Supplementary Material for:

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

Authors:

Sean Kates, University of Pennsylvania*
Tine Paulsen, New York University
Sidak Yntiso, New York University
Joshua A. Tucker, New York University

*Corresponding Author: sk5350@nyu.edu

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 may even be 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 are limited to shifting 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. [17] 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

[17] 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 Grader 3 handles each of the final papers.

Table A1. 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)

instructor, placing intermediaries at the forefront of one of the core aspects of instruction.

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 A1 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 Bayesian Aldrich-McKelvey (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.

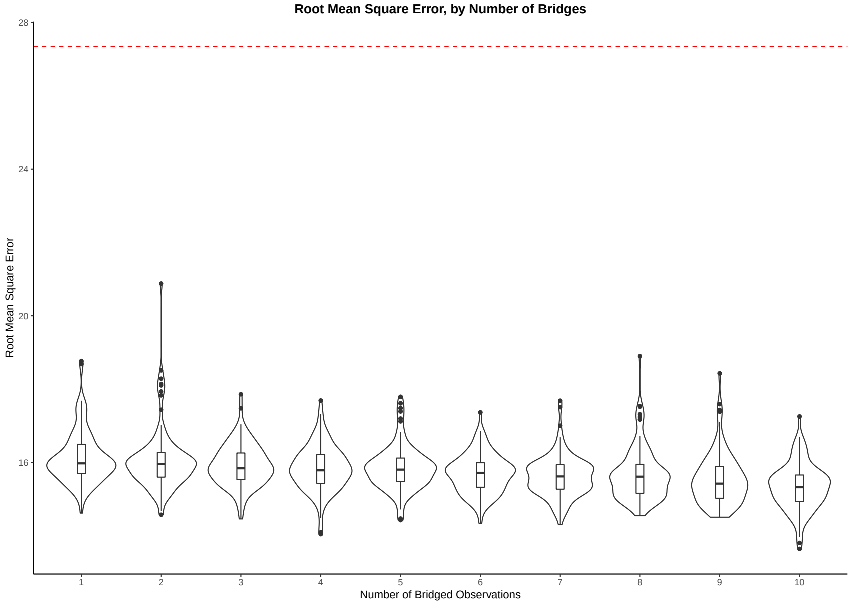


Figure A1. RMSE for estimates of student placement (rank) on midterm exam, across number of bridged exams. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

Figure A1 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 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 A1, 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.[18]

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.

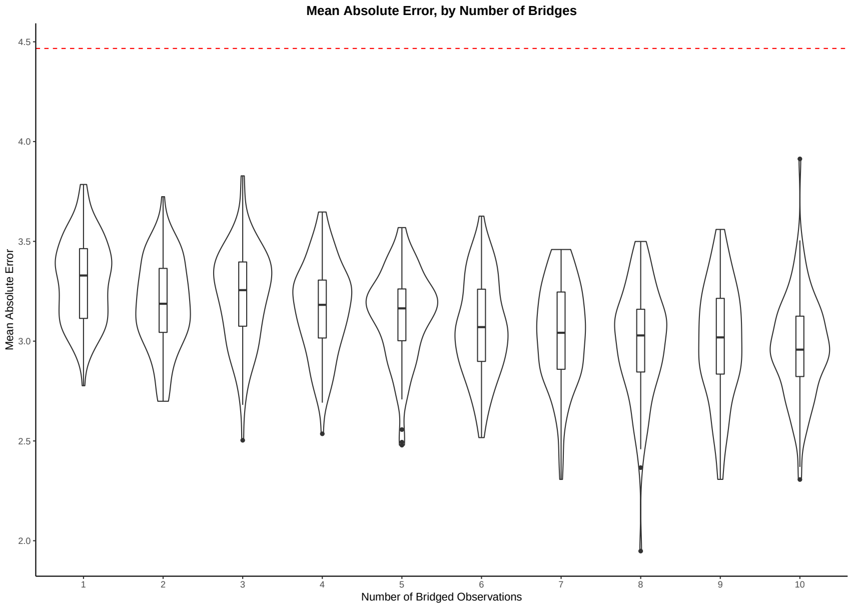


Figure A2. MAE for estimates of student score (out of 45) on midterm exam, across number of bridged exams. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

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 A2, 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.

