Bridging the Grade Gap: Reducing Assessment Bias in a Multi-Grader Class

Sean Kates* **Tine Paulsen** **Sidak Yntiso** **Joshua A. Tucker†**
NYU NYU NYU NYU

Abstract

Many large survey courses rely on multiple professors or teaching assistants to judge student responses to open-ended questions. Even following best practices, students with similar levels of conceptual understanding can receive widely varying assessments from different graders. We detail how this can occur, and argue that it is an example of differential item functioning (or interpersonal incomparability), where graders interpret the same possible grading range differently. Using both actual assessment data from a large survey course in Comparative Politics and simulation methods, we show that the bias can be corrected for by a small number of "bridging" observations across graders. We conclude by offering best practices for fair assessment in large survey courses.

*Corresponding Author: sk5350@nyu.edu

†SK had the original idea for the paper and supervised the data collection, data analysis, and writing of the paper. SK and JT produced the original research design. SK, TP, and SY conducted the analysis and wrote the first draft of the paper. JT facilitated the incorporation of the grading plan into his 'UA 500: Introduction to Comparative Politics' lecture course in the fall of 2019. SK, TP, and SY all graded three times as many essays and exams for that class than they would have had they not been involved in this research project. All of the authors contributed to the revision of the manuscript. The research carried out for this paper was ruled "exempt" from IRB oversight by ruling IRB-FY2019-2483.

Introduction

Fairness of evaluation is a primary concern in education. Students in large university classes often complain about unfair and disparate grading practices. Such practices can distort students' major choice, performance, labor market outcomes, self-evaluation, and motivation (Lavy and Sand 2018, Lavy and Megalokonomou 2019, Papageorge et al. 2020). In this manuscript, we describe a pernicious form of unfairness that arises when students are assigned different graders with varying severity levels.

We introduce an intuitive method for reducing this kind of bias: a Bayesian implementation of the Aldrich-McKelvey scaling model, where multiple graders grade some assignments. Using real student data from an actual university-level introductory course, we show that even a handful of bridging observations (increasing the workload of graders by less than 10%) can successfully reduce bias by over 50%.

Multiple rater issues are commonplace in political science with applications in roll call voting (Poole and Rosenthal 2000), judicial politics (Martin and Quinn 2002), expert ratings (Clinton and Lewis 2008), survey respondent ratings (Aldrich and McKelvey 1977) and even graduate school admissions (Jackman 2004). Although bridging is not a novel method to handle incomparability (Bailey 2007, Bakker et al. 2014, Pemstein et al. 2015, Marquardt and Pemstein 2018) - given the prevalence of grading bias - we believe its application to grading practice is underappreciated in political science. Alongside this paper, we introduce a new R package that flexibly implements our proposed method for grading data with any number of students, assessments, and graders.

Researchers seeking to advance this line of inquiry further might investigate the comparative benefits of more advanced models. After all, using advanced item response theory (IRT) models to analyze and reduce different types of bias in assessments is a thriving field on its own (Johnson 1996, Johnson and Albert 1999, Wang et al. 2014, Shin et al. 2019).[1] However,

[1]See also the literature on using advanced Rasch models for doing this, eg. Wind et al. 2016 and Wind and Jones 2018.

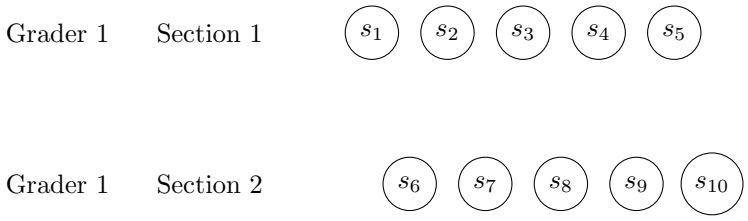
in this paper we attempt to balance not only reduction in bias with the cost of additional grading, but also simplicity of explication. Therefore, the method used to diminish bias will have to be simple enough that college students can understand the intuition behind it, meaning that we face a trade-off between simplicity and reducing bias with which pure theoretical papers do not have to contend.

The Problem: Having Multiple Graders and Achieving Fair Assessment

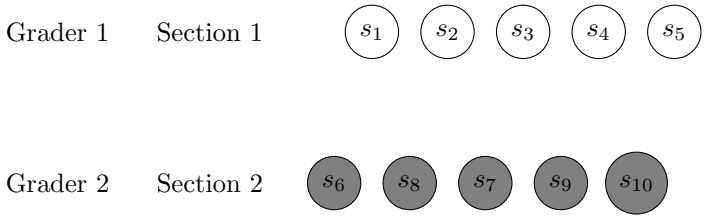
We consider grading bias stemming from having multiple graders as a violation of a specific form of fairness, fairness-through-symmetry (Blackburn 2003). In general, a process is considered fair when it can be expected to produce symmetrical outcomes for identical inputs. In our case, a grading process is "fair" when the grader assigned to a specific student does not, in expectation, affect the grade that student ultimately receives.

Figure 1: Illustration of the bias that can be introduced by having multiple graders.

(a) No grader bias: Having one grader for both sections means it is easy to rank students, even in multiple sections. Ranks for the class can be read across both lines from left to right, so s_1 is ranked as worst student in the class, s_6 is next worst, then s_2, s_7, and on until s_{10} as best student in class.



(b) Grader bias: Same students with same output as in (a), but now split into two sections, each with its own Grader. Grader 1's judgment is the same as in Figure 1 (a), but Grader 2 is more severe than Grader 1. The ranking of students differs from the one in Figure 1 (a) and is biased against those in section 2. s_6 is ranked as worst, s_5 as best. Every student in Grader 2's section loses 1 ranking spot (and every student in Grader 1's section sees their rank increase by 1 spot).



When assessments are carried out in a multi-grader environment, bias can result from a difference in severity across graders or grader error. In this paper, we focus on limiting the first source of bias (severity) and assume that all best practices (i.e. blinding to reduce student-specific error, assessment training to reduce grader-specific error) are being followed to allay the second. We illustrate how difference in grader severity can lead to unfair student assessments in Figure 1.

Proposed Solution: "Bridging" between Graders

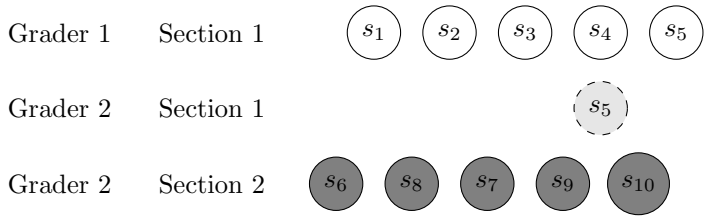
An obvious solution to this type of individual grader bias is to let the same grader review all assessments for a class. Unfortunately, this is an unrealistic solution for large lectures given common restrictions on grader time.[2]

Instead, we put forth the best solution given grader time and resource constraints: bridging across graded groups using a minimal number of bridging observations. By this, we mean that multiple graders assess the same assignment, creating a "bridge" between graders that a model can use to adjust for grader differences in the remaining unbridged observations. In a bridging scenario, some – but not all – of the students in a class will receive grades from each grader. Figure 2 displays a simplified illustration of this solution. In the example, we use a single shared (or "bridged") student to observe that one grader is stricter than another, and thus that fairness demands we adjust the students in both sections to accurately reflect relative performance.

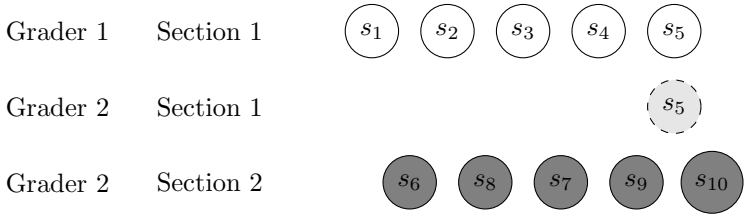
[2]We cover a host of possible alternatives to our method - and why they are found lacking - in Appendix A.

Figure 2: Illustration of how bridging can minimize bias stemming from having multiple graders.

(a) We make sure graders have at least one student assessment in common so we can adjust the assessments based on the difference in judgments between graders for the same student. Here, both Graders assess s_5's assignment, with Grader 2 doing so more harshly, placing s_5 somewhere between s_9 and s_{10}.



(b) When we adjust for the relative difference between Grader 1 and Grader 2 assessments of s_5, we calibrate the rest of the student placements so we arrive back at the same rank order as if there were only one grader.



Bridging assures that multiple graders can assess the same course and assignments while at the same time minimizing the potential bias that comes from having multiple graders.

Bridging is likely to be familiar to a wide variety of political science faculty, although they have not used it in this specific context. For example, comparativists use bridging to ensure that expert evaluators place political parties from different countries on the same left-right scale even if few experts can evaluate multiple countries' parties. (Bakker et al., 2014, p. 1097)[3]. Bridging has also been used to ensure that evaluators with different scales for "democracy" can reliably place countries on a common scale (Pemstein et al. 2015, Marquardt and Pemstein 2018, King et al. 2004). Americanists use bridging to place politicians operating in different environments in the same ideological space (Poole 2007). This famil-

[3]See also Struthers et al. (2019)

iarity with the concept means that bridging is comfortable to use and, at the same time, relatively easy to explain to curious students.

Using bridging to adjust students' grades in this way requires that we maintain some standard assumptions of student assessment. These assumptions include:

- Each grader will be internally consistent in their assessments
- Graders have to be somewhat consistent, i.e., have the same sense of what differentiates a high-quality answer and a low-quality answer
- Graders reward effort roughly "linearly." Performance is treated similarly for high- and low-quality assignments

These assumptions are not likely to be overly restrictive, and must only weakly hold to ensure the success of the process. In the next section, we explain the specific bridging approach we use, and test it on real assessment data to show its promise for reducing bias.

Data and Analysis

We assert that the grading bias described in this paper is a form of "differential item functioning" (DIF), or interpersonal incomparability. Grades given by one grader are not directly comparable to grades given by another, making it challenging to judge students' relative mastery of the material when they are assigned different graders. Aldrich and McKelvey suggested that one could overcome this issue by treating the rankings given to particular stimuli (here, grades given to specific exams) as somewhat distorted perceptions of the true, underlying latent value of the stimuli - here, student's mastery of the material (Aldrich and McKelvey 1977). This type of modeling has been frequently used in political science research to adjust for survey responses by individuals who might perceive survey questions differently, even when the questions asked and scales used are technically identical (see, e.g., Hollibaugh

et al. 2013, Lo et al. 2014, and particularly Hare et al. 2015 whom we follow in casting the process inside a Bayesian framework).

In this specific case, we adopt a simple model of assessment, where an individual student's grade on a particular assignment is a function of the student's underlying skill, attributes of the grader assessing the assignment, and some randomness. We consider two different attributes of the grader. First, we expect that graders might have different baselines for their perceptions of the underlying skill. That is, an "average" performance on a particular assignment may receive a lower or higher score for each grader, depending on this personal attribute.

Second, we also expect that graders may be more or less willing to use all parts of the scoring range. That is, even if the graders start an average student at the same score, they may be more or less rewarding to improvements on that score.[4]

Thus, we model for each grader$_i$ and student$_j$ as follows: $Grade_{ij} = \alpha_i + \beta_i\gamma_j + \mu_{ij}$, where γ_j is the underlying true skill of the student, α_i is the intercept or "shift term" assigned to each grader, β_i is the weight term assigned to each grader and μ_{ij} is the stochastic error term. The two grader-specific terms reflect what we discussed above: the intercept term (α_i) discerns whether the graders have different baseline levels for their grades, while the weight term (β_i) measures the "stretch," or how tightly or loosely an improvement in underlying skill is rewarded with an increase in grade.

