Reliability of Instructor Evaluations

Erin M. Buchanan[1], Jacob Miranda[2], & Christian Stephens[2]

[1] Harrisburg University of Science and Technology

[2] University of Alabama

Author Note

TBA

The authors made the following contributions. Erin M. Buchanan: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Analysis; Jacob Miranda: Writing - Original Draft Preparation; Christian Stephens: Writing - Original Draft Preparation.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA 17010. E-mail: ebuchanan@harrisburgu.edu

Abstract

TBA

Keywords: keywords

Word count: X

Reliability of Instructor Evaluations

The following was pre-registered: https://osf.io/czb4f

Exploratory Research Questions:

1) What is the reliability of instructor evaluations?
2) Are instructor evaluations reliable across time?
3) Is the average level of perceived fairness of the grading in the course a moderator of reliability in instructor evaluations?
4) Does the average variability in instructor fairness rating moderate reliability of instructor evaluations?

Method

Data Source

The archival study was conducted using data from the psychology department at a large Midwestern public university. We used data from 2898 undergraduate, 274 mixed-level undergraduate, and 42 graduate psychology classes taught from 1987 to 2018 that were evaluated by students using the same 15-item instrument. Faculty followed set procedures in distributing scan forms no more than two weeks before the conclusion of the semester. A student was assigned to collect the forms and deliver them to the departmental secretary. The instructor was required to leave the room while students completed the forms. In the last several years of evaluations, online versions of these forms were used with faculty encouraged to give students time to complete them in class while they were outside the classroom.

The questionnaire given to students can be found at https://osf.io/4sphx. These items were presented with a five-point scale from 1 (*strongly disagree*) to 5 (*strongly agree*). For this study, the overall instructor evaluation question was “The overall quality of this course was among the top 20% of those I have taken.". For fairness, we used the question of "The instructor used fair and appropriate methods in the determination of grades.". The ratings were averaged for each course, and the sample size for each rating was included.

Planned Analyses

The evaluations were filtered for those with at least fifteen student ratings for the course (Rantanen, 2012). We performed a robustness check for the first research question on the data when the sample size is at least $n = 10$ up to $n = 14$ (i.e., on all evaluations with at least 10 ratings, then at least 11 ratings, etc.) to determine if the reliability estimates are stable at lower sample sizes. We first screened the dataset (two evaluation questions, sample size for course) for accuracy errors, linearity, normality, and homoscedasticity. The data is assumed to not have traditional "outliers", as these evaluations represent true averages from student evaluations. If the linearity assumption fails, we considered potential nonparametric models to address non-linearity. Deviations from normality were noted as the large sample size should provide robustness for any violations of normality. If data appears to be heteroscedastic, we used bootstrapping to provide estimates and confidence intervals.

This data was considered structured by instructor; therefore, all analyses below were coded in *R* using the *nlme* package (Pinheiro, Bates, Debroy, Sarkar, & Team, 2017) to control for correlated error of instructor as a random intercept in a multilevel model. Multilevel models allow for analysis of repeated measures data without collapsing by participant [i.e., each instructor/semester/course combination can be kept separate without averaging over these measurements; Gelman (2006)]. Random intercept models are regression models on repeated data that structure the data by a specified variable, which was instructor in this analysis. Therefore, each instructor's average rating score was allowed to vary within the analysis, as ratings would be expected to be different from instructor to instructor. In each of the analyses described below, the number of students providing ratings for the course was included as a control variable to even out differences in course size as an influence in the results. However, this variable was excluded if the models did not converge. The dependent variable and predictors varied based on the research question, and these are described with each analysis below.

RQ 1. In this research question, we examined the reliability of instructor evaluations on the overall rating and separately on the fairness rating. We calculated eight types of reliability using course (same or different) by instructor (same or different) by semester (same or different). The dependent variable was the first question average with a predictor of the comparison question average, and both sample sizes (first sample size, comparison sample size). Instructor code was used as the random intercept for both ratings (i.e., two instructor random intercepts, first and comparison). The value of interest was the standardized regression coefficient for the fixed effect of question from this model. Given that the large sample size will likely produce "significant" *p*-values, we used the 95% CI to determine which reliability values were larger than zero and to compare reliability estimates to each other.

RQ 2. We used the reliability for the same instructor and course calculated as described in RQ1 at each time point difference between semesters. For example, the same semester would create a time difference of 0. The next semester (Spring to Summer, Summer to Fall, Fall to Spring) would create a time difference of 1. We used the time difference as a fixed effect to predict reliability for the overall question only with a random intercept of instructor. We used the coefficient of time difference and its confidence interval to determine if there was a linear change over time. Finally, we plotted the changes over time to examine if this effect was non-linear in nature and discuss implications of the graph.