[18] 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.

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.[19] 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.[20] 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 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 A3.

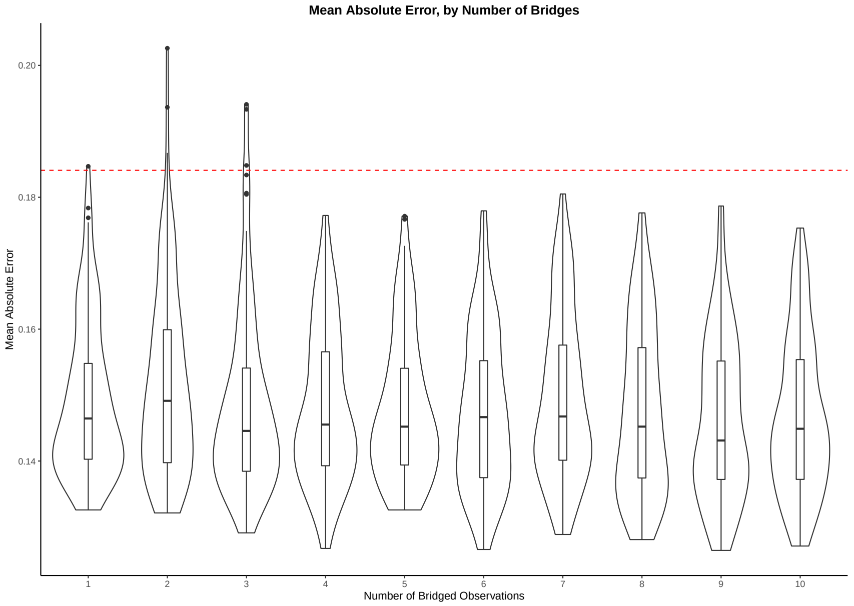


Figure A3. MAE estimates of paper score (out of 4.0), across number of bridged exams. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

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

[19] 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.

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

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.[21] In this final subsection, we show the 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 A4 displays the results. Here, the bridging method eventually reduces bias, but the number of bridges required to do so is somewhat larger and the amount of bias reduction is small relative to earlier gains. This occurs for a simple reason. As we show below, the three graders in this class were much closer to each other in terms of assessment strictness for the final exam, which in turn reduces the amount of bias there is to correct via the algorithm. As the final exam accounted for a disproportionate share of the overall grade, this reduction in possible bias correction bled through to the final grade. Still, there is some value added to bias correction in this case, and it may be worthwhile to pursue in classes when you have potential for a greater number of bridges.

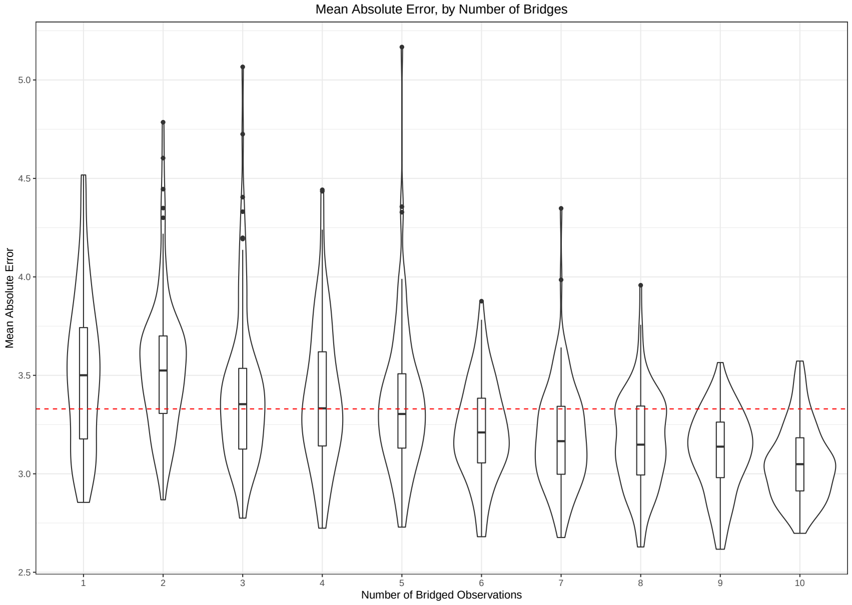


Figure A4. MAE estimates of student course grade (out of 100), across number of bridged evaluations. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