In this context, bridged exams serve as common stimuli that allow the grader to approximate a student's ability, adjusting for distorted grader perceptions. By adding bridges, we are adding information for the algorithm that allows it not only to better estimate the underlying latent skill of the student, but also to compare how graders filter the same input through different attributes, and estimate those attributes. By then extending those

[4]In a simple example, imagine one grader who gives poor students failing grades and strong students top marks, versus a different grader who never uses the extreme ends of the scale, even when faced with the same performances by the same students.

attributes to students who did *not* serve as bridges, we can judge the relative performance of all students.

We estimate this model in a Bayesian framework (as opposed to using a maximum likelihood approach) for two reasons. First, the Bayesian approach better handles inherently missing data, which is important here because all students who do not serve as bridges have grades "missing." The Bayesian approach also allows us to understand the uncertainty around our estimated grade distribution better, as we draw from a posterior distribution of all estimated parameters.

In order to estimate the model using a Bayesian approach, we require priors on our estimated parameters - here α, β, γ, and μ. We employ weak priors on each of the grader-specific parameters ($\alpha_i \sim \mathcal{N}(0, 30)$ and $\beta_i \sim N(0, 30)$) and provide a standard normal prior ($\gamma_j \sim \mathcal{N}(0, 1)$) on the underlying skill of a student, so that one can easily rank students, as well as perceive large jumps in the distribution via differences in deviations from the mean.[5] In each estimation, we utilize five chains, each running 30,000 iterations, with the first 2,000 iterations serving as burn-in and thinning the remaining iterations in intervals of 20. Although more complex approaches exist,[6] we believe that simplicity can better ensure that the method is explicable to students with minimal statistics knowledge. Interested practitioners can consult our R package that implements the method.

To evaluate the gains from bridging, we use a simulation exercise on real-world student grading data. We collected grades during a Fall 2018 semester Introduction to Comparative Politics course at a large private research university located in the Northeast,[7] a relatively large course involving 140 students, six review sections, and three teaching assistants. Each teaching assistant graded all students' performance on a midterm examination, a short 5-7

[5] μ's prior is drawn from a gamma distribution over a shape and rate parameter that themselves are each drawn from a gamma distribution $\mathcal{G}(.1, .1)$.

[6] For example, one could allow DIF to occur non-linearly at specific thresholds rather than through linear transformations of the latent skill (see Appendix C).

[7] Student assessments were de-identified. We received IRB approval (IRB-FY2019-2483) to use the de-identified grades for research purposes.

page paper covering material from the course, and a final exam. Free-response and essay-style answers accounted for the vast majority of the available points in all assignments, suggesting that differences in grader perceptions were likely, and likely to be impactful.

We use the averaged grades of all graders on each assessment as the baseline of fairness, as it removes the possibility that grader assignment affected the student's grade. Despite implementing best-practice grading protocols to reduce the potential for bias - graders were trained on a rubric, discussed possible assessments for the same answer, and graded de-identified papers - we find that the traditional grade attribution process produced significant bias.

Considering the difference in student placement on the midterm exam[8] when assessed by a single grader versus the three-grader average, the mean absolute error ("MAE") is approximately 22.5. This means that the average student's rank is 22.5 positions (out of 135 ranked positions for non-missing midterm exams) above or below their "earned" ranking.[9] Rank deviations of this magnitude can ultimately move students multiple grade categories in a class that is curved by the instructor's choice or university rules.

In the simulation exercise, we repeatedly recreate "new" classes, made up largely of students whose only grade we have access to is the one provided by their assigned teaching assistant. However, for some number of students (the number of bridges), we also have their grades as given from the other two teaching assistants - three grades for each of these bridged students. We then estimate the model above, extract the needed attributes to calculate a score for each student, as well as the relative rank of each student compared to all other students. We compare this estimated rank for each student to their rank in our "gold standard" case - where each student receives grades from all three teaching assistants and their

[8]The midterm exam was the first assessment in this course. As such, it provided the best opportunity to see the type of bias described in this paper. Graders were unaware of their own "grader type" - whether they were generally more or less strict than the other graders or had a larger or smaller range of potential grades.

[9]The root mean square error ("RMSE"), which places extra weight on considerable deviations from the appropriate rank, is 27.3.

total grade and rank is determined by the average of these three scores.[10] We calculate the MAE and RMSE for this difference in ranks across this simulation.

We repeat this process 100 times for each number of bridges, which allows us to see how variable our improvement can be depending on the students randomly chosen as bridges. We can also thereby roughly establish upper and lower bounds for how much bias can be reduced given a particular number of bridges. We discuss the results of this exercise in the next section.

Results: Even a Small Number of Bridges Reduces Bias

The simulation exercise tests the extent to which bridging observations can reduce bias stemming from grader assignment, and how efficiently this reduction is carried out. As each bridging item added involves an increase in the amount of grading by one less than the number of graders, increasing their number can quickly lead to a multiplication of work that overcomes the value of bias reduction.

However, as one can see from the violin plot in Figure 3, this concern turns out to be unnecessary. On the figure's x-axis, we plot the number of bridging observations, while the y-axis measures the MAE of a given simulation. For each level of bridging observations, we show the median, interquartile range, and MAE's kernel density derived from the 100 runs of the exercise. Thus, the thicker part of each violin represents where the largest mass of our 100 runs for each bridging level fell in terms of MAE.

Figure 3 shows that the returns to bridging are nearly immediate and quite substantial. We mark the bias associated with the traditional method with a dotted horizontal line (at 22.5). Adding even one bridging observation (which involves two graders grading one additional student each) halves the bias in the large majority of cases. As one continues to add

[10] It is possible that the use of the average grade as a baseline instead of a "true" grade could somehow skew our results. To allay this concern we include a simulation exercise where the "true" grade of each student is known in Appendix C. The exercise shows that our approach vastly reduces bias even in cases where the baseline is the "true" grade.

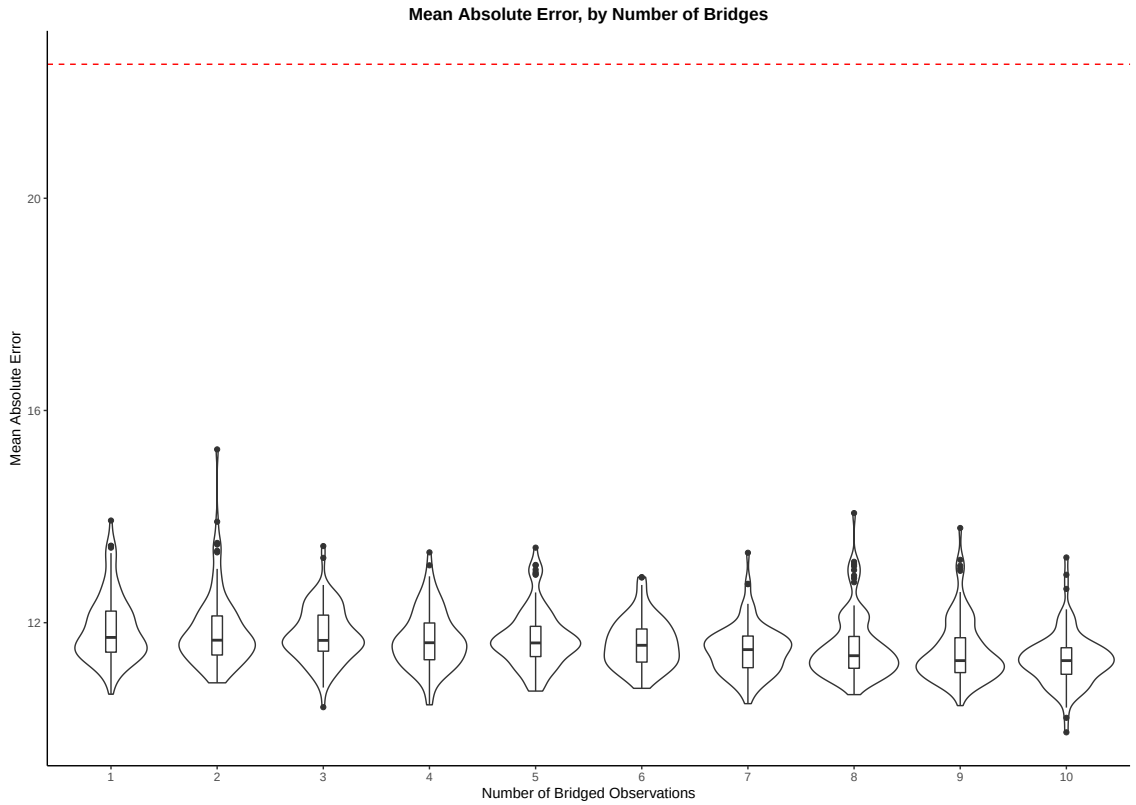


Figure 3: MAE for estimates of student placement (rank) on midterm exam, across number of bridged exams.

bridges, the bias decreases, albeit at increasingly lower rates, with only marginal improvement at the median. Increasing the number of observations can give the assessor a better chance at a "good draw" or at avoiding a poor one that does not reduce bias by as much. Still, even in the worst-case scenarios for any number of bridges, the reduction in bias is substantial.

One might ask how such an improvement is possible with relatively little resource expenditure. We offer both general and specific reasons: First, in general, in grading situations where we have a high enough number of students for each grader, we can generally place a student on a normal curve in that subset of the class. When that student is used as a bridge, we have their relative position on as many curves as we have graders, and all that is

necessary to do is shift each curve so that student is constant across them. This is a different way of saying that bridging allows us to directly extract the intercept difference for each grader. However, the parameters of the curve also gives us some information about the slope term for each grader, information that increases over the number of bridges. Thus, if most of the error is in the shift/intercept term (α), returns to bridging are going to be immediate and large.

In this particular class, the explanation is of this type. One of the three graders was consistently more strict than the other two graders, while also remaining very consistent with the other two graders in underlying rank of the students. In the unbridged situation, this grader's students would be disadvantaged unfairly by their grader assignment, *even though their grader was fairly assessing their relative quality.* However, when we apply a bridging algorithm and readjust students based on it, we immediately adjust this grader's students upward via the intercept term, and we reduce bias by over 50%.

However, we need not rely only on this individual case for evidence of our proposed solution's efficacy. In the Appendices, we pursue a variety of robustness checks and conditionality exercises, based both on the real-life data and simulations. We break down each briefly below and point to where readers can find the results.

Additional Real-Life Results: In Appendix B, we show that our improvements are robust to changes in the chosen outcome metric (RMSE vs. MAE), the item or items assessed (paper and final overall grades vs. just the midterm exam), or the outcome format (letter grade vs. rank). We find that much of these improvements in the actual data arise from accounting for one particular grader's severity; we suggest that practitioners examine the pairwise correlation and MAE between graders to determine the potential gains from the model ex-ante. Finally, we find no evidence that the performance depends on the specific students used as bridges (that is, using only low, high, or extreme scores).

Simulation Results: In Appendix C, we conduct a series of simulations to show that the bridging process results in bias reduction over a broad set of different data generating processes and potential grader pools. These findings demonstrate that the improvements from bridging are not an artifact of treating the average as the "true" grade.

First, we simulate 27 data sets reflecting various degrees of grader reliability (β) and grader shift (α) parameters, as well as grade-level error (μ).[11] The Bayesian Aldrich-McKelvey model outperforms the traditional evaluation method across all data sets except in the limiting case in which graders perfectly agree on each exam (no variability in reliability, shift, or error).

In our secondary simulation analysis, we vary the number of graders (2-5), the number of students per grader (12, 30 and 60), and the level of bias (low versus high variability for all parameters). Our approach outperforms a standard unbridged approach in every combination, but the reductions in bias increase in the number of graders, the number of students, and in variability of the parameters.

