Grades are not Normal: Improving Exam Score Models Using the Logit-Normal Distribution

Noah Arthurs
Stanford University
narthurs@cs.stanford.edu

Ben Stenhaug
Stanford University
stenhaug@stanford.edu

Sergey Karayev
Gradescope
sergeyk@gradescope.com

Chris Piech
Stanford University
piech@cs.stanford.edu

ABSTRACT

Understanding exam score distributions has implications for item response theory (IRT), grade curving, and downstream modeling tasks such as peer grading. Historically, grades have been assumed to be normally distributed, and to this day the normal is the ubiquitous choice for modeling exam scores. While this is a good assumption for tests comprised of equally-weighted dichotomous items, it breaks down on the highly polytomous domain of undergraduate-level exams. The logit-normal is a natural alternative because it is has a bounded range, can represent asymmetric distributions, and lines up with IRT models that perform logistic transformations on normally distributed abilities. To tackle this question, we analyze an anonymized dataset from Gradescope consisting of over 4000 highly polytomous undergraduate exams. We show that the logit-normal better models this data without having more parameters than the normal. In addition, we propose a new continuous polytomous IRT model that reduces the number of item-parameters by using a logit-normal assumption at the item level.

1. INTRODUCTION

Historically, student performance on exams has been assumed to be normally distributed. Grade curving originates from the idea that students exist on a "bell curve," in which most are clustered around the mean and a small number over- or under-achieve. The field of education has many criticisms for the bell-curve mindset. A common argument is that we should not take the *observation* that student performance tends to look normal and turn it into a normative practice [4, 23]. The idea that some students will inevitably fail and only a small number can enjoy the highest level of success runs counter to the goals of the educator, who should want as many students as possible to succeed. This tension plays out in the ideological battle between those who criticize grade inflation [9] and those who suggest that students may be earning the higher grades they are receiving [11].

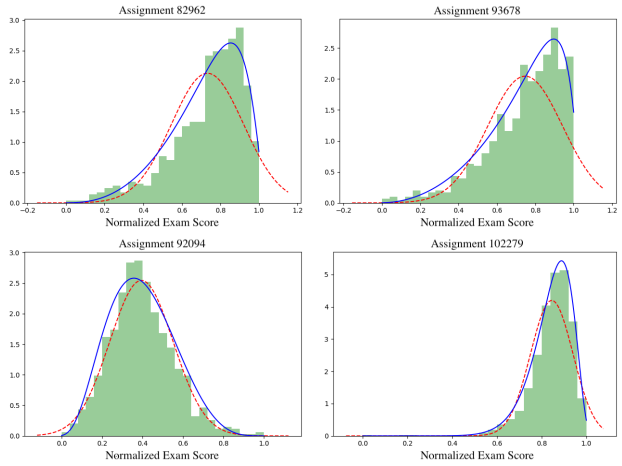

Figure 1: Score histograms of four assignments, along with the PDFs of the best-fit normals (dashed red) and best-fit logit-normals (solid blue).

The normal assumption is commonplace in modern research into educational data. Grade distributions are usually presented to students in terms of their mean and variance, and they are often visualized as normal distributions [17]. As education becomes more digitized, statistical models of grading become more widespread. For example, peer grading models allow MOOC's to assign accurate grades to students based on noisy estimates from their classmates. State of the art peer grading models use normal priors over grades [19, 18], which will result in normal-looking distributions. Both of these examples can benefit from challenging the normal assumption. In the first case, finding new ways to parameterize grade distributions can help us better interpret and visualize student behavior. In the second, a more accurate prior over grades would help peer grading models assign more accurate grades to students.

In this paper, we analyze over 4000 assignments graded on the Gradescope platform [22]. These assignments are primarily exams from undergraduate STEM courses, and as a result the data is highly polytomous (many scoring categories per question). However, principal component analysis (PCA) reveals that these exams have good low-rank approximations, meaning that the data is very structured.

First, we examine the ability of different families of distributions to model the scores in our dataset. Specifically we compare the normal to three bounded two-parameter distributions. We find that the logit-normal [1, 5] is consistently the best choice, followed by the beta, which is known to approximate the logit-normal [8].

