# Soft Grades: A Calibrated and Accurate Method for Course-Grade Estimation that Expresses Uncertainty

Juliette Woodrow
Stanford University
Stanford, USA
jwoodrow@stanford.edu

Chris Piech
Stanford University
Stanford, USA
piech@cs.stanford.edu

## Abstract

In traditional educational settings, students are often summarized by a single number—a final course grade—that reflects their performance. While final grades are convenient for reporting or comparison, they oversimplify a student's true ability and do not express uncertainty. In this paper, we introduce a new item-response model for classroom settings that infers a distribution over student abilities and uses this to represent each student's final grade as a probability distribution. This approach captures the uncertainty that comes from variations in both student performance and grading processes. Practical applications of our approach include enabling teachers to better understand grading confidence, impute missing assignment scores, and make informed decisions when curving final grades. For students, the model offers probabilistic estimates of their final course grades based on current performance, supporting informed academic decisions such as opting for Pass/Fail grading. We evaluate our model using real-world datasets, showing that the Soft Grades model is well-calibrated and surpasses the state-of-the-art polytomous IRT model in accurately predicting future scores. Additionally, we share a web application and Python scripts to make our model available to teachers and students.

## CCS Concepts

• **Computing methodologies → Model verification and validation**; **Model development and analysis**.

## Keywords

Item Response Theory, Grade Prediction, Soft Grades, Ability Inference

**ACM Reference Format:**
Juliette Woodrow and Chris Piech. 2025. Soft Grades: A Calibrated and Accurate Method for Course-Grade Estimation that Expresses Uncertainty. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706568

## 1 Introduction

Traditional course grades reduce the complexity of student performance to a single number, ignoring the inherent uncertainty in learning and assessment. As educators, we recognize that many factors unrelated to a student's true understanding of the material can affect their performance. In this paper, we introduce a novel item response model and inference method for classroom settings that captures this uncertainty by representing final grades as probability distributions, which we call "Soft Grades." Inspired by advancements in fields like medical testing, weather forecasting, and localization [1–3], where probability distributions have enhanced decision-making, our approach brings a similar shift to education, offering a richer and more informative grading system.

Consider a real student in a university level course who has completed 6 homework assignments and 2 exams. Her teacher has calculated her final grade as 90%. This 90% hides any expression of uncertainty in her final grade. For this student, our model articulates that there is a large amount of uncertainty, as shown in figure 1. For example, there is a non-zero probability that the student deserves a grade that lies outside the 85% to 95% range. The right side of figure 1 illustrates the distribution of standard deviations in "Soft Grades" for an entire introductory computer science course at an R1 university. This distribution reveals that while some students have low variance—indicating a high confidence—many students have higher variance, suggesting that their given grade may not fully capture their true understanding. If the final grade is a critical summative assessment, the teacher should be aware of this uncertainty and consider whether more data is needed. This richer representation offers teachers and students a more precise and actionable understanding of student outcomes. Teachers can use "Soft Grades" to better assess the confidence that they should have in each student's grade, improve decisions on grade curving, and impute missing grades with greater accuracy. "Soft Grades" offers students a clearer picture of their performance, enabling them to make more informed choices—such as whether to opt for Pass/Fail or how to best allocate study time across courses.

But how do we know that the uncertainty in the student's soft grade (shown in figure 1) is accurate? We show that our model is both more predictive of future scores than state-of-the-art polytomous IRT and that it is well calibrated. This means that when the model predicts an event with a probability of 70%, that event will occur about 70% of the time, validating the accuracy of the uncertainty in Soft Grades.

### 1.1 Main Contributions

(1) We develop a novel item-response model that represents grades as probability distributions. We propose a new inference technique for Item Response Theory (IRT) that enables learning a full probability distribution over student ability, rather than traditional point estimates. To the best of our



*LAK 2025, March 03–07, 2025, Dublin, Ireland*

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706568

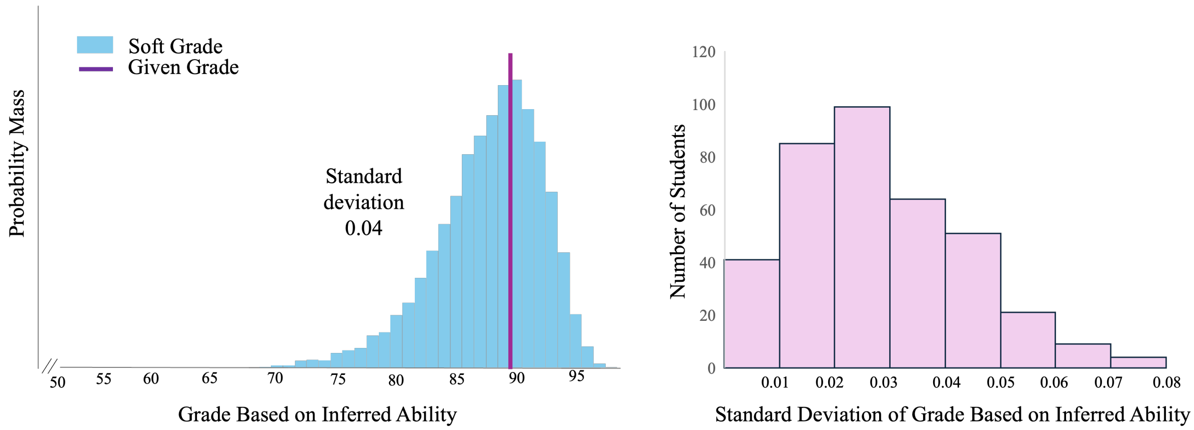



**Figure 1: Left: The "Soft" Grade for a single student, based off their inferred ability, after observing all of their course assessment grades. Even after 6 psets and 2 exams there is a lot of uncertainty as to the grade that reflects their ability. This uncertainty is inherent in the grading process (variance in student performance and grader accuracy) but is not represented by teachers in traditional classrooms. Right: the distribution of uncertainty (standard deviation) of soft grades given to students in a real class of 374 students.**

knowledge, this is the first approach that comprehensively expresses uncertainty inherent in course grades.

(2) We validate our model using real-world datasets, demonstrating that the Soft Grades model outperforms the state-of-the-art polytomous IRT model in imputing missing grades and predicting future scores. We also show that our model is well-calibrated.
(3) We present practical applications for teachers, such as representing student grades as "Soft Grade" distributions, which show the likelihood of each potential grade a student may deserve. This approach provides insights into grading confidence, assists in imputing missing assignment scores, and aids in curving final grades.
(4) We show how students can benefit from our model by receiving probabilistic estimates of their final course grades based on their current performance, helping them make informed academic decisions.
(5) We present a free web application that implements our model, making it accessible for both teachers and students.

## 2 Soft Grade Problem

Formally, the soft grade problem is to produce a prediction of a student's final grade as a probability distribution $G$. The prediction $G$ is defined by a probability mass function $P_G$. For any grade $x \in \{0, 1, 2, \ldots, 100\}$, $P_G(x)$ should return the probability, according to the model's estimate, that the student will get the final grade $x$ in the class.