Finally, we compare our preferred model's performance to an ordinal IRT model (Marquardt and Pemstein, 2018) - an alternative approach that uses similar bridging concepts to address issues of DIF. We generate nine data sets using an IRT-friendly process that incorporates grading bias via grader-specific ordinal thresholds for mapping latent ability into scores. Both BAM and IRT approaches substantially outperform a traditional regime where there is no bridging. Under moderate or severe DIF,the IRT model outperforms the Bayesian Aldrich-McKelvey model, though BAM is more impactful at lower DIF levels. We discuss how BAM's relative simplicity (both in execution and in explanation) still recommend it over a more complicated IRT approach.

[11] We explore cases with no, low, or high variability for each parameter.

Discussion and Best Practices

In this paper, we show that by creating bridged observations between graders, assessors can severely reduce the bias stemming from grader differences in severity. The reduction can be quite substantial, even at relatively low costs. While the trade-off for any individual instructor will naturally depend on the expected improvements in fairness and the burden of the additional costs, we believe this is the most economical first step in reducing potential bias.

We should note two qualifications: First, this exercise neither precludes nor eliminates all errors. Regardless of how much of the baseline differences between graders we adjust for, there are still stochastic elements and matters of taste that are hard to model. Second, it may be challenging to communicate the process to students unfamiliar with the concepts discussed in this paper and unaware of the bias lurking in more traditional ways of assigning grades. We expect that communication to students is one of the two primary obstacles to implementing this method alongside the implementer's technical know-how.

For communication, we have produced a simple set of slides that use visualizations of the problem and concrete examples to explain how the bias arises and how this method works to fix it. Alongside an instructor's guide with common FAQs and citations to more in-depth explanations of the procedures, these slides should ease communication between instructors and students.

For technical implementation, we have created a simple and straightforward R package that serves as a wrapper for rstan and takes as inputs the bare minimum amount of information from the instructor before outputting a ranking of students. The vignette and package git will have easily reproducible examples so that instructors can become comfortable with implementation before adopting the procedure.

In both cases, we believe our materials will serve as significant first steps toward making the grader adjustment process a regular part of student assessment. But these resources

should not be considered the final word. We hope to start an ongoing conversation on how best to balance student welfare between fairness and ease of understanding the grading process.

The most important contribution of this paper remains the knowledge that it takes an astonishingly small number of bridging observations to dramatically lower bias stemming from having multiple graders. This procedure is a vast improvement from doing nothing to reduce this particular form of bias, which we speculate is the norm in many classroom settings.

References

Aldrich, J. H. and R. D. McKelvey (1977). A method of scaling with applications to the 1968 and 1972 presidential elections. *American Political Science Review 71*(1), 111–130.

Bailey, M. A. (2007). Comparable preference estimates across time and institutions for the court, congress, and presidency. *American Journal of Political Science 51*(3), 433–448.

Bakker, R., S. Jolly, J. Polk, and K. Poole (2014). The European common space: Extending the use of anchoring vignettes. *Journal of Politics 76*(4), 1089–1101.

Blackburn, S. (2003). *Ethics: A very short introduction*, Volume 80. Oxford University Press.

Braun, H. I. (1988). Understanding Scoring Reliability: Experiments in Calibrating Essay Readers. *Journal of Educational Statistics 13*(1), 1–18.

Clinton, J. D. and D. E. Lewis (2008). Expert opinion, agency characteristics, and agency preferences. *Political Analysis 16*(1), 3–20.

Hare, C., D. A. Armstrong, R. Bakker, R. Carroll, and K. T. Poole (2015). Using bayesian aldrich-mckelvey scaling to study citizens' ideological preferences and perceptions. *American Journal of Political Science 59*(3), 759–774.

Hollibaugh, G. E., L. S. Rothenberg, and K. K. Rulison (2013). Does it really hurt to be out of step? *Political Research Quarterly 66*(4), 856–867.

Jackman, S. (2004). What do we learn from graduate admissions committees? a multiple rater, latent variable model, with incomplete discrete and continuous indicators. *Political Analysis 12*(4), 400–424.

Johnson, V. E. (1996). On Bayesian Analysis of Multirater Ordinal Data: An Application to Automated Essay Grading. *Journal of the American Statistical Association 91*(433), 42–51.

Johnson, V. E. and J. H. Albert (1999). *Ordinal Data Modeling*. New York, N.Y.: Springer.

King, G., C. J. L. Murray, J. A. Salomon, and A. Tanon (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review 98*(1), 191–207.

Lavy, V. and R. Megalokonomou (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. *NBER Working Paper 26021*.

Lavy, V. and E. Sand (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics 167*, 263–279.

Lo, J., S.-O. Proksch, and T. Gschwend (2014). A common left-right scale for voters and parties in europe. *Political Analysis 22*(2), 205–223.

Marquardt, K. L. and D. Pemstein (2018). IRT Models for Expert-Coded Panel Data. *Political Analysis 26*, 431–456.

Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis 10*(2), 134–153.

Papageorge, N. W., S. Gershenson, and K. M. Kang (2020). Teacher expectations matter. *Review of Economics and Statistics 102*(2), 234–251.

Pemstein, D., E. Tzelgov, and Y.-t. Wang (2015). Evaluating and improving item response theory models for cross-national expert surveys. *V-Dem Working Paper 1*.

Poole, K. T. (2007). Recovering a Basic Space From a Set of Issue Scales. *American Journal of Political Science 42*(3), 954.

Poole, K. T. and H. Rosenthal (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Shin, H. J., S. Rabe-Hesketh, and M. Wilson (2019). Trifactor Models for Multiple-Ratings Data. *Multivariate Behavioral Research 54*(3), 360–381.

Struthers, C. L., C. Hare, and R. Bakker (2019). Bridging the pond: Measuring policy positions in the united states and europe. *Political Science Research and Methods*, 1–15.

Wang, W. C., C. M. Su, and X. L. Qiu (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement 51*(3), 260–280.

Wind, S. A., G. Engelhard, and B. Wesolowski (2016). Exploring the Effects of Rater Linking Designs and Rater Fit on Achievement Estimates Within the Context of Music Performance Assessments. *Educational Assessment 21*(4), 278–299.

Wind, S. A. and E. Jones (2018). The Stabilizing Influences of Linking Set Size and Model–Data Fit in Sparse Rater-Mediated Assessment Networks. *Educational and Psychological Measurement 78*(4), 679–707.

Appendix A: Alternative Options for Addressing Bias

We do not claim to be the first to identify this particular form of bias or attempt to correct it. Concern over bias stemming from multiple graders assessing different students has forced assessors to adopt various solutions that each display some weaknesses. In this Appendix, we examine several alternative solutions and discuss why we prefer the method presented in this paper.

Perhaps the most straightforward solution is to have one grader assess the entire course. However, in the large survey courses common to both public and private universities, the resources necessary to deal with multiple assessments can escalate quickly. Assessment can demand a tradeoff from resources expended on pedagogy, decreasing lectures' caliber, and lessening education quality overall. Moreover, a single individual responsible for all grading may find it challenging to conduct the assessments in a reasonable amount of time, such that the returned assignments serve as a learning tool for students while the material remains fresh. In general, we find that in large courses, a single-grader regime is both practically infeasible and pedagogically undesirable.

Instructors can also mitigate bias by reducing assignment subjectivity - that is, by constraining the opportunities for the graders to assess identical students or responses differently. For instance, one might design an assignment where all questions are closed-ended, with one specific correct answer (e.g., multiple-choice questions, or "fill-in-the-blank" questions). This practice eliminates graders' capacity to see similar answers differently; graders can adjudicate differences that somehow arise by reference to an answer key. However, there exists a wide range of "knowledge or skills that may not be easily or plausibly assessed" by using only multiple-choice questions (Braun 1988, p. 1). This is particularly true for large survey courses, where instructors likely want to see students engage with the material in various fashions, not merely through rote memorization.

Given the practical and pedagogical benefits of having multiple graders assessing more

subjective responses, instructors must shift to reducing the bias we describe in the article rather than eliminating it. One commonly applied solution is to structure the assessment process in such a way that each grader is only responsible for grading a portion of the overall assignment, and each part of the assignment is assessed by only one grader. [12] If we assume that a single grader is internally consistent in assessing questions (an assumption held throughout this paper), then this solution would seem to reduce the potential for bias due to grader subjectivity.

However, there are both practical and theoretical concerns with this solution. Practically, one might be concerned about three different issues of increasing severity. First, the instructor has to construct each assignment to equalize grader difficulty across sections. This concern is fundamentally a matter of fairness across graders, which is perhaps a secondary concern in assessment, but not entirely trivial.

Second, logistics for assessors become more difficult under this assessment scheme. Because each student's assignment is graded in part by n assessors, every assignment must be exchanged at least *n-1* times. Assignment swapping increases both the probability of adverse outcomes (misplaced exams, exposure of students' private data if exchanges are anything other than face-to-face, etc.) *and* the time between the completion of the assignment and its assessment and return. Timely remediation of mistakes is essential for many learning outcomes, particularly in subjects that build on a shared understanding of basic principles.

The final practical concern involves this remediation more directly. Students who question their assessment and investigate how they can improve must pursue multiple sources for information and feedback. Time-consuming assignment remediation places unnecessary barriers to learning and can be a source of frustration. It can also break the natural relationship between student and instructor, placing intermediaries at the forefront of one of the core aspects of instruction.

[12] Consider as an example, an exam comprised of three essay questions, where all students have their first essay assessed by Grader 1, all of the second essays are evaluated by Grader 2, and Grader3 handles each of the final papers.

Table 1: Example of Potential Unfairness Produced when Instruments are Divided by Graders

	Section 1 (rank)	Section 2 (rank)	Section 3 (rank)	Final Score (rank)
Student A	10 (1)	10 (1)	1 (3)	21 (3)
Student B	9 (2)	7 (2)	6 (2)	22 (2)
Student C	7 (3)	6 (3)	10 (1)	23 (1)

These practical concerns are real, but they could be addressed and overcome if they were the only issues confronting this approach. However, this method also surreptitiously creates a different form of bias. While eliminating the bias stemming from differential item functioning, this method concentrates the bias stemming from differences in grader reliability. Intuitively, this method makes the section graded by the highest variance assessor more determinative of the overall grade than it otherwise might be.

Consider a simple example where three graders each assess one part of a 30-point exam, with each part worth 10 points. If two graders perceive the "actual" range of viable grades to be between 6-10, but the third grader utilizes the whole range, there are a host of outcomes that will seem (and arguably *be*) unfair. Table 1 proposes one such distribution, where the third assessor uses the full scale, but the other two graders use the truncated scale.

Here, a student who excelled in two of the three sections (student A) receives the lowest score in the class because they were "unlucky" to have done the poorest in the section of the exam graded by the assessor with the largest range. Similarly, a student who did poorly in two of the three sections still achieves the highest overall grade in the class by excelling in the one section with the highest variance. While this outcome *may* reflect the underlying overall ability of the students, it can just as likely be an artifact of a multi-grader setup. The method we propose in this paper addresses not only the severity of the grader, but also

their natural variance, attempting to bring both in line in order to achieve fair results for all students.

Appendix B: Additional Results and Robustness Checks

This appendix provides robustness checks for our analyses. First, we show that our proposed solution is robust to the choice of error metric, replacing mean absolute error (MAE) with root mean square error (RMSE) as a measure of grade bias. We then show that our method reduces bias across a range of possible assignment and grading situations, including a second exam (the final paper from the class), the aggregate final grade in the course, and the assigned letter grade in a class with a designated grading curve/distribution. In each case, the method requires only a small number of bridging observations to dramatically decrease bias in the assessment of interest.

We also create "simulated" classrooms with fewer graders than in our observed classroom. We show that even in this case, there are gains to be made by applying the BAM algorithm, and that these gains are contingent on the relationship between graders.

Finally, we attempt to contextualize the efficacy of the algorithm by varying the types of inputs it receives. We test whether being able to bridge only on specific types of grades (low scores, high scores, extreme scores) can improve our performance. In general, we find little difference in performance across these regimes, though again all three potential worlds see vast improvement from a grading scheme that uses no bridges.