In the second part of this paper, we build a simple continuous polytomous IRT model using a logit-normal assumption at the item level. Our model outperforms both the Generalized Partial Credit Model [16] (a standard discrete IRT model) and the Continuous Response Model [20] (a standard continuous IRT model) on the Gradescope data, despite having fewer parameters than either. This indicates that we can simplify and improve polytomous IRT models using structural assumptions about assignment data.

1.1 Related Work

When analyzing student behavior, it can be difficult to distinguish between cases where data is actually normal and cases where an assumption of normality is influencing the distribution. For example, SAT scores are known to be normally distributed, but this is because raw SAT scores are translated into final scores using a system that enforces a normal distribution [3]. More subtly, probabilistic models for determining scores based on peer grades often use normal priors over their output [18, 19]. As a result, they will push grade distributions to be normal. The question then remains about whether these distributions *should* look normal in reality or another prior needs to be found.

Of course, grade curving is the most direct way in which student performance is influenced to be normal. Although it remains a common practice, research has shown that most students prefer not to be graded on a curve [6], that both students and professors find indiscriminate grade curving unethical [15], and that grade curving can amplify the randomness of test-taking as a measure of student aptitude [12]. It has also been argued that educators should be striving to avoid normally distributed student outcomes, rather than enforce them [4, 23]. If this is the case, then we need to actively seek out new distributions for describing and understanding test scores.

Polytomous IRT models generally fall into two categories. Discrete models like Generalized Partial Credit [16] and Graded Response Models [21] model each point on each question separately, scaling with the number of scoring categories per question. These models make very few assumptions about the relationship between different scores and thus do not take advantage of any underlying structure in the data. Continuous models like the Continuous Response Model [20] are used less frequently, but they scale only with the number of questions in the assignment. They do this by making assumptions about the structure of the item characteristic curves (ICC's). This means that if a dataset's ICC's follow a consistent pattern, then a continuous model can thrive.

The Gradescope data we are working with is much more polytomous than most IRT datasets. This is because it comes from a wide variety of college-level courses rather than standardized tests. Despite this heterogeneity, past work on Gradescope data has found patterns in question ordering and the interpretation of the first several principal components [13]. This indicates that there may be underlying structure that a continuous IRT model could take advantage of.

2. THE DATASET

Our initial dataset consists of 6,607 assignments submitted to Gradescope, an online tool for uploading and grading student work [22]. Typically students will do their assignments by hand using a template provided by the instructor. After the assignments have been scanned and uploaded to Gradescope, instructors can grade them using a digital rubric that is always visible and modifiable. To ensure that the majority of the data consists of college-level exams, all included assignments:

- are instructor-uploaded[1]
- have a fixed template
- have at least 3 questions
- have titles that do not include terms that describe other kinds of student work (e.g. "HW" or "Quiz")
- have titles that do not include terms that are indicative of non-college-level work

The assignments were graded between Spring 2013 and Spring 2018, and come from 2748 courses at 139 different higher-education institutions[2]. The top three subject areas in the dataset are Electrical Engineering & Computer Science (50% of assignments), Math & Statistics (22%), and Chemistry (11%). The median number of courses per school is 4, and the median number of assignments per course is 2.

When fitting curves to exam scores, we want there to be enough values that the distribution is interesting/nontrivial and enough data points that the observed histogram is somewhere near the underlying distribution. For this reason, we filter out all assignments that have fewer than 10 unique scores or fewer than 75 students. We are then left with 4115 assignments. Figure 2(b) shows the joint distribution between student count and number of unique scores, as well as their marginals. Observe that the vast majority of assignments have 75-200 students and 20-80 unique scores.

Throughout this study we store each assignment as a matrix A where A_{ij} represents student j's score on question i. We use the term "exam score" to refer to the sum of a student's question scores, so to analyze exam scores, we sum the rows of the assignment matrix A.