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 A5, we show that using the bridging method on a final letter grade reduces bias by approximately 25% in a limited number of bridges. This is equivalent to 12.5 students (in a class of 135) receiving a proper letter grade

[21] 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.

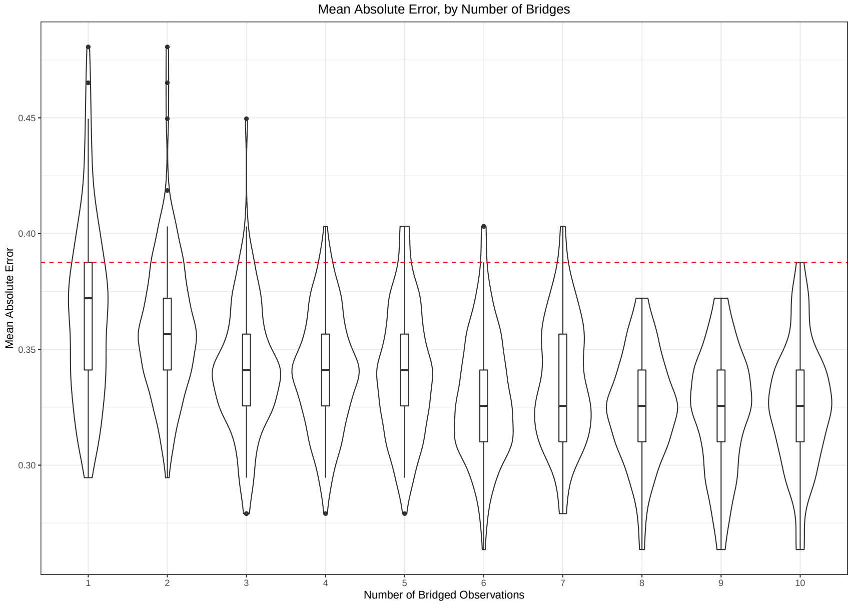


Figure A5. MAE estimates of student course letter (out of 4.0), across number of bridged evaluations. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

a full step above/below their incorrect grade (i.e. an "A" when a "B" was given) or 37.5 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 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 (4̃.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 A6 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,

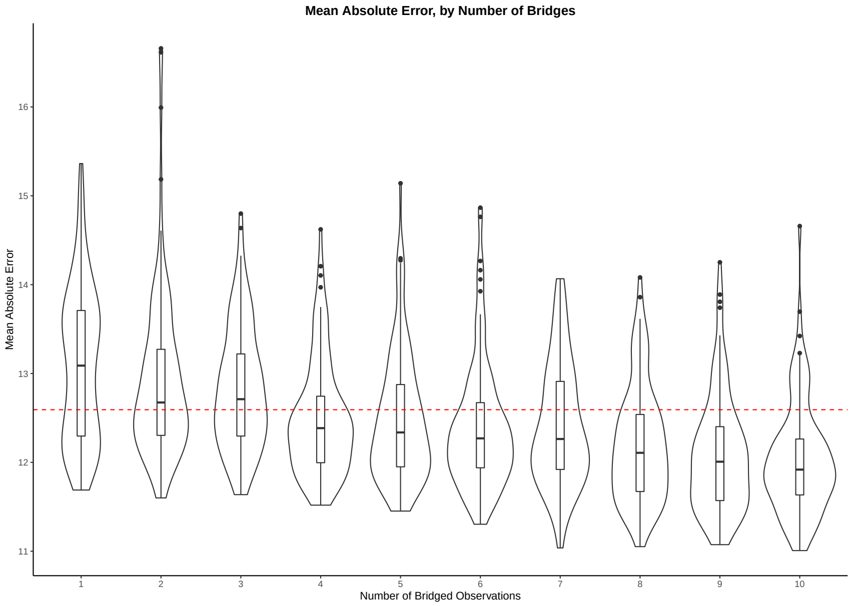


Figure A6. MAE estimates of student rank on final exam (out of 135), across number of bridged evaluations. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