The soft grade prediction is based on the grades that the student has received in the class: $g_1, g_2, \ldots, g_n$, where $n$ represents the number of assessments the student has completed. Additionally, to make a grounded prediction, information about the scores of other students on these $n$ assessments is also necessary.

We present two variants of the problem for calculating the soft grade prediction: one from the teacher's perspective and one from the student's perspective. We formalize the inputs to the soft grade problem as follows:

- Student's observed scores: $g_1, \ldots, g_n$ for all $n$ assessments in the course
- Assessment characteristics (two variants):
  (1) Instructor's perspective: $S_j$ (set of all students' scores) for each assessment $j$
  (2) Student's perspective: mean score and standard deviation of scores for each assessment $j$

A good Soft Grade prediction should be accurate and well-calibrated. A calibrated model ensures that the predicted probabilities align with the actual frequencies of outcomes. For instance, if the Soft Grades Model predicts a 70% chance of a student achieving a grade of 80 or higher, calibration ensures that, across all students for whom the model predicts a 70% probability of achieving 80 or higher, approximately 70% of them actually score 80 or higher. Accuracy measures how closely the model's predictions match the true final grades. We evaluate our Soft Grades model by comparing it to traditional grading baselines and a Continuous Response Model from Item Response Theory, using both calibration and accuracy metrics. This approach assesses how well the model quantifies uncertainty and accurately predicts final grades, key factors for its practical use in real-world classrooms.

### 2.1 Downstream Impacts of Soft Grades

The main applications of Soft Grades will be discussed in detail later, but we introduce them here to provide context for the following sections. For teachers, the Soft Grades model can assist in imputing missing assignment scores, offering a more nuanced alternative to simply dropping the grade. It also provides insight into how confident they can be in the grades they assign, informing decisions about curving grades and assigning grades to students on the borderline between two grade levels. For students, Soft Grades offers personalized predictions of their final grade, allowing them

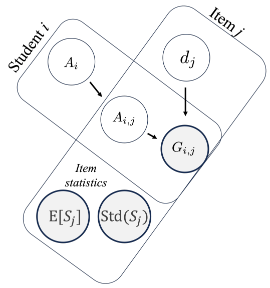



**Figure 2: Graphical model for CGRT (Student Perspective). The teacher perspective would have the entire set of scores $S_j$ for item statistics. Observed variables are shaded. Arrows represent dependency between random variables, and each rectangle represents a plate (i.e., repeated observations).**

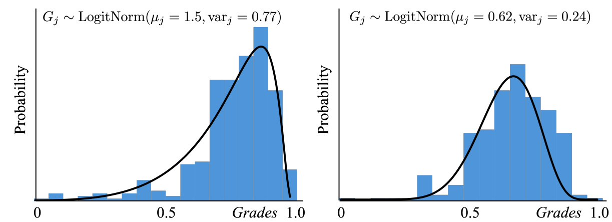



**Figure 3: Histograms of exam grades (normalized to be between 0 and 1) from two different R1 university courses, with the best-fitting logit-normal distributions overlaid. These examples highlight how logit-normals can effectively model grades.**

to explore different scenarios for upcoming assessments and make informed decisions about studying, grade targets, and whether to opt for Pass/Fail or letter grades.

## 2.2 Related Work

Item Response Theory (IRT) has been a key framework in educational assessment. It models the probability of a correct response based on a student's latent ability and item characteristics such as difficulty. Foundational models like the Rasch Model primarily focus on binary outcomes (correct/incorrect), while polytomous extensions handle multiple ordered categories or partial credit scenarios [4, 5]. Continuous Response Models (CRM) further extend IRT to continuous outcomes by modeling probability density functions [6–9].

However, traditional IRT models predominantly operate at the question level, requiring detailed item-level data that may not always be available in real-world classroom settings. Moreover, most IRT inference techniques, such as Maximum Likelihood Estimation (MLE) or Expectation-Maximization (EM), provide single-point estimates of latent traits like ability. These methods fail to explicitly capture the uncertainty inherent in student performance and grading processes. Although rare, Bayesian inference techniques have been used in IRT, primarily for analyzing multiple-choice test scores [10], but to the best of our knowledge, they have not been applied to course grades.

In addition to IRT models, various predictive models have been developed to estimate student performance over time [11–21]. These models typically track skill acquisition using machine learning or theoretical approaches. They often focus on mastery of specific skills based on practice performance. These models are often designed for controlled environments and may not fully capture the complexities of real-classroom settings, where assessments involve multiple skills and are influenced by factors such as stress, grading inconsistencies, and cumulative evaluations.

We propose a novel approach that extends IRT by incorporating uncertainty into the inference process. Unlike traditional models that focus on individual questions as items, our method considers entire assessments and ultimately computes final course grades, offering a more comprehensive representation of real-world classroom settings.

# 3 Methodology

Having defined the Soft Grades problem and its evaluation criteria, we now present our solution: Course-Grade Response Theory, a statistical model that links student abilities, assessment parameters, and observed scores to generate Soft Grades. To implement this approach, we follow three steps: first, we estimate the model parameters for each assessment; second, we infer each student's ability based on their performance; and third, we use the inferred ability to generate Soft Grades. The following sections detail each of these steps, demonstrating how we move from theoretical foundations to practical predictions.

## 3.1 Course-Grade Response Theory

Course-Grade Response Theory (CGRT) extends Item Response Theory (IRT) by modeling student performance across entire assessments and computing a probability distribution over the final course grade for each student. An assessment refers to any academic evaluation in a course, such as an assignment, quiz, or exam. The CGRT model accounts for inherent uncertainties in both student performance and grading processes. To build this framework, CGRT uses parameters to represent important features of both assessments and students. Each assessment $j$ has a difficulty parameter $d_j$, which measures how challenging that assessment is compared to others. Similarly, each student $i$ has a true ability level, $A_i$, which reflects their overall proficiency in the course material. We assume that $A_i$ follows a normal distribution, as shown in equation 8, with mean 0 and variance $\sigma_a^2$. The parameter $\sigma_a$ represents our prior belief about the range in student abilities across the class. A lager $\sigma_a$ suggests a more diverse range of abilities in a course, whereas a smaller $\sigma_a$ indicates a group of students with more similar ability levels.