RQ 3. Using the reliability estimates from RQ 2, we then added the average rating for the fairness question as the moderator with time to predict reliability. Fairness was calculated as the average of the fairness question for all courses involved in the reliability calculation for that instructor and time difference. Therefore, this rating represented the average perceived fairness of grading at the time of ratings. If this interaction effect's coefficient does not include zero, we performed a simple slopes analysis to examine the effects of instructors who were rated at average fairness, one standard deviation below average, and one standard deviation above average (Cohen, Cohen, West, & Aiken, 2003).

RQ 4. Finally, we examined the average standard deviation of fairness ratings as a moderator of with time to predict reliability. This variable represented the variability in perceived fairness in grading from student evaluations, where small numbers indicated relative agreement on the rating of fairness and larger values indicated a wide range of fairness ratings. The variability in fairness ratings was calculated in the same way as the mean fairness, which was only for the instructor and semester time difference evaluations that were used to calculate the reliability estimate. This research question was assessed the same way as research question three.

Results

Data Screening

The overall dataset was screened for normality, linearity, homogeneity, and homoscedasticity using procedures from Tabachnick, Fidell, and Ullman (2019). Data generally met assumptions with a slight skew and some heterogeneity. The complete anonymized dataset and other information can be found online at https://osf.io/k7zh2. This page also includes the manuscript written inline with the statistical analysis with the *papaja* package (Aust et al., 2022) for interested researchers/reviewers who wish to recreate these analyses.

Descriptive Statistics

3214 evaluations included at least 15 student evaluations for analysis. Table 1 portrays the descriptive statistics for each course level including the total number of evaluations, unique instructors, unique course numbers, and average scores for the two rating items. Students additionally projected their course grade for each class ($A = 5$, $B = 4$, $C = 3$, $D = 2$, $F = 1$), and the average for this item is included for reference. Overall, 231 unique instructors and 70 unique courses were included in the analyses below across 94 semesters.

RQ 1

Each individual evaluation was compared to every other evaluation resulting in 5163291 total comparisons. Eight combinations of ratings were examined using instructor (same, different), course (same, different), and semester (same, different) on both the overall and fairness evaluation ratings separately. One of the individual ratings was used to predict the comparison rating (i.e., question 1 was used to predict a comparison question 1 for the same instructor, different instructor, same semester, different semester, etc.), and the number of ratings (i.e., rating sample size) per question were used as fixed-effects covariates. The instructor(s) were used as a random intercept to control for correlated error and overall average rating per instructor. The effects were then standardized using the *parameters* package (Lüdecke et al., 2023). The data was sorted by year and semester such that "predictor" was always an earlier semester predicting a later semester's scores, except in cases of the the same semester comparisons. Therefore, positive standardized scores indicate that scores tend to go up over time, while negative scores indicate that scores tend to go down over time.

As shown in 1, reliability was highest when calculated on the same instructor in the same semester and within the same course for both overall rating and fairness. This reliability was followed by the same instructor, same semester, and different courses. Next, the reliability for same instructor, same course, and different semesters was greater than zero and usually overlapped in confidence interval with same instructor, same semester, and different courses. Interestingly, the same instructor with different courses and semesters showed a non-zero negative relationship, indicating that ratings generally were lower for later semesters in different courses.

For different instructors, we found positive non-zero reliablities when they were at least calculated on the same semester or course. These values were very close to zero, generally in the .01 to .05 range. Last, the reliabilities that were calculated on different courses, semesters, and instructors include zero in their confidence intervals. Exact values can be found in the online supplemental document with the robustness analysis.

ROBUSTNESS REVEALED:

RQ 2

The paired evaluations were then filtered to only examine course and instructor matches to explore the relation of reliability across time. Reliability was calculated by calculating the partial correlation between the overall rating for the course first evaluation and the overall rating for the course second evaluation, controlling for the number of ratings within those average scores. This reliability was calculated separately for each instructor and semester difference (i.e., the time between evaluations, 0 means same semester, 1 means the next semester, 2 means two semesters later, etc.). The ratings were filtered so that at least 10 pairs of ratings were present for each instructor and semester difference combination (Weaver & Koopman, 2014). Of 36084 possible matched instructor and course pairings, 30728 included at least 10 pairings, which was 1009 total instructor and semester combinations.

The confidence interval for the effect of semester difference predicting reliability did not cross zero, b = -0.004, 95% CI [-0.005, -0.003], R^2 = .04. The coefficient, while small, represents a small effect of time on the reliability of instructor ratings. As shown in 2, reliability appears to decrease across time.