RMSE vs. MAE in the Midterm Exam

In the main body of the paper, we show that our bridging method dramatically reduces MAE in the assessment of the midterm exam (see Figure 3 in the Results section). In this part of the appendix, we show that the same bridging method also reduces RMSE.

Figure 4 replicates the graph produced in Figure 3, with RMSE replacing MAE as the metric of error. As a measure of error, RMSE is more responsive to very large errors, and it may be that in our practical situation, this is the type of error that is most desirable to avoid. In particular, large errors in rank are likely to have large grading consequences

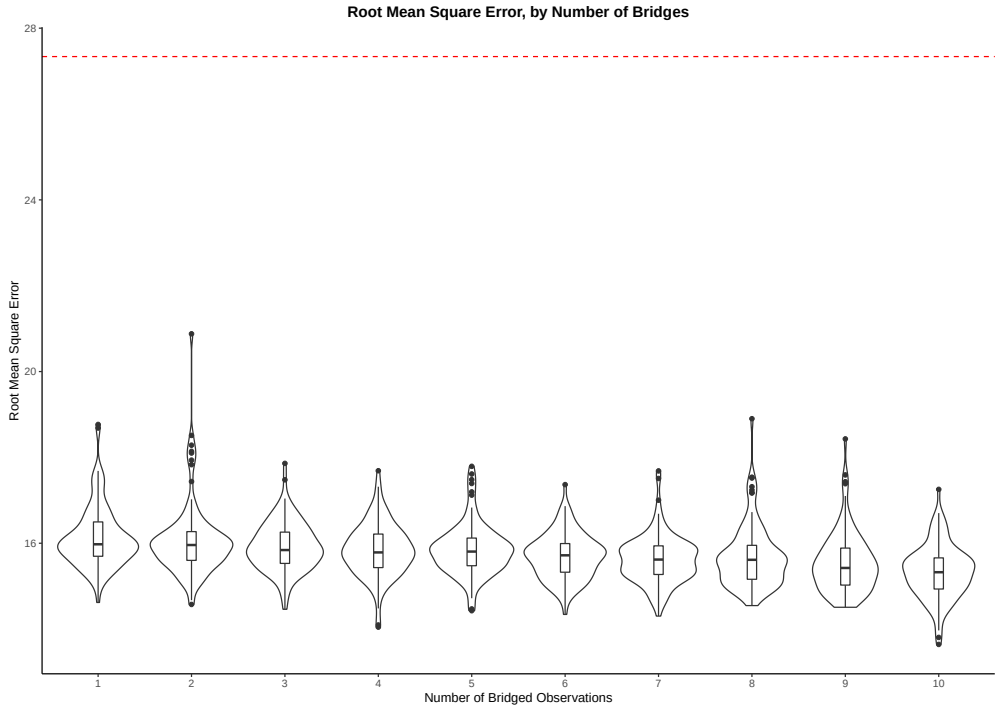


Figure 4: RMSE for estimates of student placement (rank) on final exam, across number of bridged exams.

where continuous ranks are converted into discrete letter grades. As was mentioned in the main body of the paper, the RMSE for students without bridging was calculated to be 27.3, a substantial error in a class of 135, roughly equivalent to 20% of the entire distribution and likely reflective of large numbers of students being assigned the wrong letter grade. As we see in Figure 4, our bridging method vastly reduces this error, and again within a very limited number of bridging observations. Most importantly, we reduce likely error such that students should expect to be assigned proper letter grades in courses where grade ranges encompass more than 10% of the distribution.[13]

[13]That is, when the suggested distribution looks something like: 15% of the class gets A's, 20% get A-/B+, 30% get a B, etc.

Numerical Score on the Midterm Exam

Regardless of how we measure error, then, our approach captures a more accurate assessment of the students' relative positions on the assignment. Conversely, one might be more concerned with merely getting the "correct" numerical assessment, on a common scale across students. This is, of course, an underlying input to appropriately ranking the student, but it presents another strength of our approach. Not only do we accurately rank students, but we more faithfully capture the distance between students at an absolute level.

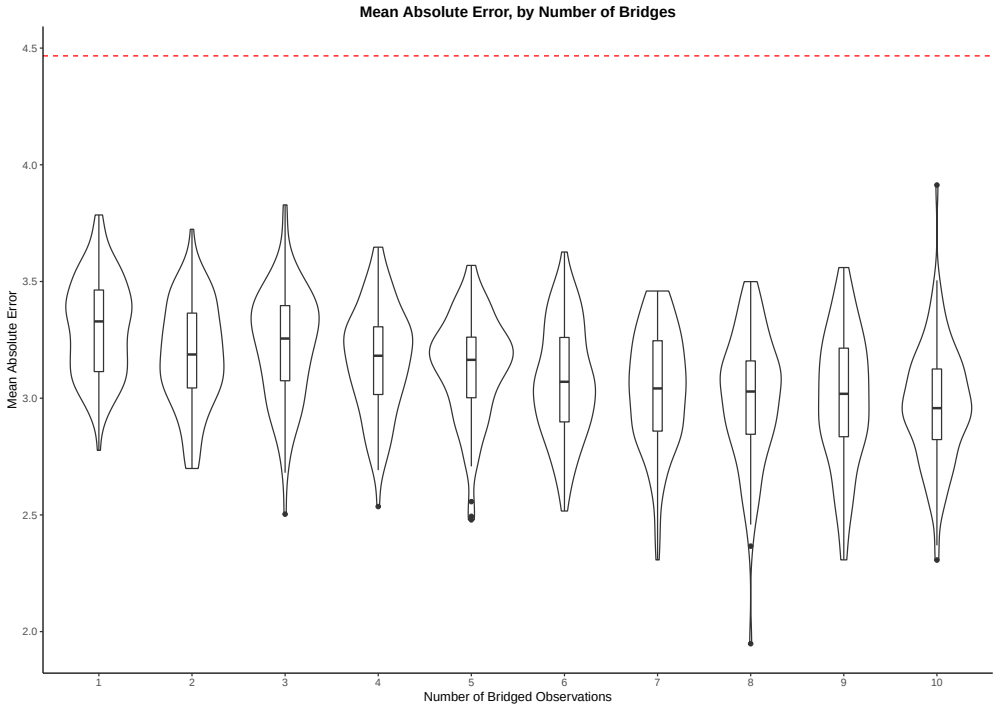


Figure 5: MAE for estimates of student score (out of 45) on midterm exam, across number of bridged exams.

To test whether we also show improvements in this vein, we use an identical process as described in the main text of the article, but with the intention of measuring how far the estimated numerical grade from our bridged model is from the average of the three numerical grades given by the three graders. In Figure 5, we visualize the results in our traditional violin plots. Again, we see clear and near immediate improvement. In the traditional grading regime, the average error between a student's assigned grade and the grade averaged across

the three graders was nearly 4.5 points, or 10% of the grade range (4.46 points, out of a maximum score of 45 points).

When we use the bridging method, this average error decreases dramatically. The median improvement after only 2-3 bridges is nearly 30%, with the potential for much larger improvements depending on the luck of the draw, particularly if we further increase the number of bridges. The bridging method helps to decrease error in raw scores, as well as ranks.

Other Assessment Types - Paper with Letter Grade

Our analysis in the main paper focused on one assessment - the midterm exam of the course. In this subsection of the appendix, and those following, we show that this choice does not drive our results. Rather, regardless of the instruments we use, or the way we think about assessment in the aggregate, the bridging process helps to reduce grading bias.

In addition to the midterm, students also completed a paper of between 5-7 pages, where they were asked to assess a particular scenario using information gleaned from the course. Each paper was graded by all three graders, with the graders assigning the paper a letter grade that could contain a plus or a minus.[14] One might be concerned that assignments of this type - naturally more subjective, but also with a more discrete grade distribution - would trouble our approach, but we show this not to be the case.

Once again, we take as a baseline "correct" grade the average of the numerical equivalents for each of the letter grades given a paper by the graders.[15] We run simulations of the type described above in both the main portion of the paper and the previous subsections of this Appendix, and measure how each run of a specific number of bridged observations improves assessment by reducing the bias produced by grader assignment. Because the ranks here are naturally chunky (all students are ultimately given one of six or seven letter grades and

[14]The possible choices for grades then, were A (4.0), A- (3.67), B+ (3.33), B (3), B- (2.67), C+ (2.33), C (2), C- (1.67), D+ (1.33), D (1), D- (0.67), and F (0). However, no one who completed the paper on time received a grade lower than a C+, so the realized range was more limited. For the purposes of this exercise, we focus only on students who completed the paper on time.

[15]Note that this average score, in itself, might not always easily translate back into a letter grade.

thus there are large numbers of ties), we focus on the difference in score from the "correct" average. These results, in our traditional violin plots, are in Figure 6.

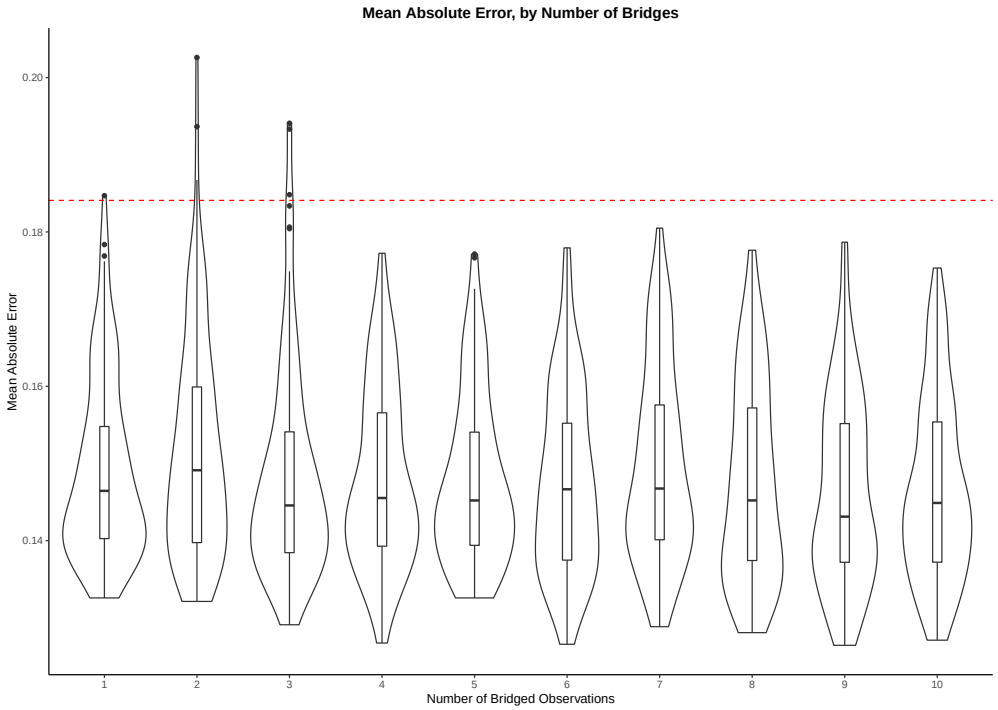


Figure 6: MAE estimates of paper score (out of 4.0), across number of bridged exams.

As one can see, MAE for the traditional method of grading is about 0.183 points, or more than half a step in the grading scale. Adding only 3-4 bridging observations drops our expected error approximately 25%, and below where we would expect to make many errors in grade assignment. It should be noted that because of the limited scale of grades we worked with in the paper assessment, the performance is slightly more noisy than for other graded items, though again the decreased bias is large and substantively meaningful.

Other Assessment Types - Aggregation

An instructor might ultimately be most concerned not with any single assignment, but with the final assessment and ranking of students.[16] In this final subsection, we show the

[16] In general, we are not of this opinion, particularly as properly assessing student performance on individual assignments allows instructors to target properly students in need of additional attention in a timely fashion, but this may vary by educational situation.

cumulative reduction in bias at the final aggregation phase. As a reminder, this still requires that you have students whose entire work product has been graded by multiple graders, and thus does not "save" any work in that fashion.