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, with 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 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 A7 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

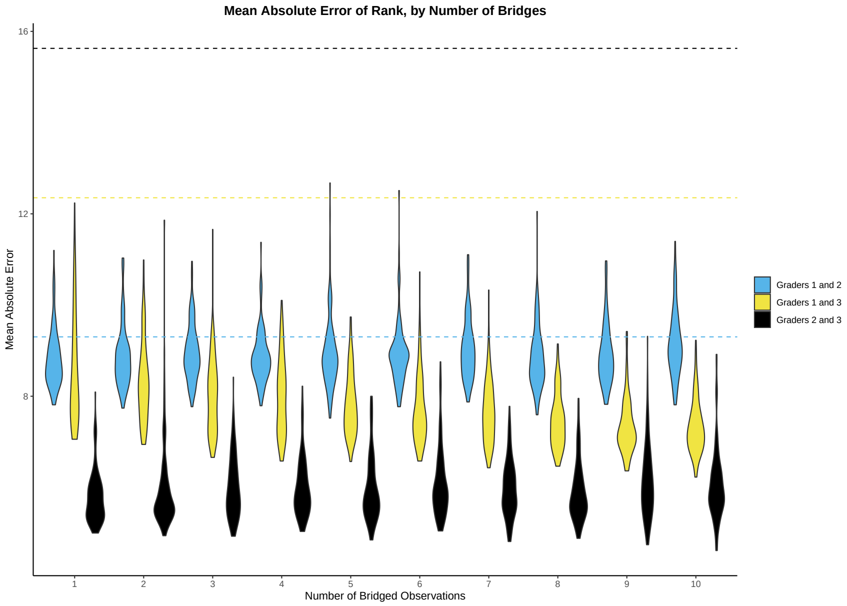


Figure A7. MAE estimates of student rank on midterm exam (out of 135), across number of bridged evaluations. Each violin plot represents a "class" combining two of the three graders for the real-life course. Each horizontal dotted line reflects the bias associated with the traditional method of grading exams in that "class."

(Grader 3 in the groups imagined in Figure A7) 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.[22]

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 A8 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

