Facilitating Student-Instructor Conversation with Bayesian Item Response Theory

ICOTS 12

Emi Cervantes

Penn State University, USA

Department of Statistics

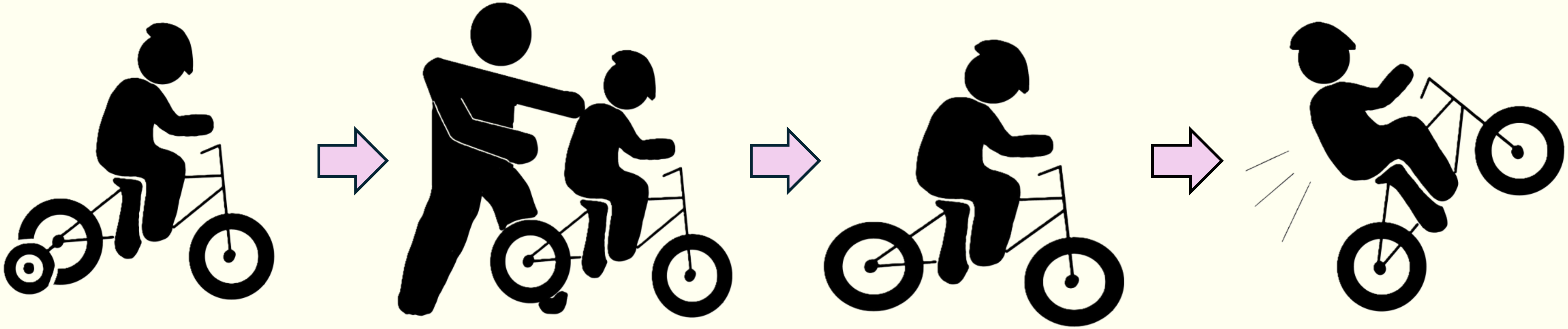
PennState

International Conference on Teaching Statistics, 2026

Rethinking Grades: Alternative Grading Approach

Alternative grading approach (e.g., Mastery, Specifications, Standards-based grading) focuses on assessing students' proficiency on learning outcomes for better feedback.

Rethinking Grades:
Standards-Based Grading (SBG)



Not Shown / Beginning

Progressing

Skillful

Highly Developed/ Adept

Standards-Based Grading in a University Statistics Course

Benefits

- More focus on mastering learning outcomes rather than points
- Transparent expectations for students
- Multiple opportunities to demonstrate understanding
- Supports a growth-oriented mindset

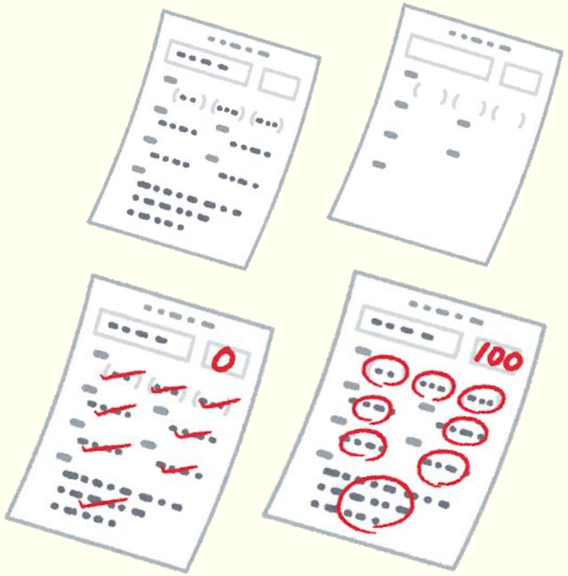
Challenges

- Requires more instructor planning and effort
- Multiple grading and regrading sessions delay feedback
- Students won't know their final letter grades until the end of term
- Students may take time getting used to the grading system