RQ 3

The confidence interval for the interaction of semester time difference and average fairness did cross zero, b = -0.001, 95% CI [-0.007, 0.005], $R^2 = .04$. Therefore, there was no effect of the interaction of average fairness with semester differences in predicting reliability. Similarly, average fairness did not predict reliability overall, b = -0.041, 95% CI [-0.226, 0.143].

RQ 4

The confidence interval for the interaction of variability of fairness and semester time difference did cross zero, b = -0.010, 95% CI [-0.022, 0.002], $R^2 = .05$. The variability of fairness also did not predict reliability overall, b = 0.291, 95% CI [-0.091, 0.672].

Discussion

- Summarize the results

What Should I Do with This Information

- Don't expect to be reliable across other classes
- Don't expect to be reliable over long period of time, people change, students change, etc.

Strengths

- a crap ton of data
- over a long period of time
- robust results

Limitations

- one item versus many
- evaluations don't mean what we want them to mean

- one uni means maybe not generalizable

Future Work

References

Aust, F., Barth, M., Diedenhofen, B., Stahl, C., Casillas, J. V., & Siegel, R. (2022). *Papaja: Prepare american psychological association journal articles with r markdown.* Retrieved from https://CRAN.R-project.org/package=papaja

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. https://doi.org/10.1198/004017005000000661

Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Højsgaard, S., Wiernik, B. M., … Luchman, J. (2023). *Parameters: Processing of model parameters.* Retrieved from https://CRAN.R-project.org/package=parameters

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., & Team, R. C. (2017). *Nlme: Linear and nonlinear mixed effects models.* Retrieved from https://cran.r-project.org/package=nlme

Rantanen, P. (2012). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(2), 224–239. https://doi.org/10.1080/02602938.2011.625471

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2019). *Using multivariate statistics* (Seventh edition). NY, NY: Pearson.

Weaver, B., & Koopman, R. (2014). An SPSS macro to compute confidence intervals for pearson's correlation. *The Quantitative Methods for Psychology*, *10*(1), 29–39. https://doi.org/10.20982/tqmp.10.1.p029

Table 1

Descriptive Statistics of Included Courses

Statistic	Undergraduate	Mixed	Master's
N Total	2898	274	42
N Instructors	223	40	10
N Courses	41	21	8
Average N Ratings	34.39	21.15	21.10
Average Overall	3.94	4.01	3.72
SD Overall	0.55	0.59	0.67
Average Fairness	4.46	4.50	4.19
SD Fairness	0.35	0.38	0.55
Average Grade	4.26	4.52	4.41
SD Grade	0.33	0.27	0.34