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

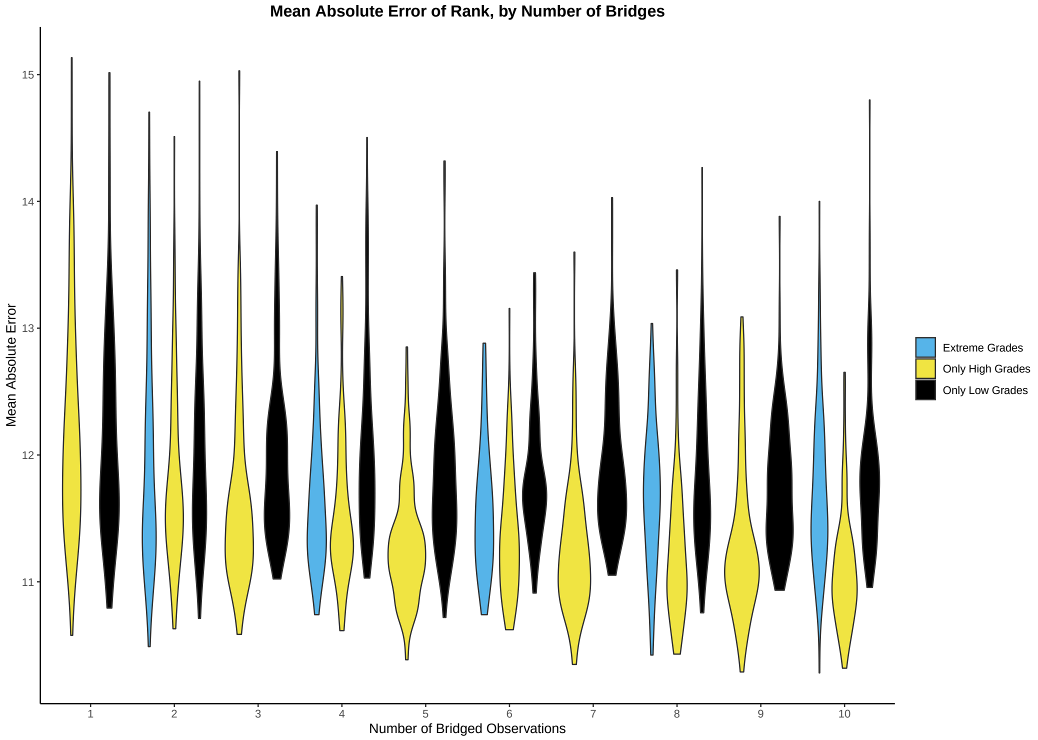


Figure A8. MAE estimates of student rank on midterm exam (out of 135), across number of bridged evaluations. Each violin plot represents a regime where the bridged observations are either all from the lowest third of the distribution, all from the highest third of the distribution, or split evenly between the highest and lowest thirds. The MAE for an unbridged process is approximately 22.5.

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 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

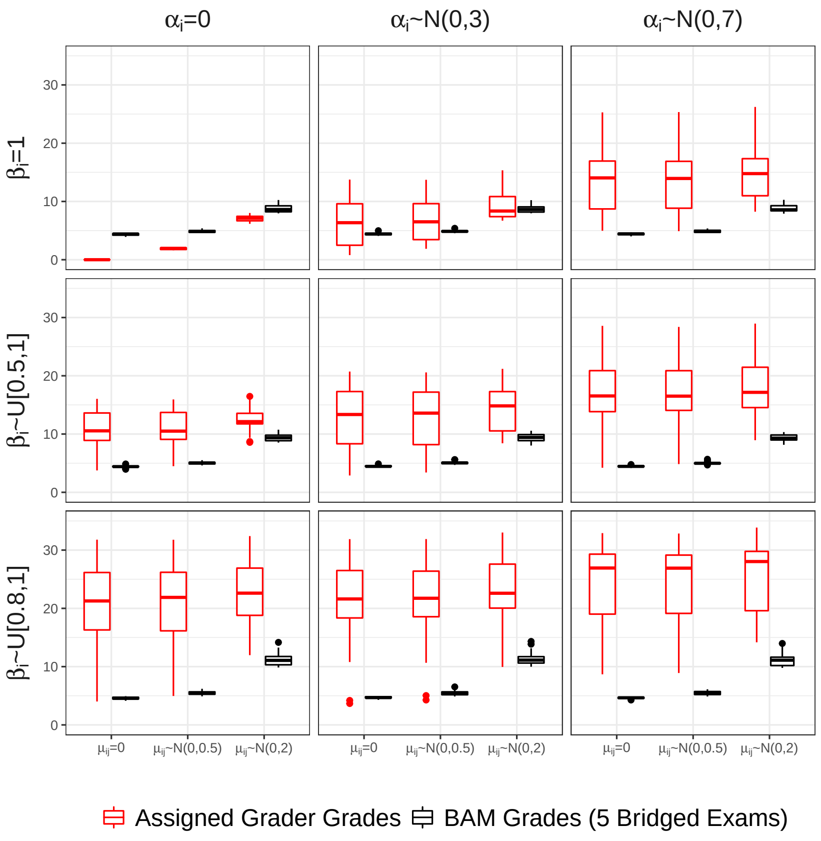
Varying Parameters

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 A9 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.