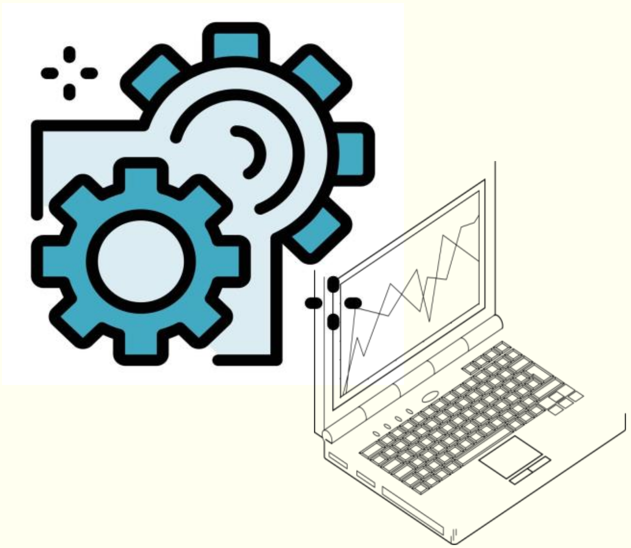
Curley, B., & Downey, J. (2024). Implementation of Alternative Grading Methods in a Mathematical Statistics Course. *Journal of Statistics and Data Science Education*, *32*(3), 272–282. https://doi.org/10.1080/26939169.2023.2249956

Standards-Based Grading is hard to implement!

Problem: Standards-based grading improves feedback, but implementation challenges prevent it from being used to its full potential.



Assessment (Grades on homework, assignments, exams, etc.)



A statistical model to track and predict students' learning outcomes

Better facilitate conversation with students using information generated from the statistical model

Longitudinal Item Response Theory (L-IRT) Model

The **Longitudinal Item Response Theory (L-IRT)** model analyzes and evaluates characteristics of assessment items and their relationship with the latent ability.

Item response data ($i = 1, \ldots, N$) of students ($j = 1, \ldots, J$) collected over time ($t = 1, \ldots, T$)

Item information: δ_j^t (item difficulty)

Latent ability: θ_{ij}^t

Probability of a successful response: $P\left(Y_{ij}^t \text{ is correct} | \delta_j^t, \theta_{ij}^t\right)$

Goal: Use L-IRT to predict students' proficiency of each learning outcome over time.

Application to SBG-based ANOVA Course

7 learning objectives, each with various number of learning outcomes assessed across 9 different timepoints in a semester (N = 74)

- **LRT**: Statistical literacy reasoning and thinking (5 learning outcomes)
- **Comm**: Communication (6 learning outcomes)
- **Tech** (5 learning outcomes)

⋮

! Not all learning outcomes were measured at every single assessment time

LRT: Students will develop their statistical literacy, reasoning, and thinking within the context of ANOVA

LRT.1: The student will learn to seek out the context for a given problem, research question, project, etc.

LRT.2: The student will learn to discuss the role that Statistics plays in life.

LRT.3: The student will learn to analyze a situation for a statistical research question.

LRT.4: The student will learn to attribute approaches to underlying philosophical positions.

LRT.5: The student will learn to attribute results to either experimental or observational designs.

Example of learning objective and learning outcomes

Application to SBG-based ANOVA Course

For each assessment time:

- Each assessment item is mapped to one or more learning outcomes
- Each item receives a discrete, proficiency score for a single learning outcome based on rubric
- Learning outcome score = average of relevant item scores
- Learning objective score = average of relevant learning outcome scores

