FACILITATING STUDENT-INSTRUCTOR CONVERSATION WITH BAYESIAN ITEM RESPONSE THEORY

Emi Cervantes, Sam Baugh, Neil Hatfield
The Pennsylvania State University, USA
exc5589@psu.edu

Low course grades can decrease motivation and contribute to STEM attrition when they fail to represent students' true understanding and growth. Given that much of STEM work touches upon ideas from Statistics, making sure that Statistics courses do not contribute to these negative impacts is a worthy cause. One way in which instructors might help mediate these impacts is through one-on-one conversations about grades and learning with students. We propose a tool for instructors teaching Statistics courses using standards-based grading to facilitate such conversations. Leveraging Bayesian Item Response Theory, the tool provides instructors with predictions of students meeting or exceeding course standards based on their current performance. Such information can assist instructors in making grades more meaningful to students and building pathways for student success.

INTRODUCTION

In the United States of America, points-based grading systems that instructors then convert to letter grades (A, B, C, D, or F) are the dominant way of communicating student learning. However, these grades often fail to accurately describe students' understandings and individual learning trajectories. These systems often result in a confounded mixture of students' understandings, growth, effort, performance, and other socioeconomical factors (Brookhart, 1991; Guskey & Link, 2019). Thus, there is a discrepancy between the letter grade and students' actual learning. Further, students' experiences with the letter grades can result in lower motivation and can cause negative impacts on their performance in future courses (Chamberlin et al., 2023; Pulfrey et al., 2011). Losing motivation in such a way can lead to students avoiding subsequent statistics courses and even dropping out of their programs. Given how a statistics course is almost universally required for any student in a STEM program, careful consideration of grades in such courses is critical.

The Guidelines for Assessment Instruction in Statistics Education (GAISE) college report stresses the importance of providing reliable assessments to improve both learning and teaching experiences (GAISE College Report ASA Revision Committee, 2016). Assessments should provide feedback to instructors on how well students are learning, so that instructors can use such information to tailor their teaching to target areas that students need to spend more time on. Moving away from the traditional points-based systems to alternatives allows instructors to implement these GAISE assessment recommendations. Alternative grading systems focus on assessing students' proficiency on learning outcomes for better feedback centered on student learning. Three of the most common forms of alternative grading include Mastery, Specifications (Specs), and Standards-based grading (SBG); all of which aim to improve the relationship between students and their grades through assessments of their skill proficiency and understandings (Brookhart et al., 2016; Campbell et al., 2020; Hatfield, 2021; Nilson, 2015; Wiggins, 1994). Each of these approaches centers on instructors calculating students' grades based on their proficiency on learning outcomes that are assessed multiple times over the course. Depending on the system, the grading team marks student work either as Proficient/Not Yet when using a Mastery/Specs approach or in an ordered category (e.g., Adept, Skillful, Progressing, Beginning) in SBG. Importantly, students have multiple chances to demonstrate proficiency for any learning outcome. By focusing on assessing the proficiency, these alternative grading approaches help students perceive their scores as feedback rather than static labels of their ability. Furthermore, they help instructors see what skills students have mastered and the areas that need further learning intervention to support their progress.

Multiple instructors have implemented SBG in the Statistics classroom (e.g., Curley and Downey, 2024; Hatfield, 2021) due to its benefits, but they often face challenges. Curley and Downey (2024) surveyed groups of students in Mathematical Statistics courses that used one form of SBG. In their surveys, many students expressed that SBG has helped them reduce their test anxiety and focus more on learning than on their ability to earn enough points to get certain grades (Curley & Downey, 2024). However, many challenges occur on the instructors' side due to difficulty in implementing SBG.

With the need for careful planning of writing out transparent learning outcomes and frequent grading sessions due to multiple reassessments, instructors must dedicate significant time to transition to and maintain the SBG system. In addition, instructors cannot give current and potential final letter grades to their students because the letter grade is not calculated until the end of the term. Since most institutions require letter grades, students will expect real-time updates of their letter grades, which are difficult to provide with SBG. These challenges highlight the need for a dedicated tool that provides instructors with real-time, meaningful insights into learning progress and projected letter grades of students that enhance the quality of feedback.