2.1 Visualizing the Data

One shortcoming of the normal is that it can only represent symmetric distributions.[3] We measure the symmetry of an exam score distribution using skew (skewness), which can be

[1]On Gradescope, students tend to upload their own homework, while exams tend to be scanned and uploaded by the instructor.

[2]However, UC Berkeley, UC San Diego, Stanford University, University of Michigan, and University of Washington account for half of all assignments in the dataset.

[3]This may not be a problem on all forms of test data. For example, if questions were equally weighted and (relatively) independent, the Central Limit Theorem would predict a symmetric distribution.

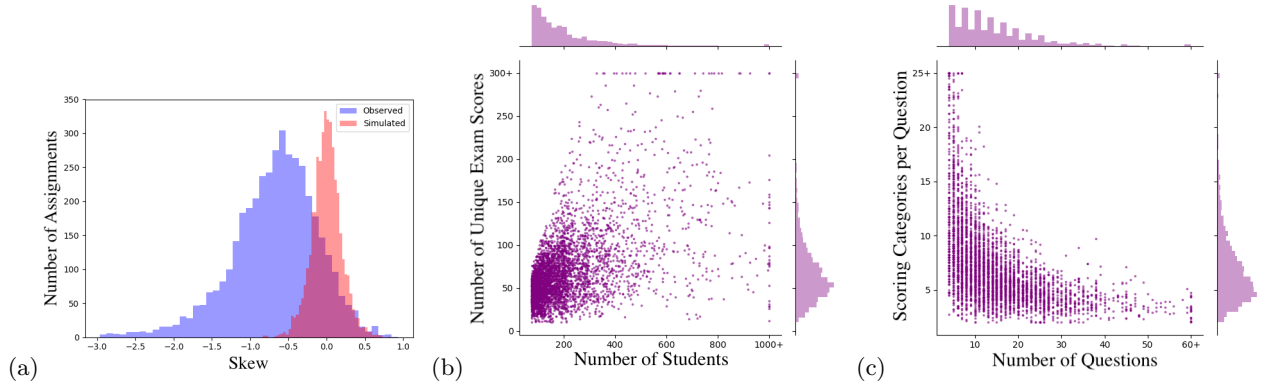

Figure 2: (a) Normal doesn't fit. Blue histogram: assignment skew distribution. Red histogram: skews of assignments had they been drawn from normal distributions (as described in section 2.1). (b) Number of students vs. number of unique exam scores for each assignment in the filtered dataset. (c) Number of questions vs. average number of scoring categories per question for each assignment in the filtered dataset.

a good indicator of how normal a distribution is [2]. A skew value of (or near) 0 indicates a symmetric distribution, while negative and positive skew values indicate large tails on the left and right respectively. The blue histogram in Figure 2(a) shows the skews of the exam score distributions in our dataset. Note that they tend to be negative, meaning that exams tend to have larger tails of below-mean students than of above-mean students. To generate the red histogram in Figure 2(a), we performed the same experiment on a simulated dataset created by redrawing the scores of each assignment from its best-fit normal[4]. The large difference in both mean and variance of these histograms shows that our observed skews would not be very likely were the data normally distributed. In order to quantify this intuition, we performed a D'Agostino's K-squared test of normality to determine how likely it would be for each assignment's skew to arise from a normal distribution [2]. We found that 73% of assignments had a p-value of 0.05 or lower, indicating that (just on the basis of skew) the vast majority of assignments are very unlikely to have come from a normal distribution.

When fitting IRT models to the assignments, we are interested in how polytomous each assignment is. Figure 2(c) shows the joint distribution between the number of questions an assignment has and the average number of scoring categories per question. The negative correlation between these two stats is unsurprising[5], but it means that we can test our models on highly polytomous items or large question-counts but not both at the same time.

2.2 Dimensionality

PCA [10] can give us insight into the dimensionality of our data. A previous use of PCA on Gradescope data [13] found that the first principal component distinguishes between high and low scoring students, while later principal components correspond to skill at particular types of questions (e.g. multiple choice, free response).

[4]If an assignment had n students, sample mean $\bar{x}$ and sample variance S, our simulated version of that assignment would consist of n draws from $N(\bar{x}, S)$.