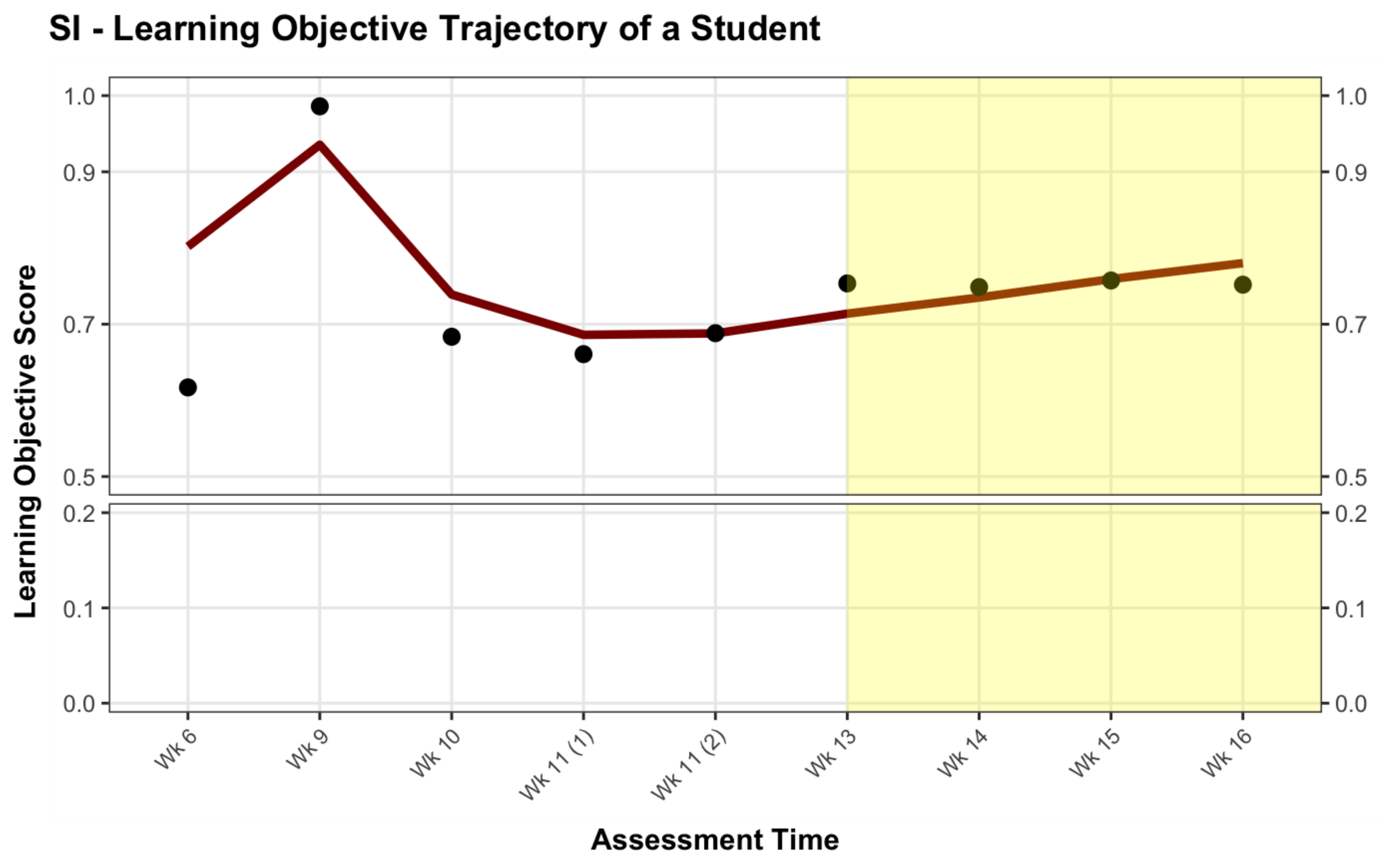
Table 1: Proficiency levels and their three categories used in the SBG-based ANOVA course with score intervals and discrete item scores.