For this exercise, we consider the MAE of the final grade (theoretically on a 0-100 scale, but practically in the range from 63-99) as calculated at the assessor level. Thus, we calculate a final grade for each student from each assessor as the result of all the inputs to a final grade from that assessor. We then use these final grades in the same fashion as the exam scores in the main analysis and the above subsection. Figure 7 displays the results. Again, the bridging method reduces bias by over 50%.

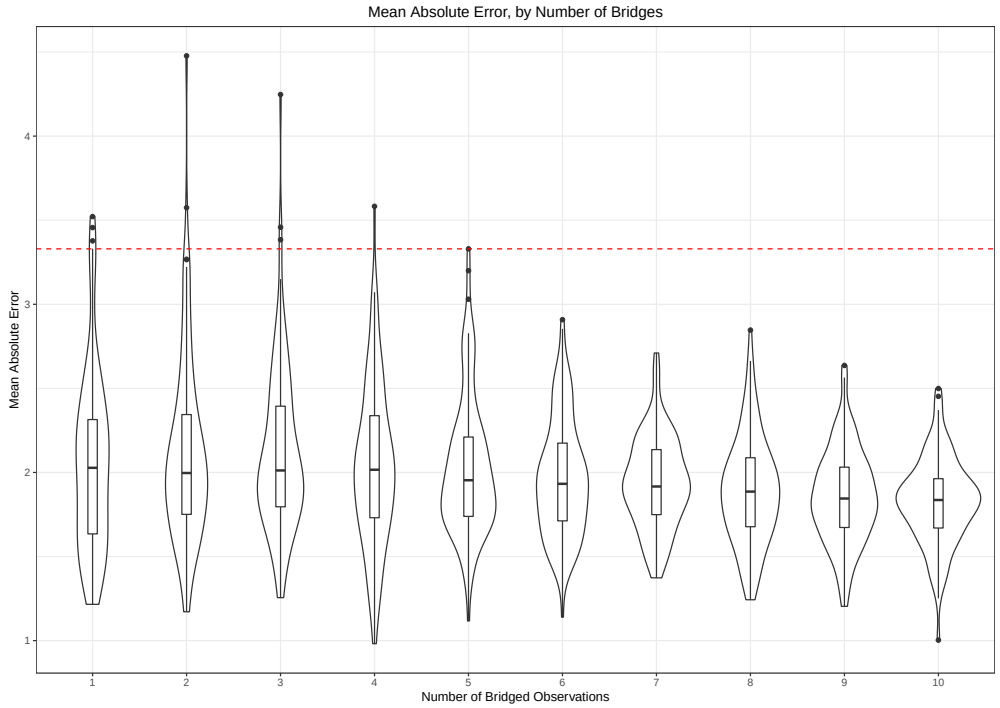


Figure 7: MAE estimates of student course grade (out of 100), across number of bridged evaluations.

Finally, we extend this analysis to look at the final raw grade letters (i.e. A, B, C, etc.) that students would receive under alternate assessment schemes. In Figure 8, we show that using the bridging method on a final letter grade reduces bias by approximately 65%. This is equivalent to 32.5 students (in a class of 135) receiving a proper letter grade

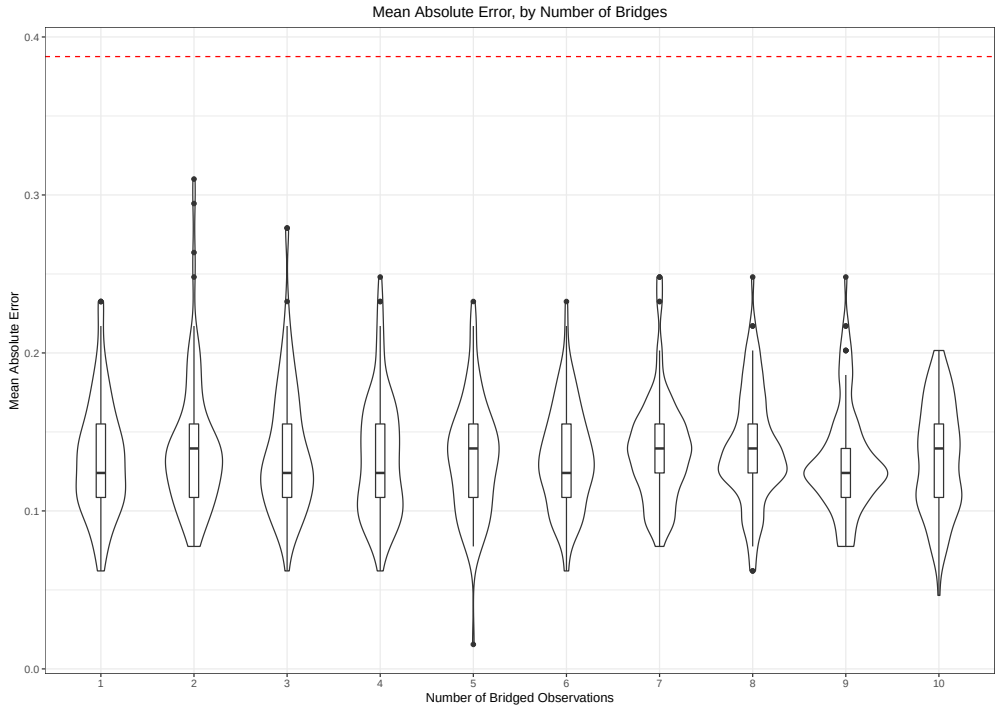


Figure 8: MAE estimates of student course letter (out of 4.0), across number of bridged evaluations.

a full step above/below their incorrect grade (i.e. an "A" when a "B" was given) or 97.5 students moving a small step in the correct direction (receiving a "B+" after being given a "B."). While we ultimately suggest applying the method to each individual instrument of assessment (which would have long-term improvement in the overall score as well), there are important gains to be had simply by applying it for one final grade.

Efficacy of the Approach over Multiple Assessments

Student assessment can and should be a dynamic process, where assessors can learn from past outcomes as naturally as students do. Many of the biases we attempt to identify and correct for using our approach can also be proactively reduced if assessors are a) informed of the discrepancies between their levels of strictness and variance, and then b) use that information to adjust their own behavior. In the course that served as the source of data for this paper, all three graders were well-informed of the grades their assigned students

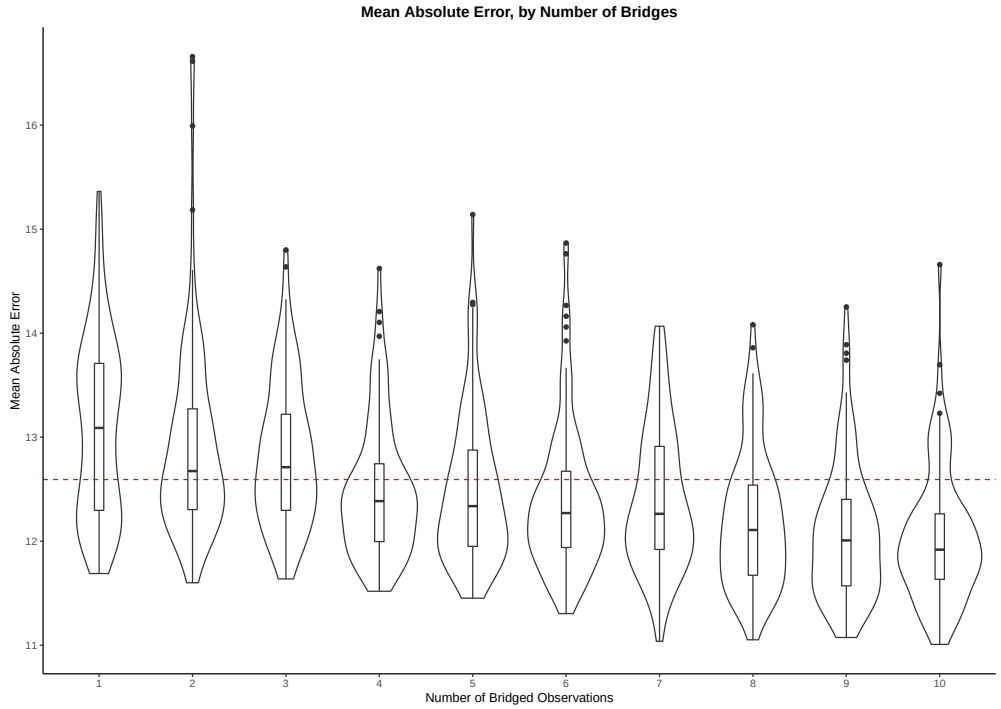


Figure 9: MAE estimates of student rank on final exam (out of 135), across number of bridged evaluations.

had received from the other two graders, and could easily gauge their relative position on the laxity and variance scales. Ultimately, this led to a reduction in the bias from even traditional grading methods over time, and a concomitant reduction in the efficacy of our method in the final stages of assessment.

The students' final exam was identical in style to the midterm that served as the first assessment for the students, but was worth 75 total points. While the average error (in points) for the traditional method on the midterm was nearly 10% of the grade ($\tilde{4}.5$ out of 45 possible points), the MAE for the traditional grading format on the final exam was just more than a third of that, at 2.84, or 3.7%. At that level, there is very little room for reduction in bias, even as we extend the number of observations, as well as very little substantive reason to do so.

And in fact, Figure 9 shows how our approach only slightly outperforms the traditional

grading method in assigning student ranks, and even that improvement is conditional on getting a not unlucky draw of bridges.

We display these results as a reminder that adopters should recognize the benefits and limitations of our approach. Our approach adjusts mechanically for bias that can, at least in some instances, be eliminated with increased information and dedicated efforts by assessors. However, that information (and specifically, information about the relative laxity of assessors) itself must come from somewhere, and we would suggest that adopting a bridging technique for at least the first or first few assessments may allow assessors to recognize their differences and adjust their behaviors.

Classroom with Fewer Graders

The course from which our main data is gathered has a constant structure: 3 main graders, without approximately 45-50 students assigned to each. Thus, our analysis of the real-life ramifications of grading bias is somewhat restricted for some variables we believe may matter. One such measure is the number of graders in the classroom. We cannot reasonably *add* graders to our real-life data retroactively, and so are limited in that direction. However, we can artificially construct classes with fewer graders, and see how well our proposed solution performs. In this subsection of Appendix B, we do so for the three possible combinations of two graders, and verify that the algorithm succeeds in reducing grading bias in each of those scenarios. We then more systematically explore the performance of our proposed solution under varying number of graders and students in the simulations presented in Appendix C.