A student's performance can vary from day to day due to factors unrelated to their true ability, such as fatigue, illness, or even grading errors. To account for this, we introduce $A_{ij}$, a noisy ability that represents the performance of student $i$ on the day they completed assessment $j$. $A_{ij}$ is modeled as the student's true ability plus a noise

**Reference 1: Key Variables in CGRT**

| | |
|---|---|
| **Parameters:** | |
| $\sigma_a$ | Standard deviation in student abilities within the class. |
| $d_j$ | Difficulty of assessment $j$. |
| $\epsilon_j^2$ | Variance of noise for assessment $j$. |
| **Random Variables:** | |
| $A_i$ | True ability of student $i$. $A_i \sim N(0, \sigma_a^2)$ |
| $A_{ij}$ | Noisy ability of student $i$ on assessment $j$. $A_{ij} = A_i + M_j$ |
| $M_j$ | Noise term for assessment $j$. $M_j \sim N(0, \epsilon_j^2)$ |
| $G_{ij}$ | Grade of student $i$ on assessment $j$. $G_{ij} = \text{Sigmoid}(A_{ij} - d_j)$ |
| $G_j$ | Distribution of grades for all students on assessment $j$. $G_j \sim \text{Logit-Normal}(-d_j, \sigma_a^2 + \epsilon_j^2)$ |
| **Composites:** | |
| $\text{var}_j = \sigma_a^2 + \epsilon_j^2$ | Variance for assessment $j$. |
| $Z_{ij} = A_{ij} - d_j$ | Difference between combined ability and difficulty for assessment $j$. |
| $g_{i1}, \ldots, g_{in}$ | Scores that student i received on assessments 1 through n |

term $M_j$, as shown in equation 9. $M_j$ captures assessment-specific variability such as fluctuations in performance or inconsistencies in grading, and is assumed to follow a normal distribution with mean 0 and variance $\epsilon_j^2$. The parameter $\epsilon_j$ quantifies the degree of uncertainty introduced by these external factors for assessment $j$. Larger values of $\epsilon_j$ indicate that the assessment is more affected by outside variability, while smaller values mean it is a more consistent and reliable observation of the student's true ability. Introducing noisy ability $A_{ij}$ allows us to account for variations and explicitly capture the uncertainty inherent in a student's performance on any given assessment. This is similar to established practices in psychometrics and statistical modeling where accounting for measurement error leads to more robust inferences [22]. Another example is Elo ratings [23] in competitive gaming, which model a player's observed performance as fluctuating around their true skill due to transient factors.

In academic assessments, a student's performance depends on both their ability and the difficulty of the assessment. The Rasch Model [24], a foundational approach in Item Response Theory (IRT), models the probability of a correct response based on the difference between a student's ability and an item's difficulty. In the Rasch model, the probability $P_{ij}$ that student $i$ answers item $j$ correctly is given by: $P_{ij} = \text{Sigmoid}(\theta_i - d_j)$ where $\theta_i$ is the ability of student $i$, $d_j$ is the difficulty of item $j$, and the sigmoid function (also known as the logistic function): $\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$, maps real numbers to values between 0 and 1. The logistic function converts the difference between the student's ability and the item's difficulty into a probability. For example, if a student's ability ($\theta_i$) is much higher than the item's difficulty ($d_j$), the resulting probability will be close to 1, suggesting that the student is likely to answer the question correctly. Conversely, if the difficulty exceeds the student's ability, the probability will be close to 0, reflecting a low likelihood of a correct response.

In CGRT, we take the same intuitive idea: higher ability compared to difficulty leads to better performance. However, instead of modeling the probability of a correct response on a single question, CGRT extends this approach to handle continuous grades on entire assessments. We use the same principle: squashing the difference between ability and difficulty to a number between 0 and 1 using the logistic function. Instead of interpreting this as a probability, we treat it as a score (a fraction out of 100) that a student receives on an assessment. Additionally, rather than using a student's true ability directly, CGRT incorporates their noisy ability ($A_{ij}$). In CGRT, The grade for student $i$ on assessment $j$ is defined as the result of applying the logistic function to the difference between their noisy ability and the assessment's difficulty, as shown in equation 10. This approach allows CGRT to model grades in a way that aligns naturally with standard grading systems while capturing inherent uncertainties in performance.

The graphical model in Figure 2 illustrates the dependencies in Course-Grade Response Theory. Each student's true ability, $A_i$, is drawn from a normal distribution and influences their noisy ability, $A_{ij}$, on each assessment. $A_{ij}$ and assessment-specific difficulty, $d_j$, affect the grade, $G_{ij}$, that student $i$ receives for assessment $j$. The observed variables are shaded to indicate that they are directly measured, while the latent variables, such as the true ability and assessment noise, remain unobserved but inferred. The arrows between variables represent the dependencies within the model. The lower part of the diagram shows the item statistics, where the mean and standard deviation of the grades for each assessment are used to estimate the parameters $d_j$ and $\epsilon_j^2$, respectively. This plate structure highlights the repeated observations across multiple students and assessments, which are needed to estimate model parameters and infer student abilities.

In summary, we formalize our three key assumptions as:

(1) Student abilities are normally distributed:

$$A_i \sim N(0, \sigma_a^2) \tag{8}$$

(2) Abilities on assessments include noise from daily fluctuations and grading inconsistencies:

$$A_{ij} = A_i + M_j \quad \text{where} \quad M_j \sim N(0, \epsilon_j^2) \tag{9}$$

(3) Grades are modeled using a logistic function, noisy ability, and assessment difficulty:

$$G_{ij} = \text{Sigmoid}(A_{ij} - d_j) \tag{10}$$

**Derivation 1: Grades as Modeled by CGRT Follow a Logit-Normal Distribution**
To understand how grades in CGRT follow a Logit-Normal distribution, we derive this result using the assumptions and equations defined in the model. The derivation starts by substituting $A_{ij} = A_i + M_j$ into equation 10,

$$G_{ij} = \text{Sigmoid}(A_i + M_j - d_j). \tag{4}$$

Since $A_i$ and $M_j$ are independent and normally distributed, their sum is also normally distributed, where the resulting mean is the sum of the means and the resulting variance is the sum of the variances:

$$A_i + M_j \sim N(0, \sigma_a^2 + \epsilon_j^2). \tag{5}$$

Subtracting the assessment difficulty $d_j$, which is a constant, we define:

$$Z_{ij} = A_i + M_j - d_j \sim N(-d_j, \sigma_a^2 + \epsilon_j^2). \tag{6}$$

Substituting $Z_{ij}$ into the grade equation, we derive:

$$G_{ij} = \text{Sigmoid}(Z_{ij}) \sim \text{Logit-Normal}(-d_j, \sigma_a^2 + \epsilon_j^2). \tag{7}$$

When a normally distributed variable is transformed by the sigmoid (logistic) function, the resulting distribution is a Logit-Normal. This result shows that when we consider the distribution of grades across all students for a particular assessment $j$, the grades follow a Logit-Normal distribution characterized by mean $-d_j$, reflecting assessment difficulty, and variance $\sigma_a^2 + \epsilon_j^2$, combining the variability from student abilities and assessment noise. This aligns with prior research and empirical observations indicating that assessment grades often follow distributions resembling the Logit-Normal.

To initially validate these assumptions, we show in Derivation 1 that they lead to the claim that grades on assessments are distributed as Logit-Normals, which aligns with prior research and empirical observations [25].

The model's ability to reproduce this characteristic distribution initially supports the validity of Course-Grade Response Theory. With this understanding of how grades are distributed according to our model, we can proceed to estimating the model parameters.

## 3.2 Estimating Model Parameters

In CGRT, we first estimate all parameters, then given estimated parameters we can compute soft grades for each student. The key parameters we estimate are: the difficulty $d_j$ of each assessment $j$, the variance in noise $\epsilon_j^2$ for each assessment $j$ and the variance in our prior belief over student abilities $\sigma_a^2$. We outline the parameters in Reference 1.