Item Response Theory

Item Response Theory (IRT) consists of a set of popular tools that model the assessment items to analyze and evaluate their characteristics and their relationship with the latent variable. One of the most salient elements of IRT in education application is the latent variable that represents students' ability (Cai et al., 2016). Commonly denoted by θ_{ij}, where $i = 1, \ldots, N$ and $j = 1, \ldots, J$ denote each individual and item, respectively, this latent ability is used to compare individual abilities and to design better assessments. It is indirectly computed with item responses and is used to compute the probability of answering an item correctly. When computing this probability, other parameters, including the item difficulty, δ_j, are used. Item difficulty parameters are relevant in practice, as high performance on a difficult item can indicate strong latent ability. Furthermore, many forms of longitudinal IRT (L-IRT) have been developed to analyze items and latent abilities that are measured repeatedly over time (Wang & Nydick, 2020). Such information from the model output can be used in a classroom to identify certain skills that students are struggling to improve in their learning over time.

The proficiency scores in SBG can be modeled using IRT, but no literature has explored this opportunity. Since the true proficiency level of students is an unobserved, latent ability, L-IRT can model the trajectories of proficiency levels in each learning outcome using the observed learning outcomes from multiple assessments over time. This approach with L-IRT will help resolve the challenge with implementation, allowing instructors to better communicate with their students about their progress as well as their current and potential final grades. Nonetheless, while the potential for IRT to predict the latent ability and item responses has been established (Ayers & Junker, 2008; Park et al., 2023), there are currently no practical applications of these models in classrooms using either traditional or alternative grading systems. In our study, we leverage the potential that L-IRT offers as a prediction tool to resolve the challenges posed to instructors and further support student learning in a statistics classroom that uses SBG. With the insights on students' current and future latent ability of each learning outcome through L-IRT, our goal is to help instructors have more meaningful conversations with their students and provide timely feedback that harnesses the benefits of SBG.

METHODS

We propose a new longitudinal IRT model to be implemented as a tool to help instructors monitor learning progress and facilitate one-on-one conversations with students in an SBG-based statistics course. Our study uses a dataset containing learning outcome scores collected at multiple timepoints throughout a single-semester ANOVA course implementing SBG.

Course Structure

In this course, the instructor created seven different learning objectives: Literacy Reasoning and Thinking (LRT), Communication (Comm), Technology (Tech), ANOVA, Data Analysis (DA), Statistical Inference (SI), and Design of Studies (Design). These learning objectives reflect the course objectives, which represent the broad competencies that the course intends to cover. Each learning objective is linked to multiple learning outcomes in the form of specific, assessable skills that fall under each learning objective. In this course, the instructor implemented a variation of SBG in which the proficiency level for each learning outcome was altered to contain three categories. There are two proficiency levels: Proficiency and Not Yet, and each level contains three categories (see Table 1).

Each learning outcome score is based on the average of the discrete scores that the student receives for corresponding items assessed throughout the assignment, homework, quizzes, tests, final exam, and course project. Most proficiency categories are determined by score intervals, whereas Adept and Not Shown require 100% and 0% on all related items, respectively. Additionally, each learning objective is calculated by taking the average of the associated learning outcome scores. There are nine different timepoints in the dataset, with the first timepoint being the sixth week and the last timepoint being taken after the final project. It is important to note that not all learning objectives and learning outcomes are assessed at every single timepoint.

Table 1: Proficiency levels and their three categories used in the SBG-based ANOVA course with score intervals and discrete item scores

Level	**Category**	**Score**
Proficiency	Adept	100%
	Highly Developed	[85%, 100%) 92.5%
	Skillful	[70%, 85%) 77.5%
Not Yet	Progressing	[50%, 70%) 60%
	Beginning	(0%, 50%) 25%
	Not Shown	0%