For this exercise, we construct a "course" by eliminating one grader, as well as the students for which that grader was the primary grader. We do this once for each of the three graders, leaving us with three different courses of similar sizes. In each of these artificial courses, we conduct the same analysis as we do in the main paper, selecting some number g students to serve as bridges in each of 100 iterations. We compare the ranks of each student in the

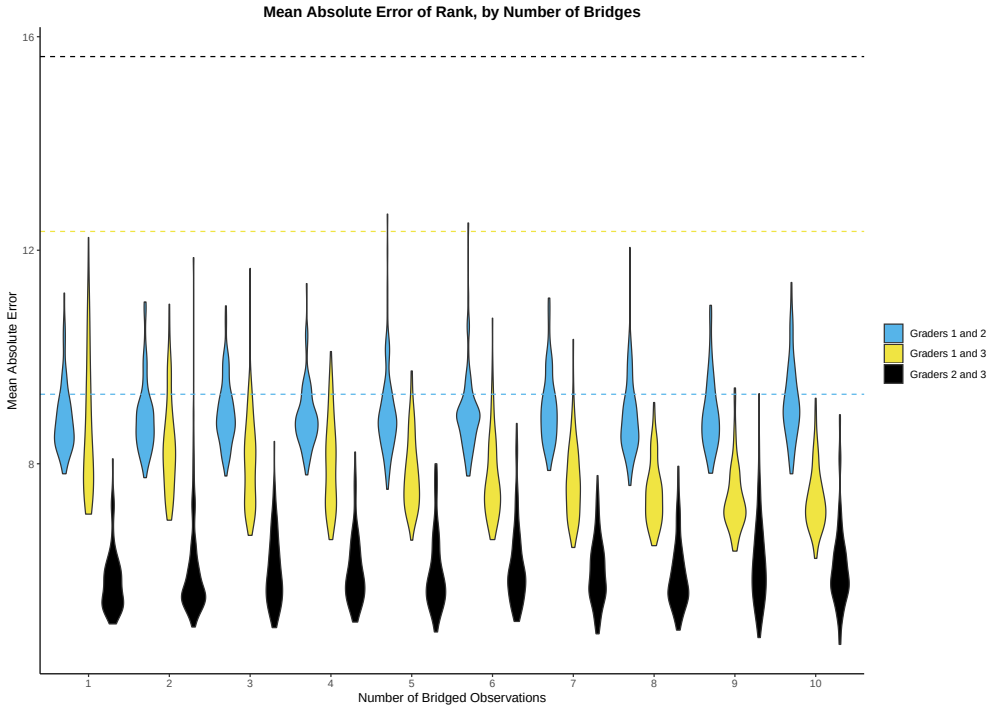


Figure 10: MAE estimates of student rank on midterm exam (out of 135), across number of bridged evaluations. Each violin plot represents a "class" combining two of the three graders for the real-life course.

bridged simulation to the student's "correct" rank, were their grades averaged across each of the graders, and calculate the Mean Average Error across students.

Figure 10 reflects our findings. In all three cases, we see improvement on grading bias in most cases, and in two of the three artificial classes the improvement is substantial. The patterns in improvement across the three classes (all comprised of two of the same three individuals) is illuminating, however.

It is a fact of the course that one of the graders was far "harsher" than the other two. This grader (Grader 3 in the groups imagined in Figure 10) consistently gave students lower marks than the other graders - which marks, if not adjusted, would have penalized students assigned to Grader 3. Thus, in the two courses where Grader 3 was one of the assessors, there is a much higher chance of grading bias of the type we describe in this paper. *BUT*, it is also the case that Grader 3's rank order of students had a higher correspondence to the

rank orders of each of the other graders than their rank orders had with each other. Thus, after we apply the bridging technique, and the shift from Grader 3 is accounted for, the classes with this grader are less subject to bias than the class without this grader. Bridging can greatly reduce bias related to shifted perceptions of the same underlying performance, but it cannot "fix" when graders fundamentally disagree on the relative performance of a student.

Varying the Attributes of Students Used as Bridges

Finally, we use the data gathered from the live course to judge whether it makes a difference *which* observations serve as a bridge. One might think that a particular type of observation or mix of observations would provide us with better bias reduction. In this subsection of Appendix B, we look at three specific possibilities. In the first, we use only scores from the bottom third of the distribution to serve as bridges. In the second, only those scores in the highest third. In the final exercise, we evenly divide the bridges over extreme scores, taking $N/2$ scores from the highest third and the same number from the lowest third to serve as our N bridges.[17]

In each of these regimes, we conduct the same analysis we have conducted throughout this paper, varying the number of bridges over 100 simulations where we select different possible combinations of bridges that follow our regime rules.

Figure 11 visualizes the results of this process. Each regime is represented by a different color, and each violin plot represents the distribution of Mean Average Error of student ranks for a particular regime under a particular number of bridges. In all cases, the default grading scheme has an MAE of 22.5, so every possibility is an improvement. However, there is no regime that outperforms the others consistently and to a significant extent.

This makes sense. The assumptions of the model roughly require graders to have the same stretch and shift parameters for students throughout the entire range of the underlying

[17]Note that this means we only analyze this regime under an even number of bridges.

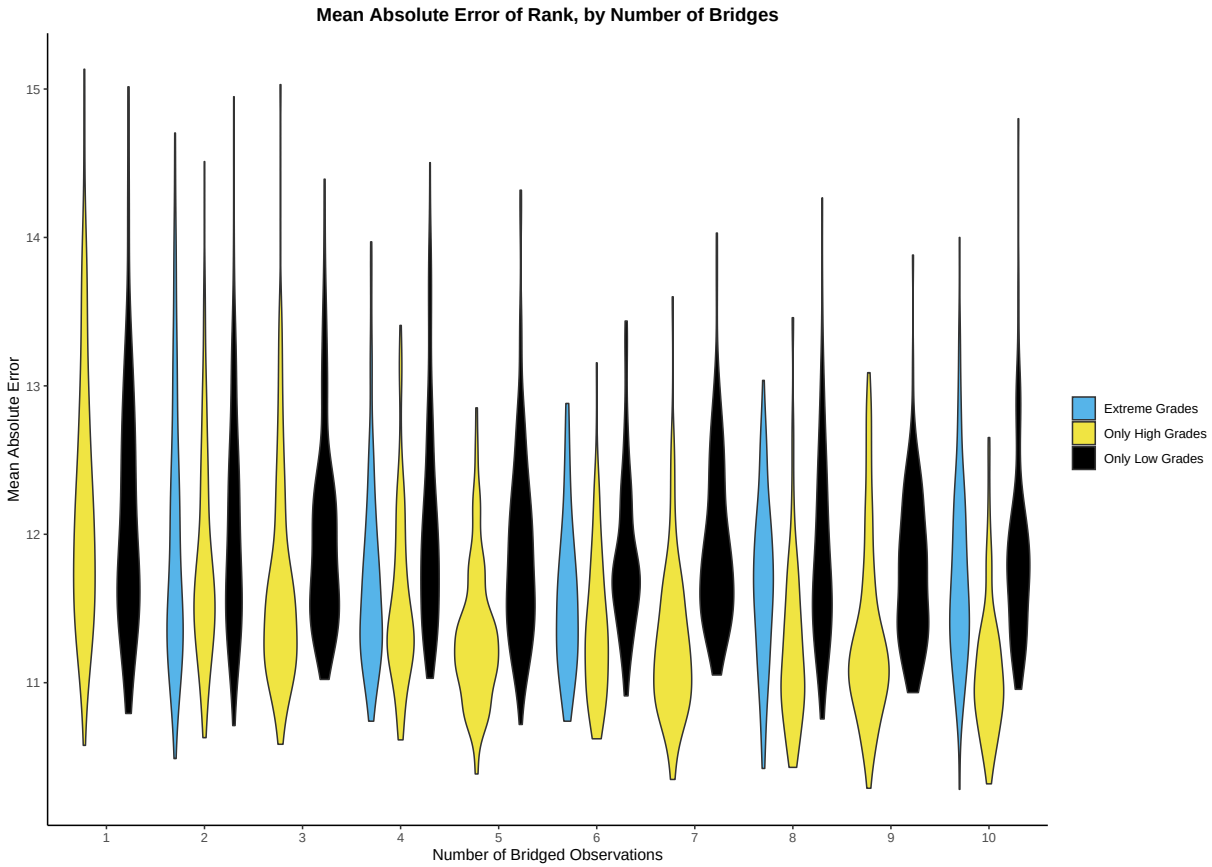


Figure 11: MAE estimates of student rank on midterm exam (out of 135), across number of bridged evaluations. Each violin plot represents a regime where the bridged observations are either all from the lowest third of the distribution, all from the highest third of the distribution, or split evenly between the highest and lowest thirds. The MAE for an unbridged process is approximately 22.5.

latent skill trait. We should then, in theory, extract the same amount of knowledge about the grader's perceptions from units at any part of that range. Generally, this is what we do observe. There is some slight evidence that inputting only bridges from the very highest third of the range may decrease bias a bit more, but not to anything approaching a significant degree, and not something we would expect to be repeated in other classes.

Rather, it is likely evidence of a peculiarity in this specific grading situation, where graders are more consistent at the highest ends of the spectrum than at the lowest. This may accord

with our natural expectation that most graders are very consistent in rewarding good work, but have varying beliefs about how harshly to punish particularly poor work.

Note also that it is not entirely clear how one might leverage differential success across regimes, even if it did exist. Doing so would require identifying *prior to bridging* those observations that would qualify as low, high, or extreme observations. Professors might be able to use pre-class GPAs or the results of a short assessment meant to group students by skill, but there would be no way to ensure that these proxies would reliably identify the best students to serve as potential bridges. Our case (where we know beforehand that the observation is of a specific quality) is the best case scenario, and still we find no reliable benefit to choosing bridges in this manner.

Appendix C: Simulation Evidence

Varying Parameters

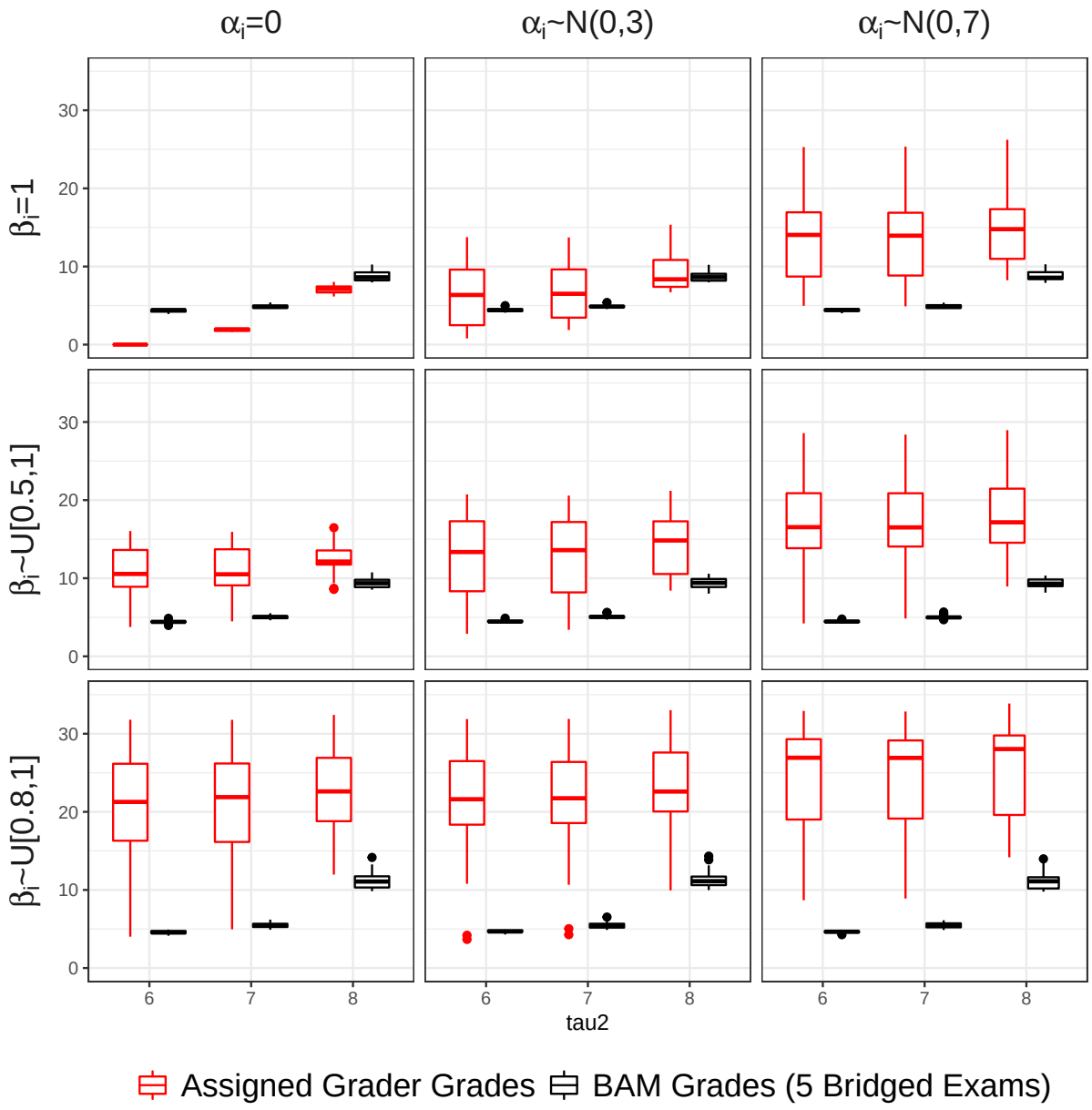
In this appendix, we apply simulation methods to bolster our argument regarding the bias-reduction qualities of bridging, and evaluate the robustness of its gains. We simulate 27 different datasets reflecting various degrees and forms of grader error. Each dataset presumes 3 graders and takes as its latent trait inputs the distribution of the average final exam grades of our course, which had a mean of 55 and a standard deviation of approximately 10 over 142 observations.