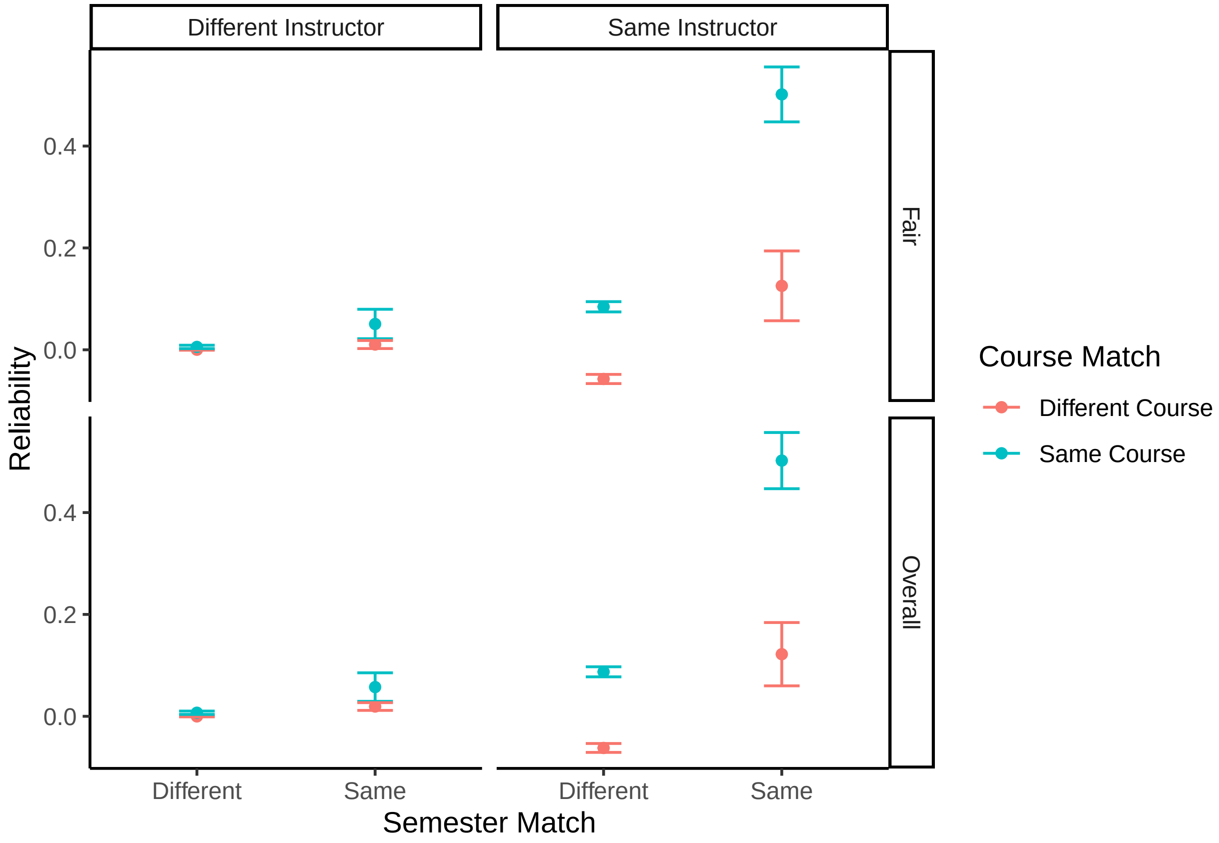

Figure 1. Reliability estimates for instructor, course, and semester combinations.

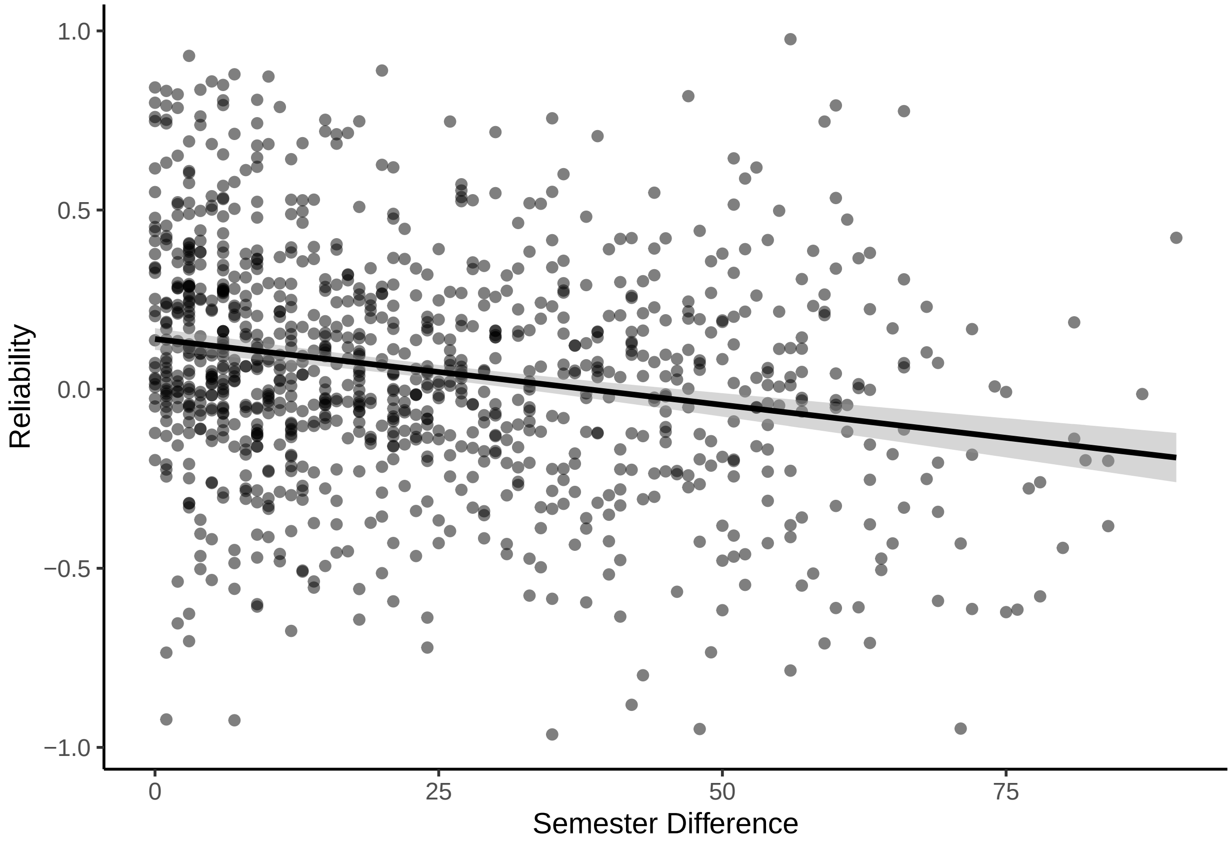

Figure 2. Reliability estimates for same instructor and course across time.