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%

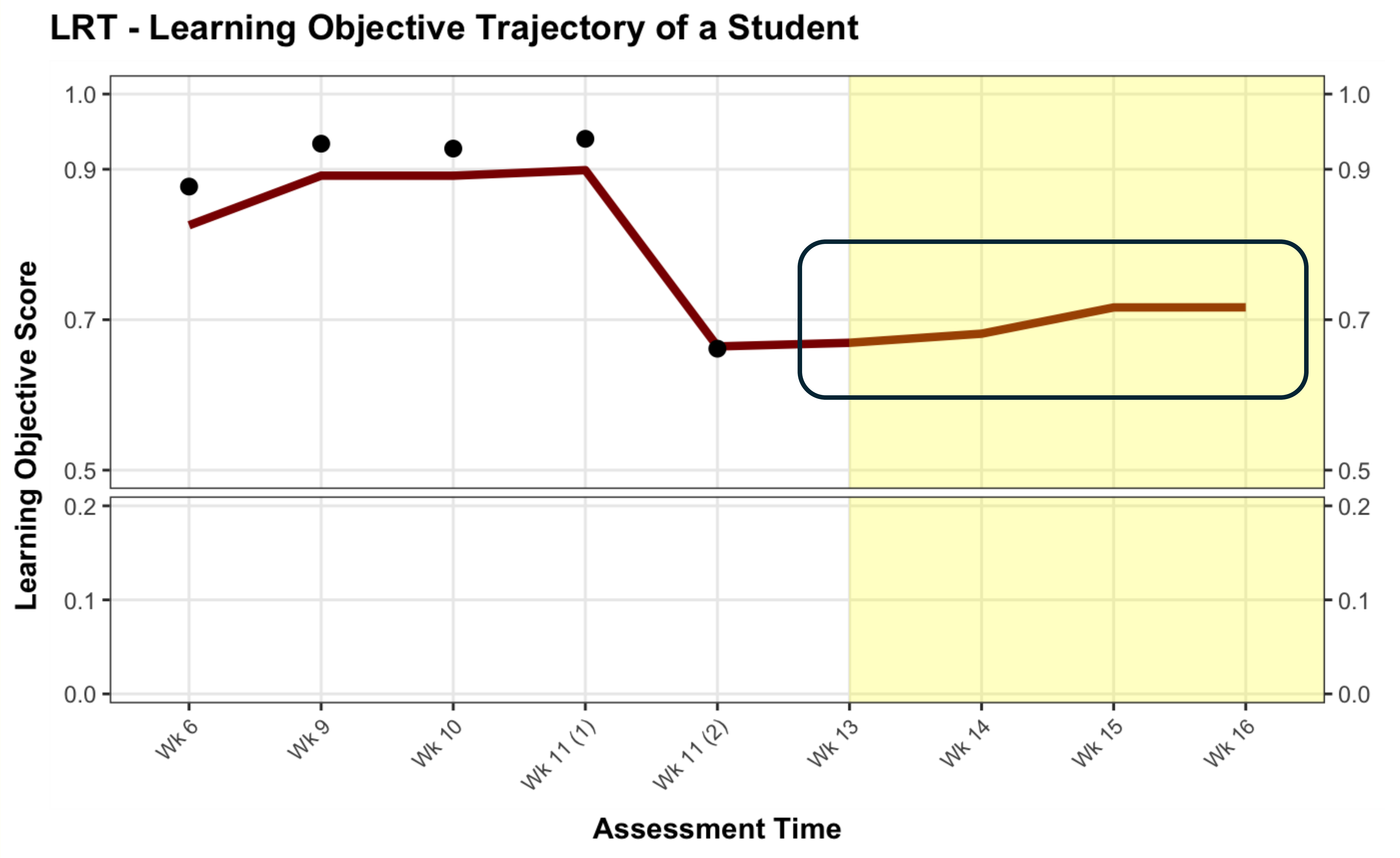
Application to SBG-based ANOVA Course

For an individual student, the model predicts learning growth for each learning outcome.