[5]There are only so many points that a student could be expected to earn over the course of a single exam.

We use PCA to describe the dimensionality of a given exam, where dimensionality is defined as the number of principal components required to account for 80% of the variance between students. Put simply, we are interested in what rank is required to form a "pretty good" approximation for the exam matrix. Intuitively, if exams have low dimensionality, then they have a large amount of structure we can exploit when modeling them.

We find that number of students and number of questions (the two dimensions of our exam matrix) do not influence dimensionality in the same way. Number of students is weakly correlated with dimensionality (0.20 Pearson correlation), and on average it requires over 250 extra students to add a dimension. Number of questions, on the other hand is more strongly correlated (0.85 Pearson correlation), and one dimension is added every 3.3 questions. This is consistent with the finding in [13] that principal components correspond to specific student aptitudes. Overall, this analysis indicates that choice of model should be based on the features of the exam itself, not on how many students are taking it.

We also examined the first principal component in isolation, and found that it generally indicated a student's score. Across our dataset, the average magnitude of the Pearson correlation between exam score and the first principal component was 0.965. In addition, the first principal component on average accounts for 43% of the variability in an assignment. The strength of the first principal component indicates that IRT models might only need one-dimensional student ability parameters to be successful on this dataset.

3. FITTING EXAM SCORES

Our first modeling task is to compare the ability of different two-parameter distributions to fit the exam scores in our dataset. Since tests have minimum and maximum scores, we choose bounded distributions. In addition, due to our findings in 2.1, we choose distributions that are not symmetric.

3.1 The Distributions

The truncated normal distribution is the result of bounding the normal above and below. It is characterized by the

	Normal	Trunc	Beta	Logit
Beats Normal	-	100%	92%	87%
Beats Trunc	0%	-	67%	75%
Beats Beta	8%	33%	-	68%
Beats Logit	13%	25%	32%	-
Average LL	0.272	0.333	0.336	0.353

Table 1: Win Rates and Likelihoods: How often does each distribution outperform the others? The logit-normal, beta and truncated normal models are all better replacements for the normal distribution. logit-normal has the highest likelihood.

mean and variance of its underlying normal, and when the bounds are known, there is a closed form maximum likelihood (MLE) estimate of these parameters [7]. The truncated normal assigns a higher probability density to every value in its domain than its underlying normal does, and as a result it will strictly outperform the normal distribution in likelihood. Although it is not symmetric around its mean, if it includes the mean of the underlying normal, then its probability density function (PDF) will be mirrored across that point. The truncated normal will be a good fit if test scores are drawn not from a normal distribution but from a slice of a normal distribution.

The beta distribution is the conjugate prior of the Bernoulli, characterized by two parameters referred to as α and β. It has no closed form MLE estimates, but there is a closed form method of moments solution that can be used as a starting point for optimization. When $\alpha = \beta$, the beta has no skew, but it can achieve a wide range of skew values by varying the difference between the two parameters. The beta does not have an intuitive interpretation in this context.

The logit-normal distribution [1, 5] is the result of applying the sigmoid (logistic) function[6] to data sampled from a normal distribution. Like the truncated normal, its parameters are the mean and variance of its underlying normal. The logit-normal and beta are known to approximate each other [8], but the logit-normal comes with the advantage of having closed form MLE estimates of its parameters[7]. It also has a nice interpretation that comes from item response theory. Logistic IRT models like 1PL/2PL/3PL take normally-distributed student abilities and use a linear transformation plus a sigmoid to transform them into probabilities. If the logit-normal is a good fit for exam scores, we can see it as performing the same kind of transformation to go from an underlying unbounded variable to an observed bounded one.

In addition to these, we tried a Two Gaussian Mixture Model in case distributions were bimodal. However, it performed worse than all three of these distributions despite having more parameters, so we did not include it in our results.

[6]The sigmoid, given by $\sigma(x) = \frac{1}{1+e^{-x}}$ compresses the reals into the range $(0, 1)$. Its inverse, the logit function given by $\sigma^{-1}(p) = \log \frac{p}{1-p}$, does the opposite.