Figure A9. MAE Simulations: Final Grade

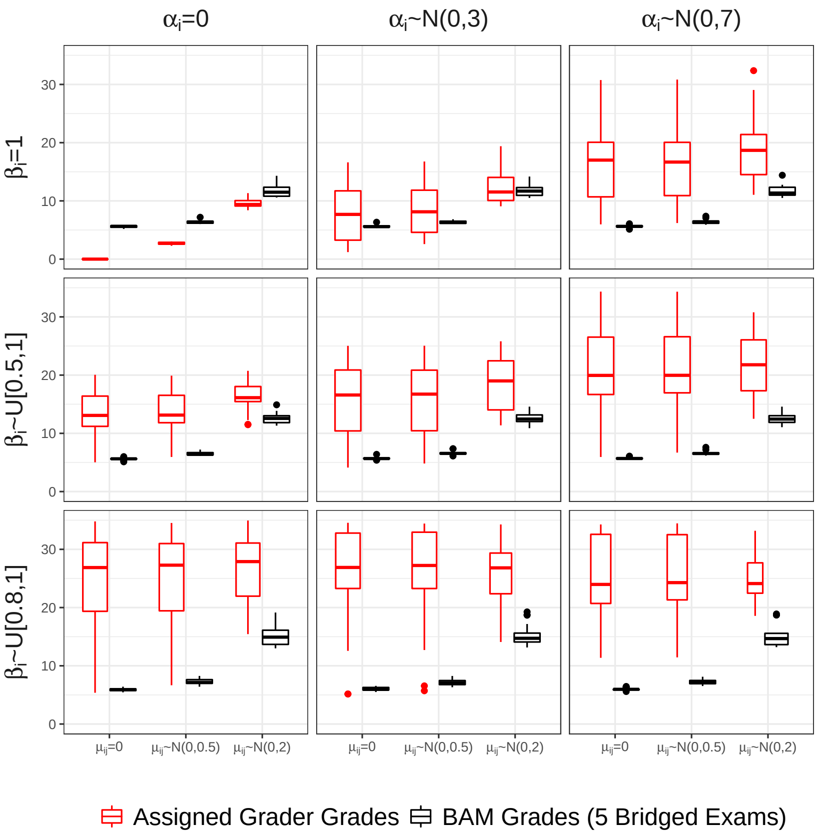


MAE estimates of final exam placement across simulated datasets. Black whiskers represent estimates from the BAM model with five bridges. Red whiskers represent estimates from the typical (assigned grader) assessment method.

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 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 A10 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 A10. 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.[23] 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 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 A11, 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

[23] 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 $\alpha_i \sim \mathcal{N}(0, 7)$.

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.

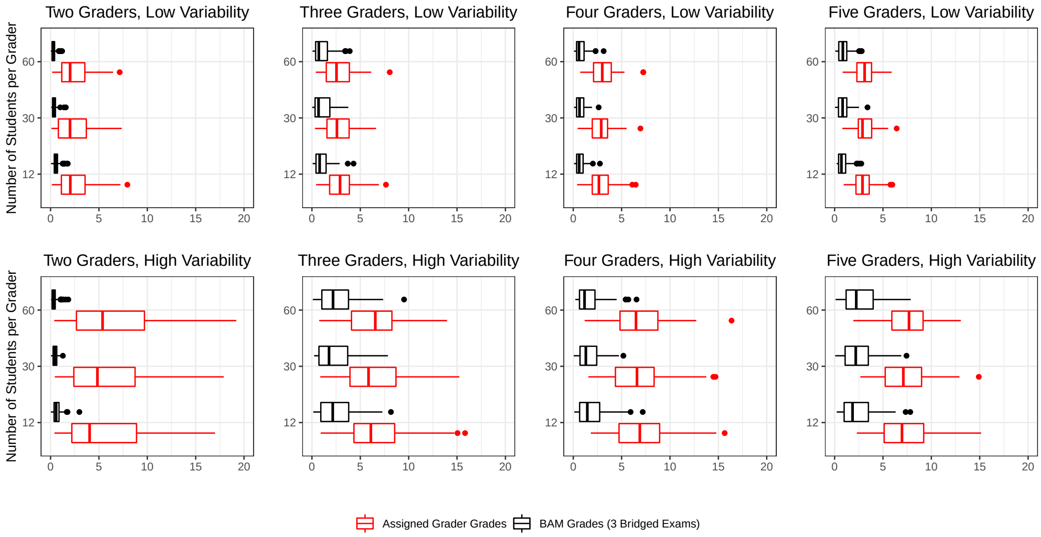


Figure A11. MAE estimates of ranked placement across simulated data sets. Black whiskers represent estimates from the BAM model with three bridges. Red whiskers represent estimates from the typical (assigned grader) assessment method.