Longitudinal Beta-IRT Model with the AR Component

The response variable that we are interested in modeling is the individual learning outcome score, which we treat as the item response data. Each learning outcome in our dataset is a continuous variable rather than a binary variable (1 = Proficient, 0 = Not Proficient) in IRT. To model our continuous response variable, we implement a Beta-IRT model that can take the item response data between 0 and 1 (Noel & Dauvier, 2007). Additionally, we extend the traditional Beta-IRT model to a longitudinal framework to model the trajectory of learning outcomes over time, called the longitudinal Beta-IRT model (L-Beta-IRT). The observed learning outcome, y_{ijt}, of a student i to the learning outcome j assessed at timepoint t is assumed to follow a Beta distribution with parameters, m_{ijt} and n_{ijt}. These parameters are described as a function of item difficulty corresponding to learning objective k, δ_k, and latent ability, θ_{ijt}. Due to computational challenges, the difficulty parameters are set to represent learning objectives, and we keep them fixed across all timepoints.

We modeled students' latent ability as a regression describing their growth in learning outcomes over time. The current ability of a student tied to a learning outcome j, θ_{ijt}, should be related to the previous ability to the same learning outcome, $\theta_{\text{ijt}-1}$, $\theta_{\text{ijt}-2}$, …, because learning is a cumulative process. We incorporate an autoregressive model of order 1, AR(1), to model latent ability and its dependence on previous ability, which yields the following regression model,

$$\theta_{\text{ijt}} = \beta_0 + \alpha_i + \alpha_{j(i)}t + \rho\theta_{\text{ijt}-1} + \epsilon_{ijt}.$$

This regression model consists of a fixed intercept, β_0, and two random effects, α_i and $\alpha_{j(i)}$. The random effect α_i captures individual-specific intercept, and the second random effect, $\alpha_{j(i)}$, represents the slope parameter for a student i with respect to the corresponding learning objective. The index $j(i)$ is used to map each learning outcome to its corresponding learning objective for each individual. Furthermore, we assume that the random slope, $\alpha_{j(i)}$, is strictly positive to account for growth in students' ability over time. The coefficient ρ is the autoregressive coefficient, which describes the effect of ability from the previous timepoint. The last term, ϵ_{ijt}, denotes the error term.

To test the predictive performance of L-IRT-Beta with the AR component (L-IRT-Beta-AR), we fit the model using training data and predict new data from testing data through a cross-validation method. The training data contains learning outcomes measured before the late course drop deadline, and the testing data contains the rest. The late course drop deadline was considered a threshold to split the dataset because often students want to make their decision about dropping a course before this deadline. Instructors need to know this information as well so that they can facilitate a conversation with students about dropping a course. We evaluate the performance of our model by comparing the mean squared errors (MSE) and classification accuracy of proficiency level and category to baseline models. The baseline models include the L-Beta-IRT without the AR component (L-Beta-IRT), the persistence model (PM), and a growth generalized linear mixed effect model (G-GLME). PM is a simple forecasting model that assumes that the prediction for a future timepoint is the same as the previous timepoint, $\hat{y}_{ijt} = y_{ijt-1}$. G-GLME models the individual proficiency level (1 = Proficient, 0 = Not Proficient) for each learning outcome using the same regression model used to model θ_{ijt} in L-Beta-IRT models.

RESULTS

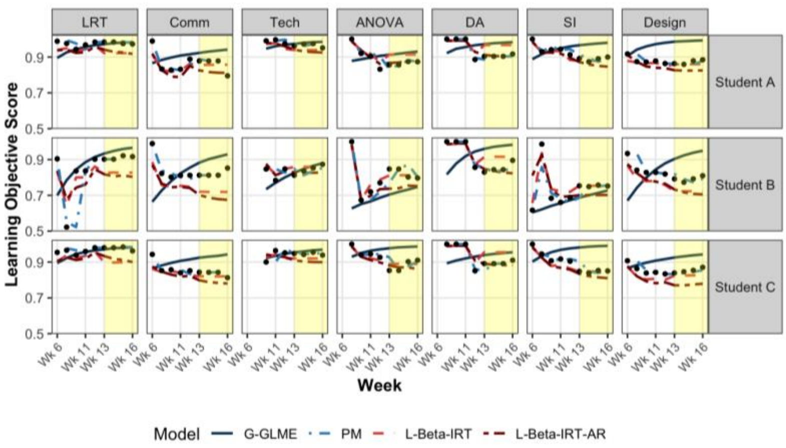
We fit L-IRT-Beta-AR, L-IRT-Beta, and G-GLME using the Stan software with No-U-Turn Sampler (NUTS) to sample the parameters. The PM predictions were generated using base R functions. All experiments are performed in R 4.6.0 and RStudio 2026.4.0.526 on a MacBook Pro with an M4 chip and 16 GB RAM. All learning outcomes were converted to decimal points, and we use a binary response variable (1 = Proficient, 0 = Not Proficient) for G-GLME. The prediction of learning outcome by G-GLME is made based on the probability of obtaining a Proficient level.