[7]Unsurprisingly, the MLE estimates for the mean and variance of the underlying normal are the mean and variance of taking the logit (σ^{-1}) of the exam scores.

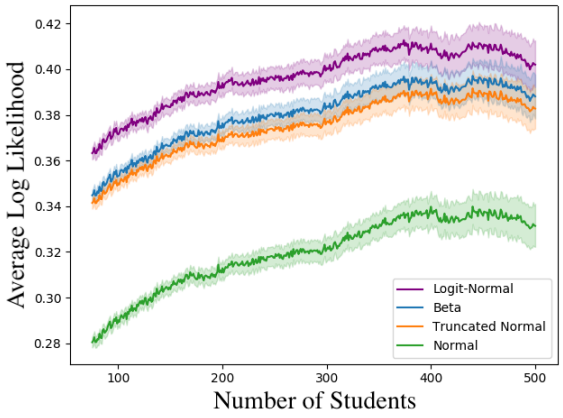

Figure 3: Difference in performance between the candidate distributions across sample sizes. Bands show standard error. Assignments were downsampled to simulate lower student counts.

3.2 Evaluating the Distributions

We fit each of the three distributions above plus the normal to each of the 4115 sets of exam scores in our filtered dataset. In order to more easily fit our bounded distributions to the data, we compress all scores into the range $[0.05, 0.95]$[8]. Specifically, if we have observed exam scores $x_1, ..., x_N$, we map each x_i to

$$x_i' = 0.9 * \frac{x_i - x^{min}}{x^{max} - x^{min}} + 0.05$$

where $x^{min} = \min_i x_i$ and $x^{max} = \max_i x_i$. We use MLE parameter estimation to fit the distributions and evaluate them using log likelihood, defined as

$$LL(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log f(x_i'|\theta)$$

where f is a PDF parameterized by θ[9]. We obtain the same semantic results when we use Earth Mover's Distance instead of Log-Likelihood as our goodness of fit metric.

3.3 Distribution Results

After performing this experiment, we find a clear hierarchy with the logit-normal performing best, followed by the beta, then the truncated normal, then the normal. Table 1 shows this hierarchy in two ways. First, the average log likelihood across assignments increases from left to right. Second, we can see that the logit-normal is a better fit than the beta 67% of the time, the beta is a better fit than the truncated normal 67% of the time, and the truncated normal is a better fit than the normal 100% of the time[10]. It is a little bit surprising that the beta outperforms the normal slightly more often

[8]$[0, 1]$ may seem like the more natural choice, but both the beta and the logit-normal perform poorly when values are close to 0 or 1 (with the logit-normal unable to produce 0's and 1's at all).

[9]Note that because the PDF's are constrained to the range 0 to 1, our log likelihoods will come out positive.

[10]As mentioned above, this is because the truncated normal's PDF lies strictly above the PDF of its underlying normal.

	Baseline	GPCM	CRM	LNM
Parameters/Item	0	B + 1	3	2
Average RMSE	0.307	0.258	0.278	**0.255**

Table 2: Loss for each IRT model measured across 4115 assignments. "B" refers to the number of scoring categories for a given item.

than the truncated normal does, but this is the only result in Table 1 that is inconsistent with our proposed hierarchy.

Figure 3 shows that our conclusion that the logit-normal is the best choice is robust to sample size. All distributions fit better as sample size increases, since larger numbers of students result in smoother score histograms.

4. LOGIT-NORMAL IRT

Many of our results so far indicate that highly polytomous exam data is also highly structured. Polytomous IRT schemes like the Generalized Partial Credit Model (GPCM) [16] and Graded Response [21] scale with the number of scoring categories per question and thus do not take advantage of structure in the shapes of the item characteristic curves. On the other hand, continuous models like the Continuous Response Model (CRM) are able to cut down on parameters by assuming a parameterized function for each ICC. In this final section, we will propose a continuous model that uses logit-normals to make such simplifying assumptions and thus take advantage of the underlying structure of our data.