In each simulated dataset, grader reliability (the stretch parameter) is held constant ($\beta_i = 1$), allowed to vary ($\beta_i \sim \mathcal{U}[0.8, 1]$) or allowed to vary greatly ($\beta_i \sim \mathcal{U}[0.5, 1]$). The grader shift parameters are similarly held constant or allowed to vary ($\alpha_i = 0$, $\alpha_i \sim \mathcal{N}(0, 3)$ or $\alpha_i \sim \mathcal{N}(0, 7)$). Finally, grade-level error is allowed to vary over wider ranges or held constant at zero ($\mu_{ij} = 0$, $\mu_{ij} \sim \mathcal{N}(0, 0.5)$ or $\mu_{ij} \sim \mathcal{N}(0, 2)$). For each simulated dataset, we evaluate to what extent five bridging observations can reduce overall grading bias for the final exam.

Figure 12 presents the MAE estimates for bridged (black box-and-whiskers) and non-bridged (red box-and-whiskers) rank placements across the different datasets. Rows represent different levels of variation in grader strictness (α_i), while columns represent different levels of variation in grader reliability (β_i). Within each cell, rows represent different levels of grade error (μ_{ij}). The plotted results are in a traditional box-and-whisker format, displaying the simulations with the lowest error, the 25th, 50th, and 75th percentile simulation, and the simulation with the highest error.

In the limiting case of no grader distortion (where $\beta = 1$, $\alpha = 0$, and $\mu = 0$), grading without bridging outperforms the bridging exercise. In this simulation, all three graders give each exam the same grade, so it does not matter which grader a student is assigned. Their

Figure 12: MAE Simulations
Final Grade

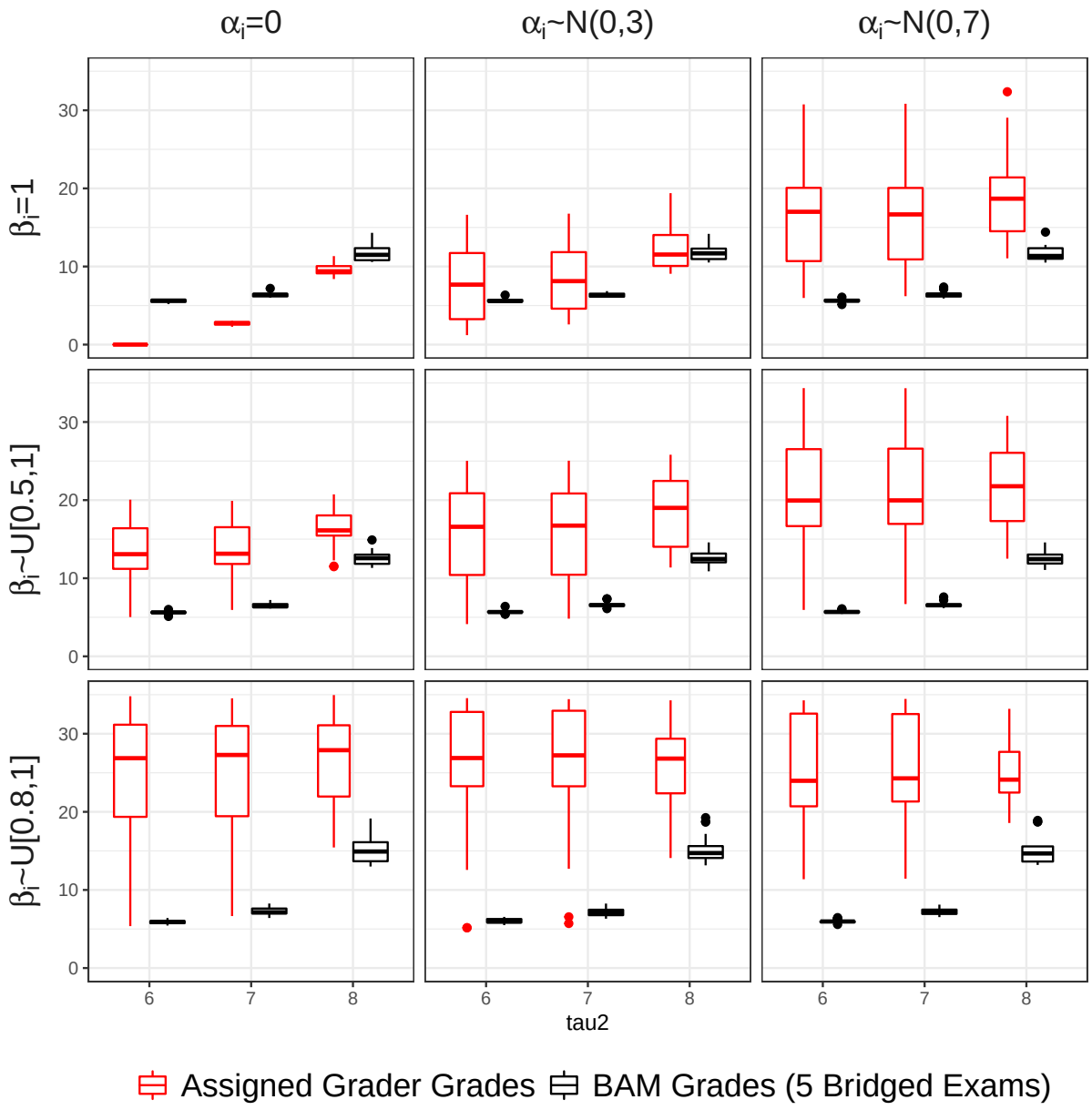


MAE estimates of final exam placement across simulated datasets. Black whiskers represent estimates from the BAM model with five bridges. Red whiskers represent estimates from the typical (assigned grader) assessment method.

latent skill will be directly translated into their grade. As soon as any error is introduced, however, bridging provides immediate and large benefits.

In the presence of any variation in grader reliability or strictness, bridging provides sizeable (two or three fold) reductions in MAE. Figure 13 presents the RMSE estimates, producing qualitatively similar results. When graders are most different (when βs are drawn from a wider range, and αs can be larger - when we move to the right in columns and down in rows), the gains are starkest and even the worst possible draw of simulations vastly outperforms doing nothing in terms of grading fairness.

Figure 13: RMSE Simulations
Final Grade



RMSE estimates of final exam placement across simulated datasets. Black whiskers represent estimates from the BAM model with five bridges. Red whiskers represent estimates from the typical (assigned grader) assessment method.

Comparison Over Size of Classes and Numbers of Graders

In Appendix B, we analyzed our real-life data in "simulated" classrooms where only two graders actually had students. We found that the proposed solution still reaped large benefits in bias reduction, but our findings were naturally limited by the course structure itself. In this subsection of Appendix C, we further push on how the efficacy of our proposed solution is conditioned by the number of graders, the number of students in the class, and the type of graders/world that the class takes place in.

Specifically, we create a simulation environment where we iteratively test how increasing either the number of graders, the number of students per grader, or the reliability and strictness of the graders affects the gains from bridging across students, for a particular number of bridges (in all of these cases, we use only 3 bridged observations). We let the number of graders vary from 2 to 5, and the number of students per grader vary between 12, 30, and 60 students per grader.

This gives us 12 possible combinations of graders and students, with the smallest (2 graders and 12 students per grader) approximating a co-taught seminar course and the largest (5 graders with 60 students per grader) more closely approximating an introductory level core course that draws hundreds of students each semester. In each of these 12 worlds, we vary whether the graders come from a distribution with low variation in grader reliability and strictness, or one with relatively high variation in the same.[18] Ultimately, then, we construct 24 different simulated environments (4 possible numbers of graders X 3 different numbers of students per grader X 2 different grader distributions).

From previous results, we expect improvement in grading bias to be conditioned strongly by the variability of graders - that as this variability increases, bias is likely to increase in the unbridged scenario, but be relatively well accounted for when we apply our bridging

[18] We construct these possible combinations from the same distributions as in the simulation above. Thus, graders in a "Low Variability" world have stretch parameters β drawn from a distribution of $\beta_i \sim \mathcal{U}[0.8, 1]$, and shift parameter α from the distribution $\alpha_i \sim \mathcal{N}(0, 3)$. Graders in a "High Variability" world have stretch parameters β drawn from a distribution of $\beta_i \sim \mathcal{U}[0.5, 1]$, and shift parameter α from the distribution $\alpha_i \sim \mathcal{N}(0, 7)$.

solution. This subsection is mainly focused on interpreting what happens at different levels of graders and students.

Each simulated environment is reconstructed 100 times, with graders and their attributes redrawn from the appropriate distribution and a new set of the appropriate number of "true" latent grades drawn from a distribution $Grade_i \sim \mathcal{N}(50, 10)$. Students with these latent skills are randomly assigned equally to graders, and the grades given by each grader to all students are calculated using the attributes as drawn from the distribution. In this way, we have grades for all students from all graders, but each student is assigned to one primary grader.

The success of the algorithm is judged by comparing the mean average error and root mean square error of the "single TA" form of grading to that of the bridged scenario (again, using 3 bridges). The baseline against which we judge both is the average grade of the student from all possible graders.

In Figure 14, we display the results of this exercise. The graph is separated into two rows, where each row corresponds to one of the states of the world. In the upper row, graders are selected from distributions with low variability for both the α and β parameters - these graders are very similar to each other, and we would expect less grading bias of the type we are attempting to reduce. In the lower row, there is higher variability, and we expect that a traditional regime with no bridging could experience quite a bit of bias. As we move across rows, we are adding graders, from a course with two graders on the far left, to a course with five distinct graders on the far right. Finally, within each plot, there are three different levels for the numbers of students each grader is assigned. For each combination of world (high vs. low) and number of graders (2, 3, 4, or 5), there are results for simulations where each grader was assigned 12 students, 30 students, or 60 students. In the bottom right panel, for illustration, the top result corresponds to a course of 300 students (60 students for each of 5 graders) under conditions likely to produce much different graders. Black box plots reflect

the distribution of MAEs pulled from each simulated combination when 3 bridges are used; red box plots reflect the distribution of MAEs when we do not utilize bridges.