Table 2: MSE comparisons of learning outcome (left) and learning objective (right) predictions on training and testing data between models.

Model	**Train Data**	**Test Data**
G-GLME	0.0488, 0.0297	0.0345, 0.0105
PM	0.0105, 0.0049	0.0051, 0.0014
L-Beta-IRT	0.0054, 0.0026	0.0341, 0.0104
L-Beta-IRT-AR	0.0051, 0.0032	0.0178, 0.0073

Table 2 shows the MSE of learning outcome and objective predictions generated by each model for both training and testing data. Each learning objective score is computed by taking the average of the predicted learning outcomes. We observe that G-GLME has the worst performance overall. This makes sense as the probability of obtaining a Proficienct level may not best represent both the learning outcome and objective scores in the dataset. PM has the lowest MSE for testing data, indicating that learning outcome scores do not change for most students over time. Lastly, L-Beta-IRT-AR shows a significant decrease in the MSE of testing data compared to L-Beta-IRT. By adding an AR component to our proposed model, we validate our previous assumption that a student's ability is influenced by their previous ability, and that this AR component plays a crucial role in improving prediction. Moreover, this result is illustrated by Figure 1, which shows the trajectories of randomly selected students' learning objectives.

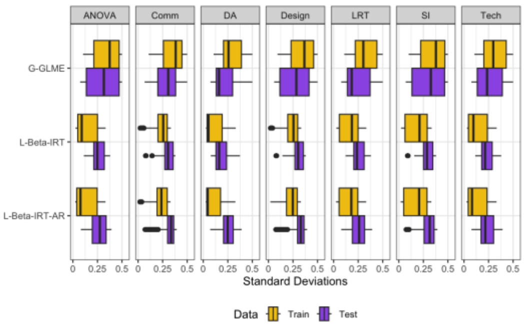
Figure 1: Model predictions of learning objective scores over time



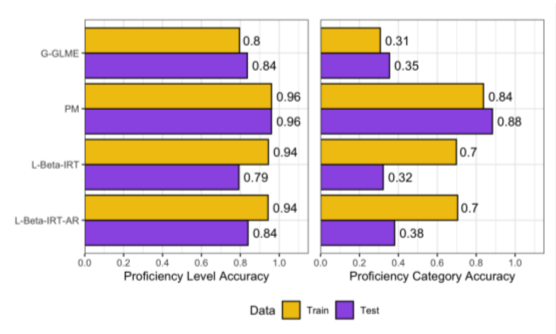
Note. Each row represents a randomly selected student (A–C), and each column corresponds to a learning objective. Observed learning objective score is indicated by black points (•), with yellow regions marking the testing data.

Figure 2(a) shows the standard deviation of the posterior draws of learning outcome predictions for all models except for PM. Although L-Beta-IRT-AR shows a lower MSE, it has about the same standard deviation across all learning objectives as L-Beta-IRT. The predictions of testing data show higher standard deviations than the training data for both L-Beta-IRT and L-Beta-IRT-AR. On the other hand, the standard deviation for G-GLME was similar between the training and testing data. Yet, the standard deviation of G-GLME overall was higher than that of our proposed models.

Figure 2: Model comparisons of (a) standard deviations of posterior draws for learning outcome predictions, and (b) classification accuracy for predicting proficiency level and category.



