean percent-correct score, by grade.

difficulty scale adequately reflects real differences in the observed performance on each skill; within grade, skills with lower observed mean percent-correct scores are harder than those with higher mean percent-correct scores.

5 Discussion

This paper presents an effective and novel strategy for constructing a vertical scale from existing skill practice data when observations are not drawn from representative samples and accumulate across an entire school year. Through this combination of machine learning and scaling methods, we were able to address a major shortcoming in online math skill practice systems, namely the lack of a difficulty metric that enables direct comparison of skill difficulty across the skill library. Most adaptive systems rely on a predefined skill hierarchy that ignores statistical difficulty. Adaptation either moves a student forward or backward, or sometimes to a predefined prerequisite skill, based on current skill

performance. This ignores research demonstrating the benefits of independent fluency-building practice when students achieve 90% accuracy. Our own internal research, not described here, found that students who encountered skills within the optimal practice zone (OPZ, defined as 85% to 100% correct) had the highest annual growth rates based on an independent interim assessment. Furthermore, they were also more likely to practice additional skills relative to students who were placed in a skill where they performed below 85% correct. By building a vertical scale that is tied directly to the vertical scale used to report performance on the placement test, we were able to use a simple logistic relationship to identify all skills within each student's OPZ with a high degree of confidence.

Vertical skill difficulty offers additional benefits. It can be incorporated into an optimal skill-selection system to identify skills of comparable difficulty, whether they belong to the same mathematics domain or different domains, and therefore make good candidates for selection. We are exploring a strategy that combines skill difficulty, conceptual relationships, and instructional sequence to select skills that are closely matched on all three dimensions and support interleaving. Interleaving pairs related skills within practice so that students must distinguish among approaches; when students achieve accuracy on both skills, this practice can promote generalization (Rohrer et al., 2015).

Because skill-specific proficiency can change frequently, we needed a mechanism to incorporate changes in proficiency over the school year. We designed an XGBoost model that combines all recent test score data and practice session results with other student characteristics (e.g., grade level), skill characteristics (e.g., skill grade, domain), and time of year. What makes XGBoost so useful is that this approach identifies the best combination of feature weights to maximize fit, and it does so in a way that captures unanticipated feature interactions, nonlinearity, and time-series dependencies. We found that this model had very little bias across the full ability range and very little within-student bias.

Limitations

Although the evidence presented in this paper provides validation of this approach for generating a vertical skill difficulty scale, it is important to recognize that, unlike traditional IRT-scaling methods, it is atheoretical. It also requires much more data

to achieve the same level of precision as traditional IRT methods. Its best application is to large-scale extant data where prior proficiency data—via a placement test, standardized test, prior practice, or a combination of these—is available. While this method may not have wide applicability, it could be very beneficial to organizations with legacy systems that hold a large amount of student-level data.

References

- Matthew K. Burns, Amanda M. VanDerHeyden, and Cynthia L. Jiban. 2006. Assessing the instructional level for mathematics: A comparison of methods. *School Psychology Review*, 35(3):401–418.
- Tianqi Chen and Carlos Guestrin. 2016. [XGBoost: A scalable tree boosting system](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Udita Gupta and Robert Z. Zheng. 2020. [Cognitive load in solving mathematics problems: Validating the role of motivation and the interaction among prior knowledge, worked examples, and task difficulty](#). *European Journal of STEM Education*, 5(1):05.
- Michael J. Kolen and Robert L. Brennan. 2004. *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd edition. Springer.
- National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common core state standards for mathematics. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers.
- J. Robar and J. Bryan. 2021. [The critical role of trip steps in math recovery](#). Renaissance. Published September 24, 2021.
- Doug Rohrer, Robert F. Dedrick, and Sandra Stershic. 2015. [Interleaved practice improves mathematics learning](#). *Journal of Educational Psychology*, 107(3):900–908.