### 3.2.1 *Estimating student ability parameter,* $\sigma_a$.

We begin by estimating the standard deviation of student abilities, $\sigma_a$. This parameter captures variability in performance levels, enabling the model to scale assessment difficulty and interpret grade distributions accurately. A common approach in Item Response Theory (IRT) is to set $\sigma_a = 1$, assuming a standard normal prior on student abilities [26, 27]. While this assumption is reasonable in many contexts, it can lead to issues in assessments where the variance in scores is small. Through our research, we found that learning $\sigma_a$ for each course significantly improves prediction accuracy and calibration, particularly in cases where assessment scores exhibit low variance.

To better understand a problem with fixing $\sigma_a^2 = 1$, recall that the total variance in assessment scores, $\text{var}_j$, is the sum of student ability variance, $\sigma_a^2$, and assessment noise variance, $\epsilon_j^2$. Now, consider an ideal assessment which has no noise ($\epsilon_j^2 = 0$). If an assessment has no noise, the variance of scores, $\text{var}_j$, is determined entirely by the variance in student abilities, $\sigma_a^2$. This implies that the smallest observed variance across assessments serves as an upper bound for the variance in student abilities. If the minimum observed $\text{var}_j$ is less than 1, setting $\sigma_a^2 = 1$ overestimates ability variance, distorting predictions and leading to poorly calibrated models.

To address this, we propose one method for estimating $\sigma_a$ per course that works well in practice. We assume that the assessment with the smallest observed variance, $\min_j \text{var}_j$, is the least noisy and use this variance as an approximate upper bound on $\sigma_a^2$. Based on this insight, we estimate $\sigma_a$ using the following formula:

$$\sigma_a = \alpha \times \min_j \text{var}_j$$

where $\alpha$ is a hyperparameter between 0 and 1. This hyperparameter reduces the variance attributed to student ability, ensuring that not all of the variance in an assessment is explained by student abilities alone. By doing so, it guarantees that every assessment includes at least some assessment-specific noise. To determine the optimal value of $\alpha$, we performed leave-one-out cross-validation on the OULAD dataset [28]. Our analysis identified $\alpha = 0.833$ as the value that provided the best results. While this one approach works well in practice, more research is needed to explore alternative strategies for learning this parameter.

### 3.2.2 *Estimating the assessment parameters,* $d_j$ *and* $\epsilon_j$.

We present two scenarios for estimating the assessment parameters depending on the user and the data that user would have available. The *Teacher Variant* assumes the user has access to a typical grade book: a score for each student and each assignment. Formally they have student scores $S_j = \{S_{ij}\}_{i=1}^{N_j}$ for each assessment $j$, where $N_j$ is the number of students who completed that assessment. In contrast, the *Student Variant* assumes the user only has access to their own scores on assessments, $g_1, \ldots, g_j$ and summary statistics (mean and standard deviation) of each assessment.

***Teacher Variant:*** In the teacher variant, we have access to the full distribution of student scores for each assessment, $S_j$. In CGRT, we assume each $S_j$ follows a logit-normal distribution. The parameters of a logit-normal distribution are the mean and variance of the underlying normal distribution. Since we have the entire distribution $S_j$, we can estimate its mean and variance directly by applying the logit transformation to all scores, and then calculating the mean and variance of the transformed values:

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} \text{logit}(S_{ij}) \qquad \text{var}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\left(\text{logit}(S_{ij}) - \mu_j\right)^2$$

Once we calculate $\text{var}_j$, we can compute $\sigma_a$ as $\alpha$ multiplied by the minimum variance. From there, we determine each $\epsilon_j^2$ using the equation: $\epsilon_j^2 = \text{var}_j - \sigma_a^2$. Since $\mu_j = -d_j$, we also recover the difficulty parameter $d_j$.

***Student Variant:*** In the student variant, we only have the mean and standard deviation of the scores $S_j$, so we cannot perform the straightforward parameter estimation used in the teacher variant. The logit-normal distribution presents two unique challenges. First, there is no closed-form solution to determine the parameters of the distribution from the mean and variance. Second, even with known parameters, closed-form expressions for the mean and variance do not exist. Fortunately, techniques for estimating the mean and variance from parameters have been well-studied [29, 30].

Our goal is to find the parameters of the logit-normal distribution that produce estimated mean and standard deviation values as close as possible to the actual observed student data. We begin by making an initial guess of the distribution parameters and then leverage the TensorFlow Probability (TFP) library to analytically estimate the mean and standard deviation based on these initial parameters. To quantify how well our estimated parameters match the observed data, we define a loss function that captures the squared differences between the estimated and observed summary stats:

$$\mathcal{L} = \left(\mathbb{E}[\hat{G_j}] - \mathbb{E}[G_j]\right)^2 + \left(\text{Std}(\hat{G_j}) - \text{Std}(G_j)\right)^2$$

We then minimize this loss function using the L-BFGS-B optimization algorithm [31]. L-BFGS-B is particularly well-suited for this task because it efficiently handles smooth, high-dimensional functions. We run the optimization multiple times with different starting points to increase the likelihood of finding a global minimum, ensuring that our estimates are as accurate as possible. While we found L-BFGS-B to be stable and effective, other methods, such as rejection sampling, can also be employed to achieve similar results. After optimization, we recover the optimal difficulty $d_j$ and variance $\text{var}_j$, for each of the assessments. From there, we do the same steps as in the teacher version to recover the variance in student abilities $\sigma_a^2$ and the $\epsilon_j$ terms.

Either the teacher or student variant can be used in our model, depending on the data available, with no significant performance differences. With the estimated assessment parameters, we now turn to the next step: inferring student abilities.

### 3.3 Inferring Student Ability

Given the estimated assessment parameters-namely the standard deviation in student abilities $\sigma_a$, the difficulties $d_1, \ldots, d_n$, and the noise variances $\epsilon_1^2, \ldots, \epsilon_n^2$, along with the student's observed scores on assessments $g_{i1}, \ldots, g_{in}$, our next task is to infer the student's true ability $A_i$. The objective is to update our belief about $A_i$ by incorporating information from the student's performance on assessments.

The first step involves transforming the student's observed grades $g_{ij}$ into noisy ability samples $a_{ij}$. Recall that equation 10 models the grade $G_{ij}$ for student $i$ on assessment $j$. Since we have specific observed grade values, $g_{ij}$, we solve for the noisy abilities $a_{ij}$, by applying the inverse of the sigmoid function, known as the logit function.

$$a_{ij} = \text{logit}(g_{ij}) + d_j \tag{8}$$

Here, $a_{ij}$ is a specific noisy value, denoting the estimated ability of student $i$ on assessment $j$ based on their grade $g_{ij}$ and the assessment's difficulty $d_j$. Note that we use lowercase letters represent observed values, while uppercase letters are used for random variables. Now we have a list of noisy abilities $a_{i1}, \ldots, a_{in}$. To infer the student's true ability $A_i$ from the list of observed abilities, we use Bayesian inference, specifically leveraging the properties of the normal distribution to perform a Gaussian posterior update.