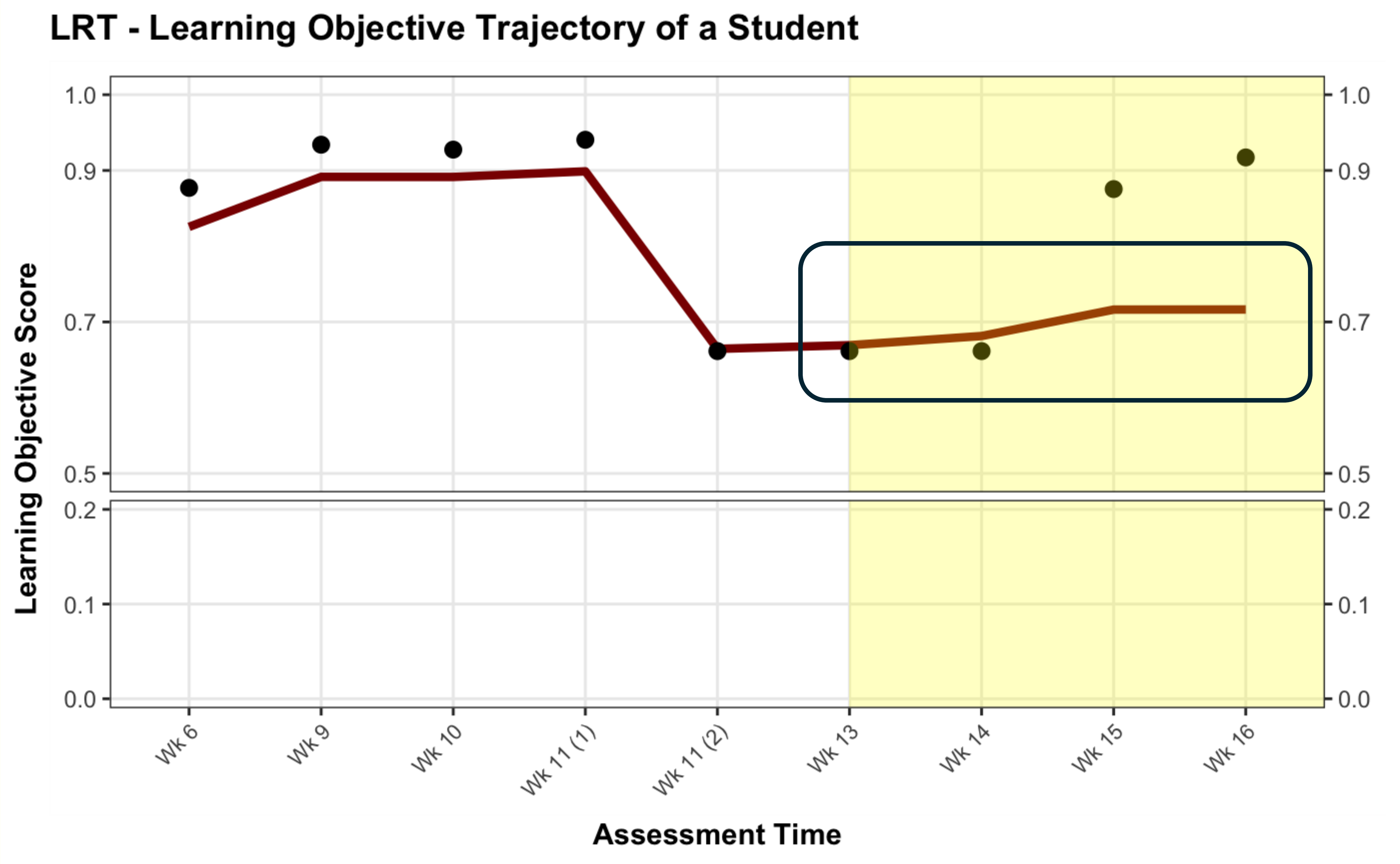
Yellow-highlighted region represents data after the *late drop deadline*, last day to drop a course during a semester at Penn State (usually occurs during week 13-14 during the 16-week semester).



Application to SBG-based ANOVA Course



Application to SBG-based ANOVA Course



L-IRT Workflow in SBG Classroom

Step 1 – Identify students who are showing struggles in learning outcome(s).

- Instructors can identify the students by looking at their latent ability and its parameters generated from the models.
- With the predicted learning outcome scores, instructors can calculate their projected letter grades at any point in the course,

Step 2 – Facilitate a conversation with the student and provide resources to guide their success.

- Share the information from Step 1 with students to give individual feedback and resources to help achieve their learning goals.

What's Next? Research and Classroom Impact

Toward Practical and Transparent Assessment

- Develop an interpretable model to predict learning objectives and course performance.
    - Use latent ability to predict student's underlying proficiency level and letter grade.
- Create an instructor-friendly implementation (e.g., LMS).
- Evaluate classroom impact through instructor and student feedback.