We find that when it comes to fitting exam scores, the logit-normal is just as successful when there are smaller numbers of unique scores available. Our model pushes that idea to its limit by modeling each question on an exam with a single logit-normal. The assumption is that highly polytomous items will behave like mini exams and as a result logit-normals will describe them well.

Our model fits a single ability $\theta_j \in \mathbb{R}$ to each student j and (as alluded to above) fits a logit-normal distribution to each item i with parameters μ_i and σ_i. Let S_i represent a random student's score on question i, and let S_{ij} represent student j's score on question i. Our parameters are then related by the following equations:

$$\begin{aligned} \theta &\sim N(0,1) \\ S_i &\sim \text{Logit-Normal}(\mu_i, \sigma_i) \\ E[S_{ij}] &= \sigma(\sigma_i \theta_j + \mu_i) \end{aligned}$$

As in section 2, we refer to our observed data as a matrix A where A_{ij} stores student j's score on question i. In addition, we assume that the data has been shifted and scaled as described in section 3.2. We fit this model in two steps:

1. Use MLE estimation to choose each μ_i and σ_i to fit the observed distribution of S_i's. Specifically, if A_i is the vector of observed scores on question i, we set μ_i and σ_i to be the sample mean and sample standard deviation of $\sigma^{-1}(A_i)$.[11]

[11]Here we are applying the logit function σ^{-1} element-wise.

2. Choose each θ_j to minimize the total squared error of the $E[S_{ij}]$'s. Specifically:

$$\begin{aligned} \theta_j &= \underset{\theta}{\text{argmin}} \sum_i (E[S_{ij}] - A_{ij})^2 \\ &= \underset{\theta}{\text{argmin}} \sum_i (\sigma(\sigma_i \theta_j + \mu_i) - A_{ij})^2 \end{aligned}$$

We use least squares to fit the θ's because we have not defined a probability distribution over S_{ij}, which would be required to perform MLE. More research is required to determine what kind of probability distribution centered at $E[S_{ij}]$ will perform best.

4.1 Evaluating the IRT Models

We will evaluate our model based on how well it can use the parameters it has learned to predict student scores on each question. Note that in-sample evaluation is the norm for IRT [14]. Specifically we will measure the RMSE between the predicted $E[S_{ij}]$'s and the observed A_{ij}'s. If an assignment has n students and m questions, this is calculated as:

$$RMSE(\theta, \mu, \sigma) = \left(\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (E[S_{ij}] - A_{ij})^2 \right)^{1/2}$$

4.2 IRT Results

Table 2 shows the comparison across the whole dataset between our Logit-Normal Model (LNM) and:

- a baseline that uses each student's normalized exam score as a prediction for each question.
- a 2PL Generalized Partial Credit Model (GPCM) [16] fit using EM.
- a Continuous Response Model (CRM) [20] fit using the EM approach described in [24].

The fact that our model has the best performance despite having the fewest parameters indicates that it is taking advantage of the structure of highly polytomous items. We can conclude that discrete models are more complicated than necessary on data like this. We can also conclude that the assumptions about item characteristic curves in CRM are not as good as the logit-normal assumption on this data.

5. CONCLUSIONS

Overall, we have shown that highly polytomous exam data has a large amount of underlying structure that can help us simplify our probabilistic models. Out of three bounded, asymmetric candidates, the logit-normal came out on top as the best prior for exam scores. In addition, the logit-normal's ability to model individual polytomous items allowed us to develop a polytomous IRT model that is simple and well-suited to this kind of data. We hope to have challenged the traditional assumption that the normal is the best prior for student behavior, and we hope that more work will be done to simplify IRT models for the highly polytomous data produced by college-level courses.