We begin with the prior belief that each $A_i$ is normally distributed as defined in equation 8. We define $\mu_{\text{prior}} = 0$ and $\sigma_{\text{prior}}^2 = \sigma_a^2$. Each observed ability $a_{ij}$ provides a likelihood function for $A_i$. Recall that $A_{ij} = A_i + M_j$. It then follows that $A_{ij} \sim N(0, \sigma_a^2 + \epsilon_j^2)$. If we condition $A_{ij}$ on $A_i$, we have $A_{ij}|A_i \sim N(A_i, \epsilon_j^2)$. This means that given $A_i$, the observed ability $A_{ij}$ is normally distributed around $A_i$ with variance $\epsilon_j^2$.

Since both the prior and likelihoods are normal distributions, the posterior distribution of $A_i$ given the observations $a_{i1}, \ldots, a_{in}$ is also normally distributed $A_i|a_{i1}, \ldots, a_{in} \sim N(\mu_{\text{posterior}}, \sigma_{\text{posterior}}^2)$. The parameters of the posterior distribution are calculated using standard formulas for Bayesian updating with normal distributions shown below.

$$\sigma_{\text{posterior}}^2 = \left(\frac{1}{\sigma_{\text{prior}}^2} + \sum_{j=1}^{k}\frac{1}{\epsilon_j^2}\right)^{-1},$$
$$\mu_{\text{posterior}} = \sigma_{\text{posterior}}^2\left(\frac{\mu_{\text{prior}}}{\sigma_{\text{prior}}^2} + \sum_{j=1}^{k}\frac{a_{ij}}{\epsilon_j^2}\right). \tag{9}$$

The outcome of the Gaussian posterior update is an updated normal distribution for $A_i$ that incorporates the evidence from the student's performance. The posterior mean, $\mu_{\text{posterior}}$ is our updated best estimate of the student's true ability and the posterior variance, $\sigma_{\text{posterior}}^2$ quantifies our uncertainty about this estimate. This process generalizes to any number of assessments allowing for flexible application across different courses and grading structures.

### 3.4 From Ability to Grade Distributions

Our next step is to use the inferred ability to estimate a distribution over the student's final grade, referred to as their Soft Grade. This distribution represents the likelihoods of all potential scores the student could receive, given the current belief of their ability and the characteristics of the assessments. We construct the Soft Grade using Monte Carlo simulation, sampling from the student's ability distribution to simulate assessment performance. This process generates an empirical distribution of final grades, effectively propagating the uncertainty in the student's ability.

We begin by drawing a large number of samples $a_i^{(s)}$, where $i$ identifies the student and $s$ is the sample index, from the inferred posterior distribution of the student's ability. Each sample represents a possible true ability level the student might have. For each sampled ability, $a_i^{(s)}$, we simulate the student's performance on each assessment $j$ by first sampling an observed ability $a_{ij}^{(s)}$ from the conditional distribution $A_{ij}|A_i \sim N(0, \epsilon_j^2)$. This step models the fluctuations in the student's performance on assessment $j$ due to factors unrelated to their true ability. Next, we compute the simulated grade, $g_{ij}^{(s)} = \text{Sigmoid}(a_{ij}^{(s)} - d_j)$. For each sample, we then have $n$ simulated grades, one for each of the $n$ assessments. We then compute a final grade as an average of all $n$ grades, denoted as $f_i^{(s)}$ and computed:

$$f_i^{(s)} = \frac{1}{n}\Sigma_{j=1}^{n} g_{ij}^{(s)} \tag{10}$$

Each final grade sample is rounded to two decimal places to ensure consistency in reporting and to reflect real-world grading practices. We repeat this for a large number of samples (e.g., 10000 iterations) to create a set of final grades $f_i^1, f_i^2, \ldots f_i^S$. This collection of discrete simulated final grades forms the student's Soft Grade distribution.

## 4 Experimental Setup

To evaluate our model, we conducted experiments to predict students' future grades. We outline the datasets, evaluation metrics, and baseline models used for comparison.

### 4.1 Datasets

We conducted experiments using both synthetic and real-world datasets. This dual approach allowed us to assess the model's performance under controlled conditions with known parameters, as well as its applicability in real university courses. For the synthetic dataset, we generated data from a simulated course environment to evaluate our model's capability to recover the true parameters. The simulated course consisted of 9 assignments and 200 students. To enhance the robustness of our results, we created ten different offerings of this course, each with varying parameters. The outcomes are averaged across all offerings.

For the real-world evaluation, we used two datasets: the first is non-public data from two courses at a large R1 research university (denoted as datasets C1 and C2), and the second is the publicly available Open University Learning Analytics Dataset (denoted as OULAD) [28], consisting of seven courses spanning two years. Grades were clipped to the range 0.005 to 0.995 to ensure numerical stability with the sigmoid function, and only students with scores for all assessments were included. C1 is an introductory computer science course with 9 assessments and 378 students. C2 is a probability course with 8 assessments, drawn from three offerings with 392, 307, and 239 students respectively. The OULAD dataset includes 7 courses (OULAD-1 through OULAD-7), each with 2-4 offerings each. Each offering had 5 to 13 assessments and 200 to 2,000 students. For all offerings, we inferred student ability from the first half of assessments (rounding down when the total number of assessments was odd) and predicted performance on the rest.

### 4.2 Evaluation Metrics

We used likelihood as a metric to assess how closely the predicted grades matched the true final grades, measured at different levels of precision (exact match, within ±1, ±3, etc.). Next, we used calibration plots to compare predicted probabilities of achieving certain grades with observed frequencies. A well-calibrated model's predictions should align with the diagonal $y = x$ line, indicating that predicted probabilities match actual outcomes. We quantified calibration using Expected Calibration Error (ECE), with lower ECE indicating better calibration. Finally, we conducted a hard prediction evaluation, generating a single grade estimate instead of a distribution. This was done by using the mean of the inferred true ability $A_i$ to produce a final grade prediction for each student. Hard predictions provide a more straightforward comparison to traditional IRT models, which also produce single grade estimates. We evaluated these predictions using Root Mean Square Error (RMSE), which measures how closely the predicted final grades match the given final grades.

### 4.3 Baseline Comparison

We implemented a baseline model for comparison to our soft predictions. In this baseline, each student's grade is represented as a normal distribution, with the mean set to their average score across all assessments and a fixed, small standard deviation to reflect minimal uncertainty. This Fixed Mean Normal (FMN) baseline reflects current grading practices, ignoring individual abilities and assessment difficulties, and having little to no uncertainty.

To compare to the hard predictions, we implemented a Continuous Response Model (CRM) [6] where we fit the parameters with the EM approach described in this paper [9]. We learned parameters and student abilities from the first half of the assessments. For the remaining assessments, we used the average parameter values learned from the first half to predict future scores. This approach is referred to as the CRM baseline.

## 5 Results

### 5.1 Synthetic Data

We first evaluate our model using synthetic data, where the true parameters for each assessment are known. This allows us to compare our model, which learns these parameters, to an oracle model that uses the true parameter values for all predictions. By doing so, we can validate whether our model accurately learns the correct parameters.