Key Questions for Future Works

- *Instructors: How does the tool assist instructors in adopting and managing standards-based grading practices?*
- *Students: How does the tool help students better understand their performances in the course? How does it affect their learning experience?*

Research Team

Sam Baugh
Assistant Professor
Penn State University
Department of Statistics

Neil Hatfield
Assistant Research Professor
Penn State University
Department of Statistics

ICOTS 12

PennState

Links to paper and slides

Thank you!

Emi Cervantes · *emic@psu.edu*
Penn State (USA), Department of Statistics
PhD Student

Appendix

LRT: Students will develop their statistical literacy, reasoning, and thinking within the context of ANOVA

LRT.1: The student will learn to seek out the context for a given problem, research question, project, etc.

LRT.2: The student will learn to discuss the role that Statistics plays in life.

LRT.3: The student will learn to analyze a situation for a statistical research question.

LRT.4: The student will learn to attribute approaches to underlying philosophical positions.

LRT.5: The student will learn to attribute results to either experimental or observational designs.

Comm: Students will develop their communication skills alongside their statistical understandings so that they can present their work to peers, superiors, and clients in effective ways.

Comm.1: The student will learn to summarize the context of the research inquiry.

Comm.2: The student will learn to convey how a research question fits within the ANOVA frameworks.

Comm.3: The student will learn to interpret the values of statistics (both descriptive/incisive and inferential) within the current context.

Comm.4: The student will learn to meaningfully discuss data visualizations to support others in their learning about the current context.

Comm.5: The student will learn to assess and constructively critique the work of others to improve both his/her work and the work of others.

Comm.6: The student will learn to generate a report that tells the data's story, incorporating visualizations and statistics, and provides a basis for making a decision for answering the research question.

Design: Students will develop their skills for developing experimental and observational studies.

Design.1: The student will be able to classify attributes as being a response, factor, block, or covariate for a given study context.

Design.2: The student will be able to differentiate between measurement units and experimental units.

Design.3: The student will be able to differentiate between various ANOVA models.

Design.4: The student will be able to recommend what ANOVA models is most appropriate for a given study.

Design.5: The student will be able to produce a visual representation of underlying model for a given situation.

Design.6: The student will be able to plan out an experimental or observational study to answer a given statistical research question.

Design.7: The student will be able to recommend a potential sample size given choices for Type I and Type II risks, number of groups, and effect size.

ANOVA: Students will develop their knowledge and understanding of ANOVA models and their applicability to real-word statistical research questions.

ANOVA.1: The student will learn to transform a research question into testable hypotheses within ANOVA contexts.

ANOVA.2: The student will learn to interpret the terms in various ANOVA models both in a specific context and in general.

ANOVA.3: The student will learn to use ANOVA screens for one-way and one-way with blcoking designs to construct a model.

ANOVA.4: The student will learn to differentiate between fixed and random effects.

ANOVA.5: The student will learn to provide examples of ANOVA models/designs.

ANOVA.6: The student will learn to interpret the estimates of model terms within the present context.

DA: Students will develop their skills in data analysis to describe what the data have to say and explain what could be going on in a particular context

DA.1: The student will learn to differentiate between attributes that are necessary to answer a research question and those that are not.

DA.2: The student will learn to identify possible sources of variation in a particular context.

DA.3: The student will learn to analyze data visualizations.

DA.4: The student will learn to differentiate between attributes of a collection and statistcs that mesure those attributes.

DA.5: The student will learn to recommend which statistics(s) would be the bewst to use for emasureing certain attributes.

DA.6: The student will learn to manipulate data sets in preparation for analysis.

SI: Students will develop their skills in using the tools of statistical inference to make decisions

SI.1: The student will learn to differentiate between four types of inference errors.

SI.2: The student will learn to recommend a Type I Error Rate to control for, using what method, and at what level.

SI.3: The student will learn how to check his/her assumptions when carrying out hypothesis tests.

SI.4: The student will learn to distinguish between productive and problematic interpretations of p-values.

SI.5: The student will learn to distinguish between productive and problematic interpretations of confidence intervals.

SI.6: The student will learn to appraise effect sizes given the context.