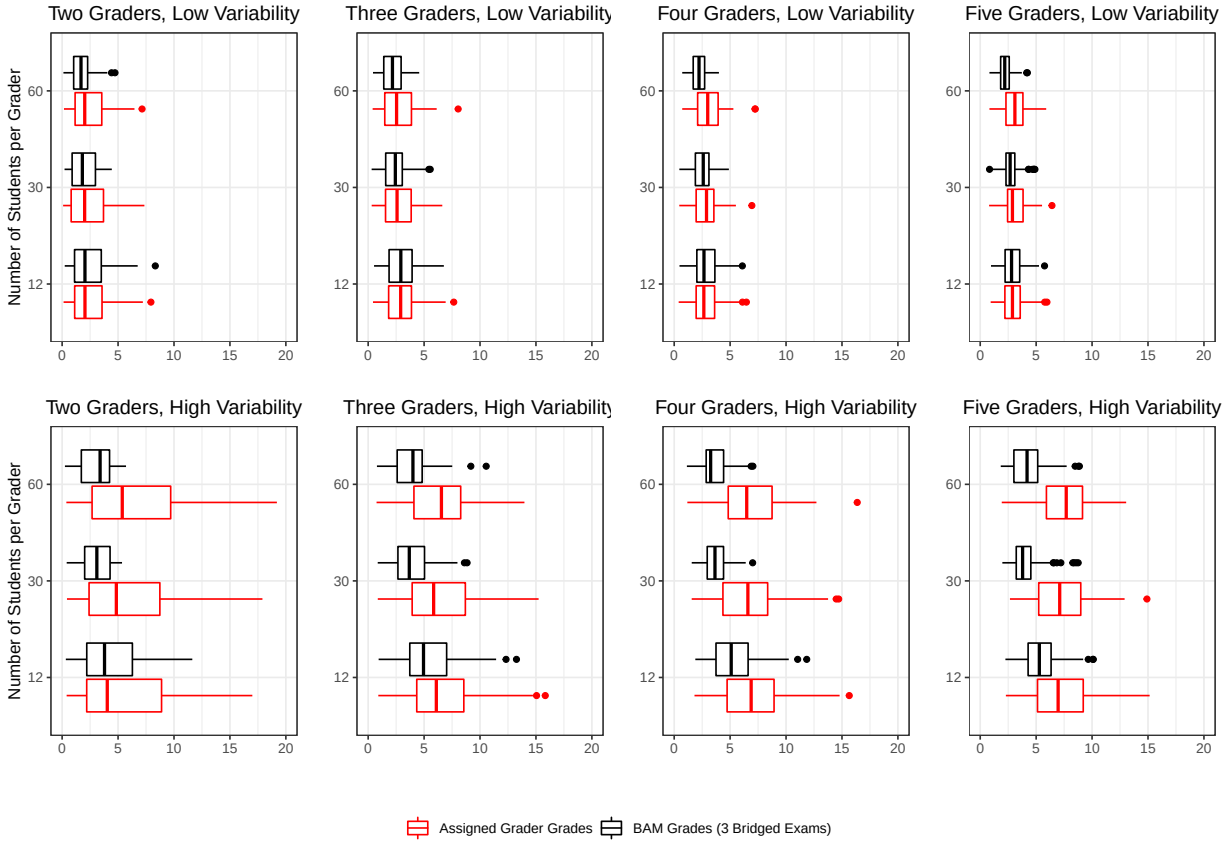


Figure 14: MAE estimates of ranked placement across simulated data sets. Black whiskers represent estimates from the BAM model with three bridges. Red whiskers represent estimates from the typical (assigned grader) assessment method.

These 24 simulations allow us to say something about the conditions under which our proposed solution is likely to be more or less likely to reduce bias. When we compare across rows, we note first that there is less baseline in the low variability world. This is as we expect. The type of bias we are attempting to address stems from this variability - from graders that have different baselines, and different functions mapping increases in perceived skill to additional reward. As we move to the right, holding variability and the number of students per grader constant, there is a slight but distinguishable increase in baseline error,

and a larger, but still relatively small increase in bias as we increase the number of students per grader, but leave the other two variables constant.

What does vary considerably across simulations is how successful our solution is at reducing that bias. In the low variability world (top row), we see some improvement in nearly every combination of grader number and number of students per grader. But this improvement is small, with reductions greater than 10-15% realized only when there are many students for each grader. In the high variability world, the story is much different. Our approach yields large reductions in bias that increase in both the number of graders and the number of students assigned to each grader. In many of the classes with 30 or more students per grader, or 3 or more graders, we reduce the bias by nearly 50%.

It is difficult to know in which setting (high vs. low variability) a specific real-life course takes place. The types of questions and expected responses on a particular assessment can affect this to a great degree, as can the experience and similarities of the graders. However, these simulations suggest it is largely true that regardless of setting, there is at least some benefit to bridging, and very large benefits in many instances.

Comparison with Alternative Models

We can also compare the performance of the BAM model with alternative approaches to modelling differential item functioning. Following Marquardt and Pemstein (2018), we focus on ordinal IRT models incorporating DIF via grader-specific ordinal thresholds for mapping latent ability into scores.

More precisely, let $\tilde{Grade}_{ij}$ denote the grader i's perception of the true grade (γ_j) and let e_{ij} denote the error of the grader's perception: $\tilde{Grade}_{ij} = \gamma_j + e_{ij}$. Assuming grader error follows a common distribution with variance σ, the cumulative distribution function of the error term is $F(\frac{e_{ij}}{\sigma})$. The grader can assign any $k \in \{1, .., K\}$ ordinal grades. Then, the probability that the grader assigns some grade $Grade_{ij}$=k given thresholds γ_k is:

$$\begin{aligned} Prob(Grade_{ij} = k) &= Prob(\tilde{Grade}_{ij} > \gamma_{k-1} \wedge \tilde{Grade}_{ij} \leq \gamma_k) \\ &= F(\frac{\gamma_k - \gamma_j}{\sigma}) - F(\frac{\gamma_{k-1} - \gamma_j}{\sigma}) \\ &= F(\kappa_k - \gamma_j\tau) - F(\kappa_{k-1} - \gamma_j\tau), \end{aligned}$$

where $\tau = \frac{1}{\sigma}$ is the grader's precision and $\kappa_k = \gamma_k\tau$ are estimated thresholds.

In the simulation procedure to follow, we examine a "threshold-DIF" IRT model that allows for variation in grader-specific thresholds ($\kappa_{i,k}$) and grader-specific precision (τ_i).

$$\begin{aligned} Prob(Grade_{ij} = k) &= \phi(\kappa_{i,k} - \gamma_j\tau_i) - \phi(\kappa_{i,k-1} - \gamma_j\tau_i) \\ \kappa_{i,k} &\sim \mathcal{N}(\kappa_k, 3) \\ \kappa_k &\sim \mathcal{N}(0, 10) \\ \tau_i &\sim \mathcal{N}(1, 1) \\ \gamma_j &\sim \mathcal{N}(0, 1) \end{aligned}$$

Grader thresholds are hierarchically clustered about global thresholds (κ_k) with a standard

deviation of 3. The global thresholds are normally distributed about zero with a standard deviation of 10. Grader precision is normally distributed around one with a standard deviation of one and restricted to positive values. Finally, latent ability follows a standard normal distribution.

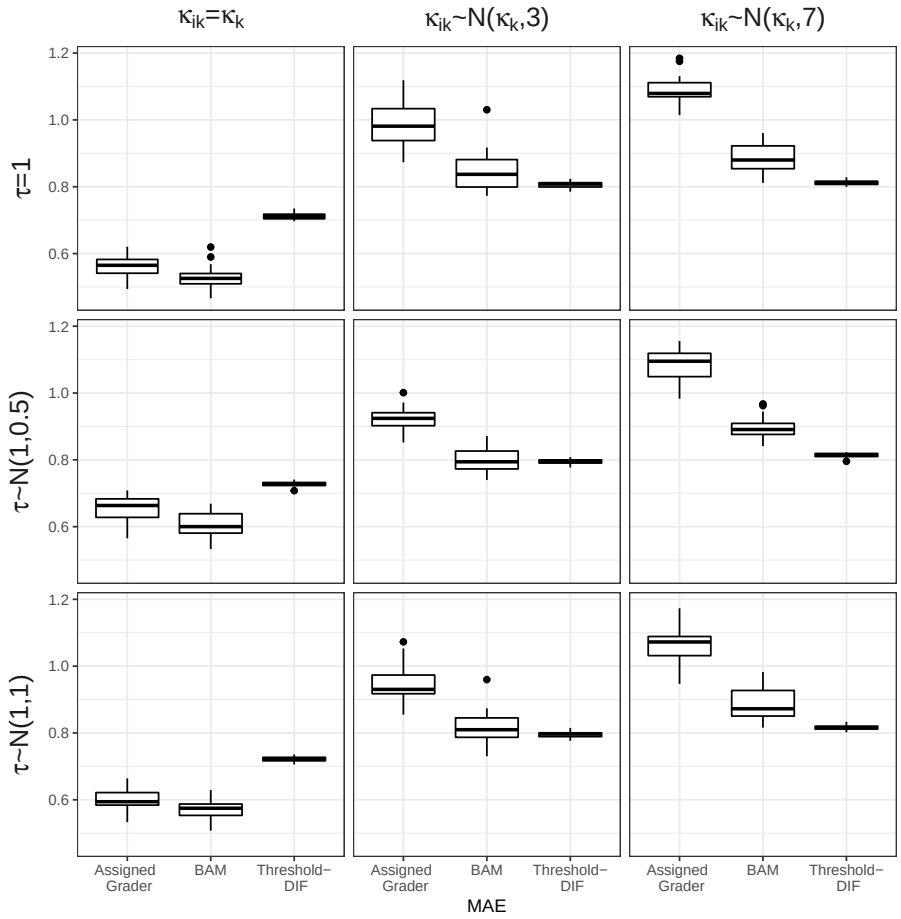
Simulation Procedure: We use the three-grader average of the midterm as the baseline (γ_j). Next, we apply the normal distribution's quantile function to the average midterm grade to estimate the (global) thresholds, κ_k. We then generate nine simulation datasets corresponding to all the combinations of three forms of grader precision and three forms of grader DIF:

- Variability in precision across three graders:
    1. No variability: $\tau = \tau_i = 1$;
    2. Medium variability: $\tau_i \sim \mathcal{N}(1, .5)$;
    3. High variability: $\tau_i \sim \mathcal{N}(1, 1)$.
- Variability in DIF across three graders:
    1. No variability: all graders use the same thresholds, κ_k;
    2. Medium variability: $\kappa_{i,k} \sim \mathcal{N}(\kappa_k, 3)$;
    3. High variability: $\kappa_{i,k} \sim \mathcal{N}(\kappa_k, 7)$.

Simulation Results: Across 20 iterations of this simulation procedure, we randomly select five exams to treat as bridging observations with the remaining exams assigned to a single grader. Figure 15 illustrates the mean absolute error of the simulated midterm score across the three levels of grader precision (rows) and three forms of DIF (columns). We find that both of the two bridging approaches improve upon the traditional method in the presence of moderate or severe DIF. In these circumstances, the IRT model produces both smaller and

less variable error than the the Bayesian Aldrich Mckelvey model. In the extreme scenario in which graders agree on the underlying thresholds, the Bayesian Aldrich Mckelvey model performs slightly better than the IRT model.

Figure 15: MAE of Simulated Midterm Grades, assuming IRT Data Generating Process



MAE estimates of midterm score across simulated datasets. The rows capture the level of variability in grader precision, from zero to high variability. The columns vary based upon the level of grading bias, from no DIF to high DIF.

Appendix D: Description of Communication Package

We have identified explaining the method to students as one of the big challenges of our approach. Therefore, we have created a communication package that is intended to help instructors explain the bridging process. The package consists of two elements:

- A set of slides
- A visualization tool

The primary intended audience of these elements are the students in classes that intend to use bridging to reduce bias in grading. The set of slides explain the potential problem with multiple graders and how the bridging process can help mediate it from a student's perspective. We have kept the slides simple so they can function as a baseline explanation for a variety of classes and instructors. We encourage instructors that intend to use our method to customize the slides to suit their needs.

The visualization tool consists of an R script and the files necessary to execute it. The tool produces a document that visualizes how the bridging method can reduce bias. The visualization is meant to give students another way of understanding the method and its benefits.