(a)



(b)

Next, we look at the accuracy of how each model can classify students' learning outcomes to the right proficiency level and category. Based on the learning outcome predictions, we classify the data into proficiency levels and categories using the thresholds in Table 1. Figure 2(b) shows the accuracy of proficiency level and category classifications. We observe that PM has the highest accuracy overall. Besides PM, the other three models seem to have similar accuracy. Furthermore, L-Beta-IRT-AR shows higher accuracy in both proficiency categories and levels with testing data compared to both L-Beta-IRT and G-GLME. Our proposed models have much lower accuracy in predicting proficiency category than proficiency level. This is because Beta-IRT cannot predict discrete values at 0 and 1, so no predictions are classified to either the Adept or Not Shown categories.

CONCLUSION

The proposed L-Beta-IRT-AR model can accurately predict scores of learning outcomes over time, providing frameworks designed to help instructors implement SBG in their statistics classroom. Our model resolves the issues in implementing SBG by enabling instructors to monitor student progress in each learning outcome and predict corresponding letter grades at any point. Instructors can use the information provided by the model to better facilitate a meeting with students to deliver timely feedback on their learning progress. We plan to conduct a qualitative study to examine how instructors translate the information generated from the model into formative feedback. Our goal is to combine the results from the model and qualitative study to provide a trustworthy tool that helps instructors to fully leverage the benefits of SBG in their own classroom.

Although PM appears to achieve better performance overall, our L-Beta-IRT models provide richer and more interpretable information about how students are learning and developing over time for each different learning outcome. Such information will help instructors to expedite the feedback process. Using the model, we suggest the following steps to facilitate the conversations with students:

1. Identify students who are showing struggles in learning outcome(s).
2. Facilitate a conversation with the student and provide resources to guide their success.

In Step 1, the instructors can identify the students by looking at their latent ability regression parameters and the predicted learning outcome scores. If a student has a lower individual intercept, α_i, and a random effect, $\alpha_{j(i)}$, than the averages for a particular learning outcome, then this indicates that the learning outcome is the skill that the student may be struggling with and needs to spend more time on. Although G-GLME has those parameters, they describe the change in observed proficiency levels, which may not represent their latent ability as generated by our L-Beta-IRT models. Additionally, if the predicted learning outcome score is below the threshold for the Not Yet level, then the student may be at risk of failing. In Step 2, the instructors can use and share this information from Step 1 with their students to give individual feedback and provide resources to help them achieve their learning goals. Through one-on-one conversations with the students, instructors should clearly communicate how the numbers from the model reflect students' underlying understanding of each learning outcome to ensure students remain well-informed about their learning progress. It is important that during this conversation, instructors give resources (e.g., practice problems, reading, etc.) to students so they can effectively bridge any learning gaps.

Unlike the baseline models, our L-Beta-IRT can model the latent ability, which should be used to predict their letter grades. Since our dataset did not include students' final letter grades, we did not explore this idea in our study. However, in IRT frameworks, latent ability is linked indirectly to their observed learning outcome scores; therefore, it provides a good summary of their proficiency in each learning outcome. Letter grades from traditional grading systems fail to capture and undermine the intellectual ability of a student (Walton & Spencer, 2009). SBG solves this issue by focusing on their grades based on their proficiency level of learning outcomes, and the latent abilities from our L-Beta-IRT models can be used as empirical evidence for students' proficiency in learning outcomes. Instructors can use the latent ability generated from the models with certain proficiency thresholds to better assign letter grades that best capture their learning experience in a course. This aspect of the model should be further investigated by comparing the projected letter grades from latent ability to their actual letter grades.

Besides the positive impacts of our models in the SBG classroom, they should be improved further to provide better accuracy and reliability. The performances of L-Beta-IRT and L-Beta-IRT-AR were worse than PM. Especially, both models did not show good prediction performance in classifying students into the right proficiency category in the future timepoints, which is important information that instructors can provide to students as feedback. Future work should improve the predictive performance of our models by improving their structure. For instance, polytomous IRT models are used to model an item response with multiple categories. Molenaar et al. (2022) implement a logistic graded response model (Samejima, 1973), which is used to model an ordinal response variable, to a Beta-IRT model, allowing for the analysis of continuous item responses bounded between 0 and 1 inclusive (Molenaar et al., 2022). Future studies should consider adding polytomous frameworks to our L-Beta-IRT models to predict learning levels and categories while predicting the continuous learning outcome scores. Another direction that we should consider is to make the models more flexible to accept different types of alternative grading systems. Since many instructors may implement different structures, such as only using the binary proficiency level instead of a continuous score from our dataset, it is important that our models can work with different variations of alternative grading systems.