SI.7: The student will learn to synthesize the results of a hypothesis test in a report.

Tech: Students will develop their skill with at least one statistical software package so that they can use technology to handle computations, allowing them to focus on designs, interpretations, and communication

Tech.1: The student will learn to use technology to create data visualizations.

Tech.2: The student will learn to use technology to perform calculations on data sets.

Tech.3: The student will learn to use technology to analyze data to answer research questions.

Tech.4: The student will learn to design a method to get to sampling distribution for his/her selected estimator.

Tech.5: The student will learn to use technology to his/her advantage when engaged in data analysis.

Longitudinal Item Response Theory (L-IRT) Model – Latent Ability

We assume that every student has **positive growth** in their ability over time, so the coefficient representing growth over time is constrained to be strictly positive.

$$\theta_{ij}^{t} = \beta + Z_{ij}^{t} \upsilon_i + \rho\theta_{ij}^{t-1} + \epsilon_{ij}^{t}$$

Proficiency Level of *i*-th student for learning outcome *j* at timepoint *t*

Fixed Intercept

Random Component

- Student-specific intercept
- Student- and learning-objective–specific growth over time

AR(1) Component
Autoregressive model of order 1 to model current ability on previous ability.

Residuals

Longitudinal Item Response Theory (L-IRT) Model – Latent Ability

We assume that every student has **positive growth** in their ability over time, so the coefficient representing growth over time is constrained to be strictly positive.

$$\theta_{ij}^{t} = \beta + \underbrace{\alpha_i + \alpha_{j(i)} t}_{} + \rho\theta_{ij}^{t-1} + \epsilon_{ij}^{t}$$

Proficiency level of *i*-th student for learning outcome *j* at timepoint *t*

Fixed Intercept

Random Components
- Student-specific intercept
- Student- and learning-objective–specific growth over time

AR(1) Component
Autoregressive model of order 1 to model current ability on previous ability.

Residuals

where $\alpha_{j(i)} \geq 0$ and $\epsilon_{ij}^{t} \sim \mathcal{N}(0, \sigma^2)$

Model Formulation

At each time point, student responses to each item are assumed to follow a Beta distribution: $Y^t_{i,j}|m^t_{i,j}, n^t_{i,j} \sim \text{Beta}(m^t_{i,j}, n^t_{i,j})$, where $m^t_{i,j}$ and $n^t_{i,j}$ are positive reals which are often interpreted as "acceptance" and "refusal" parameters, respectively.

$$P(Y^t_{i,j}|m^t_{i,j}, n^t_{i,j}) = \frac{\Gamma(m^t_{i,j} + n^t_{i,j})}{\Gamma(m^t_{i,j})\Gamma(n^t_{i,j})} y_{i,j}^{m_{i,j}-1}(1 - y_{i,j})^{n^t_{i,j}-1}. \tag{1}$$

$$P(Y_{ij}|m_{i,j}, n_{i,j}) = \prod_{t=1}^{T} P(Y^t_{i,j}|m^t_{i,j}, n^t_{i,j}). \tag{2}$$

The parameters of the beta distribution can be represented as exponential functions of person and item parameters,

$$m^t_{i,j} = \exp\left\{\frac{\theta^t_{i,j} - \delta^t_j}{2}\right\}, \quad n^t_{i,j} = \exp\left\{-\frac{\theta^t_{i,j} - \delta^t_j}{2}\right\}. \tag{3}$$

CV Experiment Results

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

Model	Training Data	Testing Data
G-GLME	0.0488, *0.0297*	0.0051, *0.0032*
Persistent Model (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*

G-GLME:

$$Y_{ij}^t \sim \text{Bernoulli}(\eta_{ij}^t)$$

$$\eta_{ij}^t = \beta_0 + \alpha_i + \alpha_{j(i)} t + \epsilon_{ij}^t$$

$$\text{where } \alpha_{j(i)} \geq 0 \text{ and } \epsilon_{ij}^t \sim \mathcal{N}(0, \sigma^2)$$

Persistent Model:

$$Y_{ij}^t = Y_{ij}^{t-1}$$

CV Experiment Results