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

Kates, Sean ; Paulsen, Tine ; Yntiso, Sidak ; Tucker, Joshua A

DOI: https://doi.org/10.1017/pan.2022.27

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

Sean Kates[1], **Tine Paulsen**[2], **Sidak Yntiso**[3] and **Joshua A. Tucker**[3]

[1] *Program in Data Analytics, University of Pennsylvania, Philadelphia, PA, USA. E-mail: sk5350@nyu.edu*
[2] *Department of Political Science and International Relations, University of Southern California, Los Angeles, LA, NY*
[3] *Wilf Family Department of Politics, New York University, New York, NY, USA*

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 by a small number of "bridging" observations across graders. We conclude by offering best practices for fair assessment in large survey courses.

Keywords: Bayesian Aldrich–McKelvey scaling, differential item functioning, assessment bias

1 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 Megalokonomou 2019; Lavy and Sand 2018; Papageorge, Gershenson, and Kang 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 more than 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; Marquardt and Pemstein 2018; Pemstein, Tzelgov, and Wang 2015)—given the prevalence of grading bias—we believe that 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.[1]

Researchers seeking to advance this line of inquiry might further investigate the comparative benefits of more advanced models. After all, using advanced item response theory (IRT) models to

Political Analysis (2023)
vol. 31: 642–650
DOI: 10.1017/pan.2022.27

Published
14 December 2022

Corresponding author
Sean Kates

Edited by
Jeff Gill

1 While awaiting approval from CRAN, the R package "bridgR" can be located here: https://github.com/sidakyntiso/bridgr.git.

analyze and reduce different types of bias in assessments is a thriving field on its own (Johnson 1996; Johnson and Albert 1999; Shin, Rabe-Hesketh, and Wilson 2019; Wang, Su, and Qiu 2014).[2] However, in this paper, we attempt to balance not only reduction in bias with the cost of additional grading, but also simplicity of explication. More specifically, the method used to diminish bias has 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.

2 The Problem: Having Multiple Graders and Achieving Fair Assessment

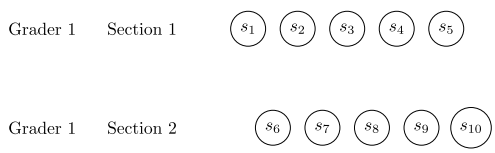
We consider grading bias stemming from having multiple graders as a violation of a specific form of fairness, fairness-through-symmetry (Blackburn 2003). In general, a process is considered fair when it can be expected to produce symmetrical outcomes for identical inputs. In our case, a grading process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

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 and 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.

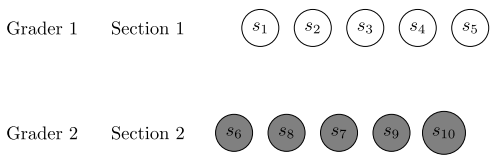
3 Proposed Solution: "Bridging" between Graders

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.[3]

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



(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.

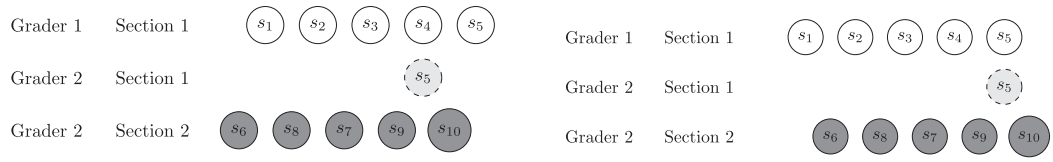


(b) Grader bias: Same students with same output as in (a), but now split into two sections, each with its own Grader. Grader 1's judgment is the same as in Figure 1 (a), but Grader 2 is more severe than Grader 1. The ranking of students differs from the one in Figure 1 (a) and is biased against those 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).

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

2 See also the literature on using advanced Rasch models for doing this, for example, Wind, Engelhard, and Wesolowski (2016) and Wind and Jones (2018).

3 We cover a host of possible alternatives to our method—and why they are found lacking—in Appendix A in the Supplementary Material.



(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.

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

(or "bridged") student to observe that one grader is stricter than another. Thus, fairness demands that we adjust the students in both sections to reflect relative performance accurately.

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).[4] Bridging has also been used to ensure that evaluators with different scales for "democracy" can reliably place countries on a common scale (King *et al.* 2004; Marquardt and Pemstein 2018; Pemstein, Tzelgov, and Wang 2015). Americanists use bridging to place politicians operating in different environments in the same ideological space (Poole 2007). This familiarity 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 the following:

- Each grader will be internally consistent in their assessments.
- Graders have to be somewhat consistent, that is, have the same sense of what differentiates a high-quality answer and a low-quality answer.
- Graders reward better quality answers 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.

4 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, Rothenberg, and Rulison (2013) and Lo, Proksch,

4 See also Struthers, Hare, and Bakker (2019).

and Gschwend (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.[5]

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, whereas 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 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 $\beta_i \sim \mathcal{N}(0, 30)$) 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.[6] 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,[7] 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.[8] We collected grades during a Fall 2018 semester *Introduction to Comparative Politics* course

5 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.

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

7 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 in the Supplementary Material).

8 All data and scripts are posted in our replication data, available in Kates *et al.* (2022) at https://doi.org/10.7910/DVN/16EYI2. The accompanying R package is available at https://github.com/sidakyntiso/bridgr.git.

at a large private research university located in the Northeast,[9] 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-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[10] 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, where the actual rank is the rank from the student's single assigned grader and the "earned" ranking is the ranking based on the average grade of all three graders.[11] 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, that is, 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 total grade and rank is determined by the average of these three scores.[12] We calculate the MAE and the 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 is 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.