REFERENCES

Ayers, E., & Junker, B. (2008). IRT Modeling of Tutor Performance to Predict End-of-Year Exam Scores. Educational and Psychological Measurement, 68(6), 972–987. https://doi.org/10.1177/0013164408318758

Brookhart, S. M. (1991). Grading Practices and Validity. Educational Measurement: Issues and Practice, 10(1), 35–36. https://doi.org/10.1111/j.1745-3992.1991.tb00182.x

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. Review of Educational Research, 86(4), 803–848. https://doi.org/10.3102/0034654316672069

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. Annual Review of Statistics and Its Application, 3(1), 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Campbell, R., Clark, D., & OShaughnessy, J. (2020). Introduction to the Special Issue on Implementing Mastery Grading in the Undergraduate Mathematics Classroom. PRIMUS, 30(8–10), 837–848. https://doi.org/10.1080/10511970.2020.1778824

Chamberlin, K., Yasué, M., & Chiang, I.-C. A. (2023). The impact of grades on student motivation. Active Learning in Higher Education, 24(2), 109–124. https://doi.org/10.1177/1469787418819728

Curley, B., & Downey, J. (2024). Implementation of Alternative Grading Methods in a Mathematical Statistics Course. Journal of Statistics and Data Science Education, 32(3), 272–282. https://doi.org/10.1080/26939169.2023.2249956

GAISE College Report ASA Revision Committee. (2016). Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaisecollege_full.pdf

Guskey, T. R., & Link, L. J. (2019). Exploring the factors teachers consider in determining students' grades. Assessment in Education: Principles, Policy & Practice, 26(3), 303–320. https://doi.org/10.1080/0969594X.2018.1555515

Hatfield, N. J. (2021, June 28). Mastery Based Grading in Upper-level Statistics Courses. 2021 United States Conference on Teaching Statistics. https://www.causeweb.org/cause/uscots/uscots21/th-17-mastery-based-grading-upper-level-statistics-courses

Molenaar, D., Cúri, M., & Bazán, J. L. (2022). Zero and One Inflated Item Response Theory Models for Bounded Continuous Data. Journal of Educational and Behavioral Statistics, 47(6), 693–735. https://doi.org/10.3102/10769986221108455

Nilson, L. B. (2023). Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time. Routledge. https://doi.org/10.4324/9781003447061

Noel, Y., & Dauvier, B. (2007). A Beta Item Response Model for Continuous Bounded Responses. Applied Psychological Measurement, 31(1), 47–73. https://doi.org/10.1177/0146621605287691

Park, J. Y., Dedja, K., Pliakos, K., Kim, J., Joo, S., Cornillie, F., Vens, C., & Van den Noortgate, W. (2023). Comparing the prediction performance of item response theory and machine learning methods on item responses for educational assessments. Behavior Research Methods, 55(4), 2109–2124. https://doi.org/10.3758/s13428-022-01910-8

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. Journal of Educational Psychology, 103(3), 683–700. https://doi.org/10.1037/a0023911

Samejima, F. (1973). Homogeneous case of the continuous response model. Psychometrika, 38(2), 203–219. https://doi.org/10.1007/BF02291114

Walton, G. M., & Spencer, S. J. (2009). Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students. Psychological Science, 20(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.x

Wang, C., & Nydick, S. W. (2020). On Longitudinal Item Response Theory Models: A Didactic. Journal of Educational and Behavioral Statistics, 45(3), 339–368. https://doi.org/10.3102/1076998619882026

Wiggins, G. (1994). Toward better report cards. Educational Leadership, 52(2), 28–37. https://www.ascd.org/el/articles/toward-better-report-cards