The Soft Grades model performs nearly as well as the Oracle, as shown in Table 1. The likelihood of exactly predicting the final grade is 0.118 for the Soft Grades model, compared to 0.119 for the Oracle. With a tolerance of ±1, the likelihoods are 0.323 and 0.324, respectively, and for a tolerance of ±3, they are 0.605 and 0.612. These results demonstrate that the Soft Grades model accurately learns the true parameters of the synthetic data, achieving predictions almost as precise as the Oracle. In contrast, the FMN Baseline model, which does not learn parameters, performs significantly worse. Its likelihood of exactly matching the final grade is 0.066,

| | Likelihood | | | Expected Calibration Error |
|---|---|---|---|---|
| | **Exact** | **±1** | **±3** | (ECE) |
| Oracle | 0.119 ± 0.007 | 0.324 ± 0.013 | 0.612 ± 0.016 | 0.018 |
| Soft Grades | 0.118 ± 0.007 | 0.323 ± 0.014 | 0.605 ± 0.016 | 0.020 |
| FMN Baseline | 0.066 ± 0.007 | 0.192 ± 0.019 | 0.407 ± 0.027 | 0.131 |

**Table 1: Likelihood of predicting the final grade within different tolerance levels (Exact, ±1, ±3) for synthetic data using Oracle, Soft Grades, and FMN Baseline models. The Oracle represents the theoretical upper bound of performance, and lower ECE values indicate better calibration.**

and for ±1, it achieves 0.192—substantially lower than both the Oracle and the Soft Grades models.

The calibration plot for synthetic data (Figure 4c) shows that the Soft Grades model is well-calibrated, with points closely aligning to the $y = x$ line. This indicates that the model's predicted probabilities closely match the actual outcomes. The Expected Calibration Error (ECE), shown on the right side of table 1 is 0.020, significantly lower than the FMN baseline's ECE of 0.131, and very close to the oracle's ECE of 0.018, indicating that the Soft Grades model provides reliable probability estimates in line with the true outcomes.

## 5.2 Real Data

The calibration plots for C1 and C2 (Figures 4a and 4b) show that the Soft Grades model is well-calibrated in predicting final grades on these two real datasets, demonstrating that the Soft Grades model is robust when applied to data from these two university courses.

The ECE values for real data (Table 2) provide further evidence of the model's calibration accuracy. For C1, the Soft Grades model achieves an ECE of 0.021, compared to the FMN baseline's 0.187. Similarly, for C2, the Soft Grades model has an ECE of 0.064, significantly lower than the baseline's 0.119. In the OULAD courses, the Soft Grades model consistently outperforms the baseline, with ECE values ranging from 0.039 to 0.174, while the baseline's ECE values are much higher, ranging from 0.242 to 0.530. These results indicate that the Soft Grades model is much better calibrated across various courses compared to the FMN baseline.

In addition to calibration, we assess the likelihood of the true final grades falling within specific ranges around the predicted grades, using the same method as with the synthetic data. In the OULAD dataset, the Soft Grades model consistently outperforms the baseline. In OULAD-1, the likelihood of exactly predicting the final grade is 0.059 for Soft Grades, compared to 0.100 for the baseline. At a tolerance of ±3, the Soft Grades model achieves 0.396, while the baseline reaches 0.574. In OULAD-7, the Soft Grades model predicts with an exact likelihood of 0.058, outperforming the baseline's 0.034, and at ±3, it achieves 0.393 compared to 0.252 for the baseline. Similar patterns emerge in the university courses. In C1, the Soft Grades model achieves a likelihood of 0.201 for exact predictions, 0.482 at ±1, and 0.762 at ±3, outperforming the baseline at every level. In C2, the exact likelihood is 0.120 for Soft Grades, with 0.341 at ±1 and 0.652 at ±3, remaining competitive with the baseline. These results highlight the Soft Grades model's ability to deliver more accurate predictions, especially at wider tolerances.

Finally, we compare the hard predictions of our model to the CRM baseline to evaluate its relative accuracy. Table 3 presents the Root Mean Square Error (RMSE) of hard predictions from the Soft Grades Model compared to the CRM baseline across different courses. The Soft Grades model consistently achieves lower RMSE values, indicating more accurate predictions. For instance, in C1, the Soft Grades model has an RMSE of 0.042, compared to the CRM baseline's 0.060. Similarly, in C2, the Soft Grades model achieves a much lower RMSE of 0.058, while the CRM baseline records a significantly higher RMSE of 0.264. This trend continues across the OULAD courses, with the Soft Grades model consistently outperforming the CRM baseline. For example, in OULAD-2, the Soft Grades model's RMSE is 0.067, compared to 0.221 for the CRM baseline, and in OULAD-6, the RMSE is 0.046 for Soft Grades, while the CRM baseline reaches 0.401. These results show the Soft Grades model is more accurate than CRM on these datasets.

# 6 Practical Applications of Soft Grades

The Soft Grades model offers powerful and practical tools for both teachers and students, providing a richer and more informative representation of student performance.

## 6.1 Soft Grades For Teachers

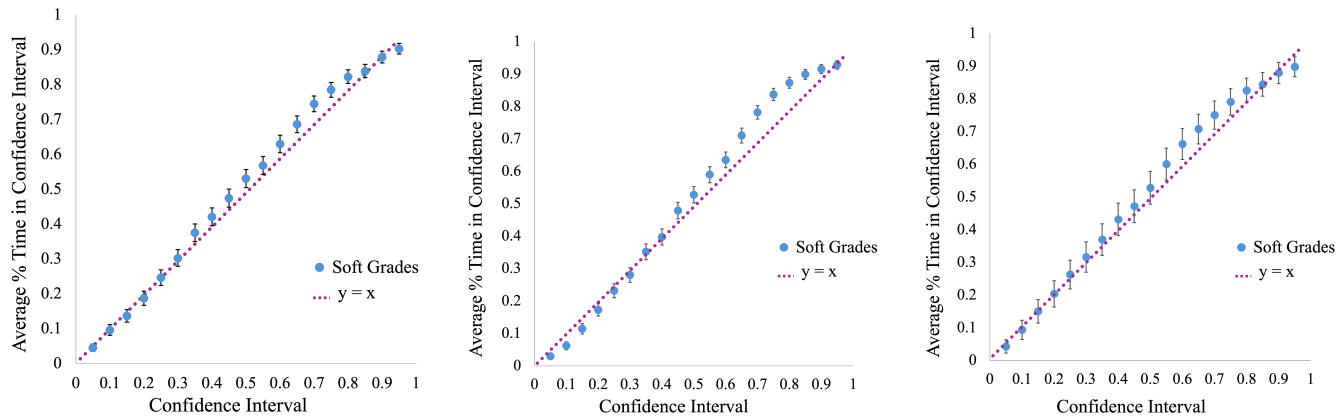
Soft Grades provides a principled way to impute missing scores for students who, due to extenuating circumstances such as family emergencies or health issues, have incomplete coursework. Instead of dropping a missing grade, teachers can infer a student's ability using the assessments they have completed. From this inferred ability, they can generate an estimated grade for that assessment using Monte Carlo sampling from the student's ability distribution. The teacher can then have more information when deciding how to handle cases like this.