6. FUTURE WORK

The main loose end from this paper is the fact that we did not define full probability distributions over S_{ij} in the logit-normal model. Without these distributions, the model is

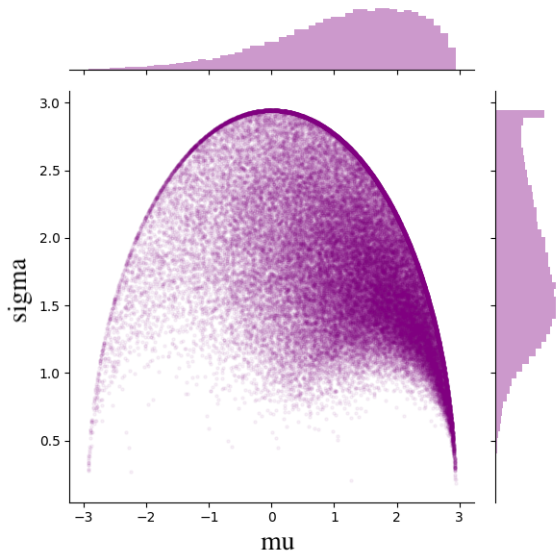

Figure 4: Joint distribution of σ and μ item parameters from the Logit-Normal Model.

able to predict scores and abilities but unable to act as a generative model. Future research needs to be done to find out what family of distributions best models S_{ij}. In addition, Figure 4 shows that there is a strong (somewhat ellipsoidal) relationship between the item parameters of our model, indicating that there may be further structure to exploit in highly polytomous items.

While our results are convincing in the highly-polytomous domain, more work is required to see how well they generalize to less polytomous data. In addition, the logit-normal needs to be tested on downstream tasks like peer grading in order to verify that it is an effective prior for exam scores.

In the interest of reproducibility, and to enable further science, the fully anonymized dataset used in this paper will be made available to other researchers for appropriate academic use. To gain access to the data, researchers must provide IRB approval documentation and must sign an agreement that ensures the anonymized data is treated appropriately.

7. REFERENCES

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

[2] R. B. D'agostino, A. Belanger, and R. B. D'Agostino Jr. A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, 1990.

[3] N. J. Dorans. Recentering and realigning the sat score distributions: How and why. *Journal of Educational Measurement*, 39(1):59–84, 2002.

[4] L. Fendler and I. Muzaffar. The history of the bell curve: Sorting and the idea of normal. *Educational Theory*, 58(1):63–82, 2008.

[5] P. Frederic and F. Lad. Two moments of the logitnormal distribution. *Communications in Statistics-Simulation and Computation®*, 37(7):1263–1269, 2008.

[6] J. F. Gaultney and A. Cann. Grade expectations. *Teaching of Psychology*, 28(2):84–87, 2001.

[7] A. Hald. Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal*, 1949(1):119–134, 1949.

[8] N. L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176, 1949.

[9] V. E. Johnson. *Grade inflation: A crisis in college education*. Springer Science & Business Media, 2006.

[10] I. Jolliffe. *Principal component analysis*. Springer, 2011.

[11] A. Kohn. The dangerous myth of grade inflation. *The Chronicle of Higher Education*, 49(11):B7, 2002.

[12] G. Kulick and R. Wright. The impact of grading on the curve: A simulation analysis. *International Journal for the Scholarship of Teaching and Learning*, 2(2):n2, 2008.

[13] P. Laskowski, S. Karayev, and M. A. Hearst. How do professors format exams?: an analysis of question variety at scale. 2018.

[14] A. Maydeu-Olivares. Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3):71–101, 2013.

[15] B. L. Morgan and A. J. Korschgen. The ethics of faculty behavior: Students' and professors' views. *College student journal*, 35(3):418, 2001.

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

[17] R. O'Dea, M. Lagisz, M. Jennions, and S. Nakagawa. Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature communications*, 9(1):3777, 2018.

[18] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

[19] K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1046. ACM, 2014.

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

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

[22] A. Singh, S. Karayev, K. Gutowski, and P. Abbeel. Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the fourth (2017) acm conference on learning@ scale*, pages 81–88. ACM, 2017.

[23] R. J. Sternberg. The school bell and the bell curve. why they don't mix. *NASSP Bulletin*, 80(577):46–56, 1996.

[24] C. Zopluoglu. A comparison of two estimation algorithms for samejima's continuous irt model. *Behavior research methods*, 45(1):54–64, 2013.