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.

Table A2. Comparison of Approaches

Approach	Scale	Grader-specific variation
BAM	Linear	Intercepts and reliability
Intercept DIF	Ordinal	Intercepts and precision
Threshold DIF	Ordinal	Thresholds and precision

More precisely, let $Gr\tilde{a}de_{ij}$ denote the grader *i*'s perception of the true grade (γ_j) and let e_{ij} denote the error of the grader's perception: $Gr\tilde{a}de_{ij} = \gamma_j + e_{ij}$. 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:

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

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

Intercept DIF: In the first IRT model, we assume grader-specific intercepts (κ_i) and grader-specific precision (β_i). This model is similar to BAM but assumes an ordinal scale:

$$\begin{aligned} Prob(Grade_{ij} = k) &= \phi(\tau_k - \kappa_i - \gamma_j \beta_i) - \phi(\tau_{k-1} - \kappa_i - \gamma_j \beta_i) \\ \beta_i &\sim \mathcal{N}(1, 1) \\ \kappa_i &\sim \mathcal{N}(0, .5) \end{aligned}$$

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).

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

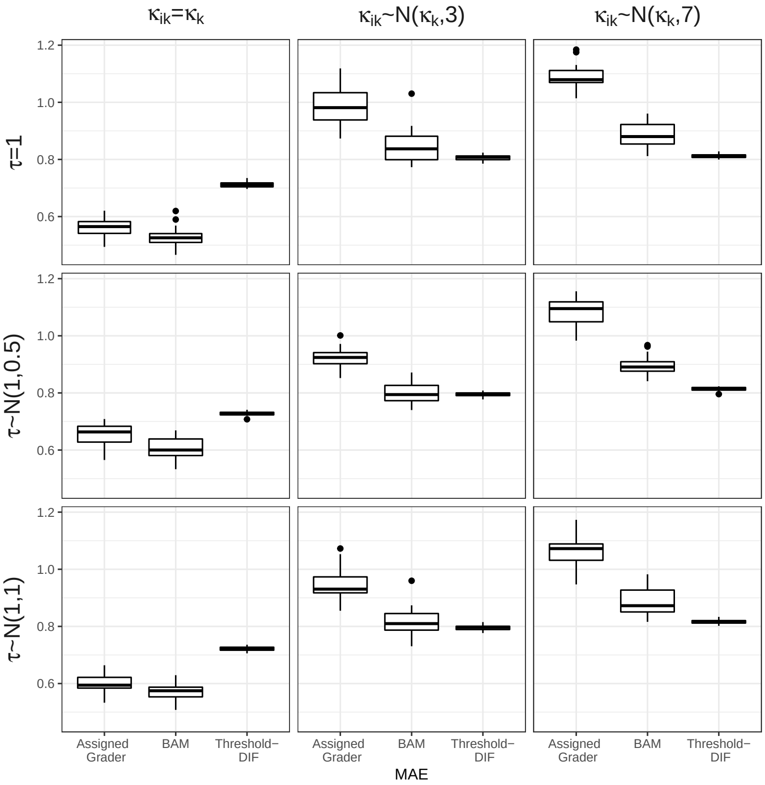
Grader thresholds are hierarchically clustered about global thresholds (κ_k) with a standard deviation of 3. The global thresholds are normally distributed about zero with a standard deviation of 10. Grader precision is normally distributed around one with a standard deviation of one and restricted to 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 A12 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 A12. 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 is included in our R package.[24] The package provides tool to visualize how the bridging method can reduce bias. The visualization is meant to give students another way of understanding the method and its benefits.

[24] https://github.com/sidakyntiso/bridgr.git