Furthermore, Soft Grades can improve grading practices by providing teachers with a visual representation of the uncertainty in the grades they assign. A wider spread in a student's Soft Grade distribution indicates higher uncertainty, offering teachers insights into how confident they should be about a student's performance. This can be particularly valuable when assigning grades to students who are on the borderline of grade boundaries.

Many teachers curve grades to align student outcomes with expected distributions. Soft Grades support this process by offering histograms of standard deviations and visualizations of scores with their uncertainties. These tools enable teachers to make more informed decisions about how to apply curves fairly, ensuring the final grade distribution aligns with the variability in student performance.

## 6.2 Soft Grades For Students

We make a few modifications to the algorithm to better serve students' needs. First, students input their scores and the statistics (mean and standard deviation) for the $k$ assessments they have completed so far. The model uses this data to infer the student's




**(a) Calibration plot for one offering of C1.** **(b) Calibration plot for one offering of C2.** **(c) Calibration plot for synthetic data.**

**Figure 4: Calibration plots comparing the Soft Grades model across two real course offerings (C1 and C2) and one synthetic dataset. The model is well-calibrated in all cases, suggesting that the Soft Grades model accurately captures uncertainty and provides reliable predictions. Error bars are standard error of the mean.**

| | C1 | C2 | OULAD-1 | OULAD-2 | OULAD-3 | OULAD-4 | OULAD-5 | OULAD-6 | OULAD-7 |
|---|---|---|---|---|---|---|---|---|---|
| **Soft Grades (ECE)** | **0.021** | **0.064** | **0.049** | **0.084** | **0.174** | **0.110** | **0.039** | **0.082** | **0.080** |
| FMN Baseline (ECE) | 0.187 | 0.119 | 0.275 | 0.375 | 0.308 | 0.435 | 0.338 | 0.242 | 0.530 |

**Table 2: Expected Calibration Error (ECE) of Soft Grades Model compared to FMN baseline averaged across all offerings of each course. The Soft Grades model demonstrates significantly lower ECE than the FMN baseline, indicating better calibration in its predictions. The p-values for all comparisons are < 0.0001, indicating that the observed differences in ECE are statistically significant.**

| | C1 | C2 | OULAD-1 | OULAD-2 | OULAD-3 | OULAD-4 | OULAD-5 | OULAD-6 | OULAD-7 |
|---|---|---|---|---|---|---|---|---|---|
| **Soft Grades (RMSE)** | **0.042** | **0.058** | **0.069** | **0.067** | **0.062** | **0.089** | **0.071** | **0.046** | **0.076** |
| CRM Baseline (RMSE) | 0.060 | 0.264 | 0.079 | 0.221 | 0.071 | 0.145 | 0.080 | 0.401 | 0.349 |

**Table 3: Root Mean Square Error (RMSE) comparison between Soft Grades and CRM Baseline across all courses. Lower RMSE is better. All comparisons are statistically significant, with p-values < 0.05.**

current ability. Next, the student provides their best estimates for the mean and standard deviation of the remaining $n-k$ assessments. With these estimates, the model runs Monte Carlo simulations to predict potential grades for the remaining assessments.

In each simulation, the final grade combines the student's actual scores from the first $k$ assessments with the predicted scores for only the remaining $n-k$ assessments. This differs from the teacher's version of the Soft Grades model, where we simulate scores for all $n$ assessments using the inferred ability. For students, it is most helpful to focus on the range of possible future outcomes rather than the uncertainty in their past performance, since this information allows them to make actionable decisions about how to approach the remainder of the course.

The model then produces a final soft grade distribution, giving the student a clear understanding of the likelihood of potential outcomes in the course. At this point, the student has two options. First, they can experiment with different assumptions about the difficulty of future assessments, adjusting the statistics to see how these changes affect the possible grades. Second, they can select a specific grade for a future assignment and ask the model to assume they received that grade. The model will then update its inference of the student's ability using the original $k$ true scores, plus the newly assumed score for the selected assignment. This feature allows students to explore how different assessment parameters, such as difficulty, can influence their final outcomes, and see how achieving specific future scores would affect their overall grade. This level of interactive exploration is not available with current tools.

This ability to visualize a distribution over possible outcomes gives students valuable information they can use to make strategic academic decisions, such as how to best allocate their study time across different courses, or even whether to opt for a letter grade or Pass/Fail based on the likelihood of outcomes. By offering a detailed look into potential outcomes, the Soft Grades model encourages proactive academic planning and helps students make more informed decisions about their study habits and goals.

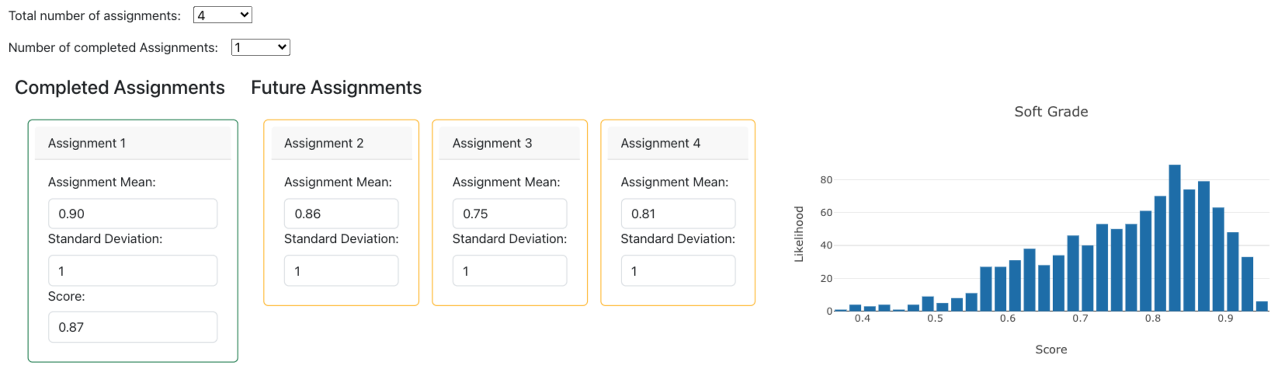



**Figure 5: Soft Grade web application user interface for students**

## 6.3 Web Application

We have developed a web application that allows students and teachers to use the Soft Grades model. For students, the app allows them to input their completed assignment scores and statistics. They can then experiment with different scenarios for upcoming assignments to predict their soft final grade or the Soft Grades for specific future assessments. The user interface for the web application student page in shown in figure 5. The teacher version of the web app links to Python scripts for imputing grades, viewing Soft Grades, and generating class-wide statistics like the distribution of standard deviations. These scripts run locally to ensure student data privacy and are available with instructions at: https://juliettewoodrow.github.io/softgrades/.

## 7 Discussion

An important extension of Soft Grades is to refine the modeling of assessment variance by separating the variance due to student abilities ($\sigma_a^2$) from the variance caused by noise ($\epsilon_j^2$, factors unrelated to ability). Perhaps with rich historical data, or when prior knowledge about assessment noise is available, we could develop models that more precisely capture how much of the uncertainty stems from the student and how much arises from external factors. This differentiation could lead to better-informed decisions about interventions and adjustments in grading practices.