5 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 leads quickly to a multiplication of work and a continuously larger cost in terms of instructor time chasing more bias reduction.

9 Student assessments were de-identified. We received IRB approval (New York University IRB-FY2019-2483) to use the de-identified grades for research purposes.

10 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.

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

12 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 in the Supplementary Material. The exercise shows that our approach vastly reduces bias even in cases where the baseline is the "true" grade.

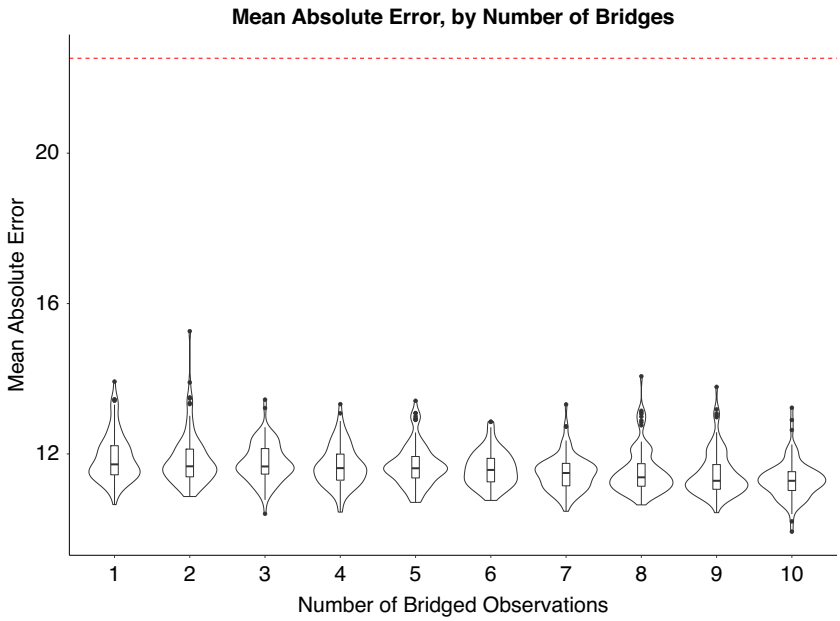


Figure 3. Mean absolute error for estimates of student placement (rank) on midterm exam, across a number of bridged exams. The horizontal dotted line reflects the bias associated with the traditional method of grading exams in our data.

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, whereas 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 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 to 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 give 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 more than 50%.

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

5.1 Additional Real-Life Results

In Appendix B in the Supplementary Material, 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 commonly graded assignments to determine the potential gains from the algorithm ex ante. This allows for a "mixed-methods" approach to decreasing bias that utilizes both the algorithm and pedagogical refocusing where necessary. Finally, we find no evidence that the performance depends on the specific students used as bridges (i.e., using only low, high, or extreme scores).

5.2 Simulation Results

In Appendix C in the Supplementary Material, 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 datasets reflecting various degrees of grader reliability (β) and grader shift (α), as well as grade-level error (μ).[13] The Bayesian Aldrich–McKelvey model outperforms the traditional evaluation method across all datasets 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 vs. 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, in the number of students, and in the 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 datasets from an IRT data-generating process that incorporates grading bias via grader-specific ordinal thresholds for mapping latent ability into scores. We find that both bridging 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, although this difference is negligible compared to the difference with the traditional method.

6 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 that this is the most economical first step in reducing potential bias.

13 We explore cases with no, low, or high variability for each parameter.

We should note three 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, this algorithm addresses a specific type of error stemming from attributes of the assessors.[14] However, we should not expect it to perform (or claim that it performs) equally well across all classrooms.[15] Finally, 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.[16]

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 that 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.

Acknowledgments

We thank Shawna Metzger and the participants at the APSA 2019 Education and Replication in Political Methodology Panel for thoughtful feedback. S.K. had the original idea for the paper and supervised the data collection, data analysis, and writing of the paper. S.K. and J.T. produced the original research design. S.K., T.P., and S.Y. conducted the analysis and wrote the first draft of the paper. J.T. facilitated the incorporation of the grading plan into his "UA 500: Introduction to Comparative Politics" lecture course in the fall of 2019. S.K., T.P., and S.Y. 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 of the New York University IRB Board.

Data Availability Statement

Replication code for this article is available in Kates *et al.* (2022) at https://doi.org/10.7910/DVN/BIORH8.

14 It is also worth explicitly noting that the method we propose here does not address the *quality* of the instruction received by the student as a function of the instructor. We might want to think, therefore, of a *maximal* definition of fairness: the grade is independent of all aspects of instructor assignment. From that vantage point, it would make sense to describe our method as addressing a more *minimal* definition of fairness: that *conditional on the answer* provided by the student to an exam question, the grade received should be independent of the assignment to a particular grader. It is for this reason that the method proposed is described as a means of reducing *assessor* or *grader* bias, as opposed to the more maximal consideration of all forms of instructor assignment bias.

15 In practice, instructors can diagnose the usefulness of the bridging algorithm after collecting grading data with common, bridged assignments. Randomly selected bridging exams permit the instructor to generate estimates of the MAE (and RMSE) or conduct hypothesis tests before applying the algorithm (the R package includes this procedure).

16 These materials are available in the replication files linked above and below.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.27.

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*, vol. 80. New York: 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: Springer.

Kates, S., T. Paulsen, S. Yntiso, and J. A. Tucker. 2022. *Replication Data for: Bridging the Grade Gap: Reducing Assessment Bias in a Multi-Grader Class*. https://doi.org/10.7910/DVN/BIORH8.

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*. New York: Oxford University Press.

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* 8: 677–691.

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.