Another extension is the concept of a "Soft GPA", which would incorporate the ideas behind soft grades into a comprehensive measure of student performance over the course of an entire academic program. Unlike traditional GPAs, which reduce student performance to point estimates, a Soft GPA would reflect both the student's ability, the varying difficulties of different courses, and the uncertainty-especially early on in a student's academic journey. This could lead to a richer understanding of students' academic trajectories, enabling employers, advisors, instructors, and students themselves to make more informed decisions.

Finally, although our current model assumes that grades follow a logit-normal distribution, it is worth exploring alternative distributions to model assessments. Other distributions, such as the beta distribution, may better fit certain types of grading patterns.

## 8 Limitations

A key assumption in our model is that a student's true ability is assumed to remain constant throughout an entire course. This simplification facilitates the modeling process, but it does not account for the dynamic nature of learning. Abilities can and should certainly evolve throughout a course as students learn course content and meta-course skills (e.g. test taking strategies). Future work should explore methods for dynamically updating ability estimates, potentially incorporating temporal models or Bayesian approaches that allow for evolving student abilities.

Another limitation is that the Soft Grades model has not yet been deployed in live classroom environments. While the model performs well in simulations and post-hoc analysis of real-world data, there is much to learn from observing how teachers and students actually engage with the tool in practice. Future work could involve usability studies and feedback from teachers and students to better understand how the model integrates into real classroom settings.

## 9 Conclusion

Soft Grades offer a new lens for understanding student performance by incorporating uncertainty into final course grades. By moving beyond single-point estimates, this model provides a richer framework that better reflects the realities of classrooms. With demonstrated state-of-the-art performance in grade prediction and imputation, we envision Soft Grades as a transformative tool for educators and students alike—empowering more informed decisions and fostering deeper insights into academic achievement.

## Acknowledgments

We would like to sincerely thank the Carina Foundation for making this research possible.

## References


[1] C. Piech, A. Malik, L. Mapstone, R. Chang, and C. Lin. 2020. The Stanford Acuity Test: A Precise Vision Test Using Bayesian Techniques and a Discovery in Human Visual Response. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence* (New York, USA).

[2] Rudolph E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82, 1 (1960), 35–45.

[3] Allan H. Murphy. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* 8, 2 (1993), 281–293.

[4] Eiji Muraki. 1992. A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series* 1992, 1 (1992), i–30.

[5] Fumiko Samejima. 2016. Graded response models. In *Handbook of Item Response Theory, Volume One*. Chapman and Hall/CRC, 123–136.

[6] Fumiko Samejima. 1973. Homogeneous case of the continuous response model. *Psychometrika* 38, 2 (1973), 203–219.

[7] Yu Chen, Telmo de Menezes e Silva Filho, Ricardo B. C. Prudêncio, Tom Diethe, and Peter A. Flach. 2019. $\beta$3-IRT: A New Item Response Model and its Applications. *ArXiv* abs/1903.04016 (2019). https://api.semanticscholar.org/CorpusID:91185736

[8] Pere J. Ferrando. 2019. A Comprehensive IRT Approach for Modeling Binary, Graded, and Continuous Responses With Error in Persons and Items. *Applied Psychological Measurement* 43, 5 (2019), 339–359. https://doi.org/10.1177/0146621618817779

[9] Cengiz Zopluoglu. 2013. A comparison of two estimation algorithms for Samejima's continuous IRT model. *Behavior Research Methods* 45, 1 (2013), 54–64.

[10] George Karabatsos. 2016. Bayesian Nonparametric Response Models. In *Handbook of Item Response Theory* (1st ed.), Wim J. van der Linden (Ed.). Chapman and Hall/CRC, 13. https://doi.org/10.1201/9781315119144

[11] Weijie Jiang and Zachary A. Pardos. 2021. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 608–617. https://doi.org/10.1145/3461702.3462623

[12] Jake Barrett, Alasdair Day, and Kobi Gal. 2024. Improving Model Fairness with Time-Augmented Bayesian Knowledge Tracing. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 46–54. https://doi.org/10.1145/3636555.3636849

[13] Paulo Blikstein. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*. 110–116.

[14] Yossi Ben David, Avi Segal, and Ya'akov Gal. 2016. Sequencing educational content in classrooms using Bayesian knowledge tracing. In *Proceedings of the sixth international conference on Learning Analytics & Knowledge*. 354–363.

[15] Shuchi Grover, Satabdi Basu, Marie Bienkowski, Michael Eagle, Nicholas Diana, and John Stamper. 2017. A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education (TOCE)* 17, 3 (2017), 1–25.

[16] Yun Huang, Yanbo Xu, and Peter Brusilovsky. 2014. Doing more with less: Student modeling and performance prediction with reduced content models. In *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22*. Springer, 338–349.

[17] Balqis Albreiki, Nazar Zaki, and Hany Alashwal. 2021. A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences* 11, 9 (2021), 552.

[18] Carrie Demmans Epp and Susan Bull. 2015. Uncertainty Representation in Visualizations of Learning Analytics for Learners: Current Approaches and Opportunities. *IEEE Transactions on Learning Technologies* 8, 3 (2015), 242–260. https://doi.org/10.1109/TLT.2015.2411604

[19] Benjamin Motz. 2023. Concentration toward the mode: Estimating changes in the shape of a distribution of student data. https://doi.org/10.31234/osf.io/6p9td

[20] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems (NeurIPS)*. 505–513.

[21] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. In *Proceedings of the 6th International Conference on User Modeling*. Springer-Verlag, 113–122. https://doi.org/10.1007/978-3-0404-79180-0_13

[22] Ben Gillen, Erik Snowberg, and Leeat Yariv. 2019. Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy* 127, 4 (2019), 1826–1863. https://doi.org/10.1086/701681

[23] Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing.

[24] Georg Rasch. 1960. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*.

[25] Noah Arthurs, Ben Stenhaug, Sergey Karayev, and Chris Piech. 2019. Grades Are Not Normal: Improving Exam Score Models Using the Logit-Normal Distribution. *International Educational Data Mining Society* (2019).

[26] Sooyeon Kim, Tim Moses, and Hanwook Yoo. 2015. Effectiveness of Item Response Theory (IRT) Proficiency Estimation Methods Under Adaptive Multistage Testing: Effectiveness of IRT Proficiency Estimation Methods. *ETS Research Report Series* 2015 (05 2015). https://doi.org/10.1002/ets2.12057

[27] Y. Lee. 2019. Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill. *Information Discovery and Delivery* 47, 2 (2019), 67–75. https://doi.org/10.1108/IDD-08-2018-0030

[28] J. Kuzilek, M. Hlosta, and Z. Zdrahal. 2017. Open University Learning Analytics dataset. *Scientific Data* 4 (2017), 170171. https://doi.org/10.1038/sdata.2017.171

[29] James B. Holmes and Matthew R. Schofield. 2020. Moments of the logit-normal distribution. *Communications in Statistics - Theory and Methods* 51, 3 (2020), 610–623. https://doi.org/10.1080/03610926.2020.1752723

[30] Jhon Atchison and Sheng M Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika* 67, 2 (1980), 261–272.

[31] Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45, 1-3 (1989), 503–528.