Understanding Student Heterogeneity Using Evaluation Panel Data

Frederik Hjorth, postdoc, Ph.D.

Department of Political Science, University of Copenhagen

Objectives

- develop method for creating evaluation panel data, i.e. linking individual student evaluations over time
- use implementation in big undergrad course ('Metode 2') to draw lessons for future teaching

Motivating example

Consider a hypothetical class with two (unobserved) types of students, whose evaluations of readings differ markedly across readings, illustrated in Figure 1:

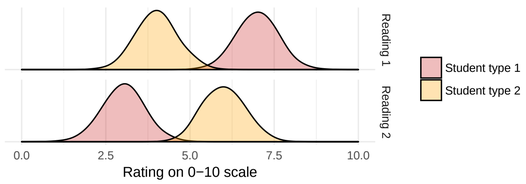

Figure 1: Ratings pattern when types' evaluations of readings vary across weeks.

We can also imagine the two types evaluating texts similarly across readings, cf. Figure 2:

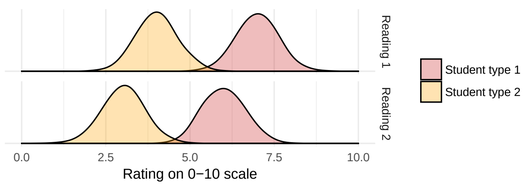

Figure 2: Ratings pattern when types' evaluations of readings are similar across weeks.

These two situations lend themselves to significantly different interpretations. In Figure 1, the evaluations suggest that different types of students prefer different texts. In Figure 2, type matters less. But in cross-sectional data, both would appear as presented in Figure 3:

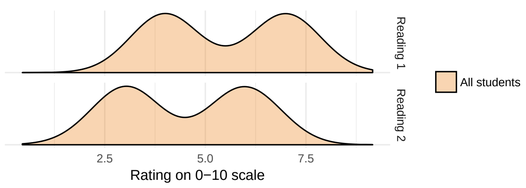

Figure 3: How both situations appear in cross-sectional format.

Without information linking individual students across weeks, the differential distribution of type-specific preferences over time is hidden to the teacher. In this poster, I present results from a pilot project **linking students' evaluations across classes**, allowing for disentangling the patterns presented above.

Theory

Student evaluation can be conceptualized as not merely a control mechanism (for study management) or an information mechanism (for the teacher), but as a constitutive component in a shared responsibility for learning between students and faculty [3]. Cook-Sather et al. conceptualize *student-faculty partnerships* characterized by respect, reciprocity, and responsibility [1].

In this project, I emphasize the formative aspect of evaluation, i.e. evaluation with the purpose of improving future teaching [2]. Focusing on evaluations of readings, the analysis can inform future selections of course readings. In contrast to standard, cross-sectional approaches, the analysis here can uncover student heterogeneity in evaluations. This can inform efforts to design more inclusive teaching environments [4].

Important result

Students **mostly agree on evaluations**, but **some heterogeneity w.r.t. English (ctr. Danish) readings**

Data

Over the course of five lectures in an undergraduate course on research methods in political science ("Metode 2") at the Department of Political Science (DPS) in spring semester 2017, I collected data on students' evaluations of the readings for each lecture. Building on the approach presented in [5], I collected evaluations using Google Forms surveys which students accessed using a customized `bit.ly` link provided at the end of the lecture.

In addition to evaluation questions, each survey asked students to provide a custom six-digit ID number calculated as follows:

- Digits 1-4: date of birth in `ddmm` format
- Digits 5-6: last two digits of KU ID number
- E.g.: birthday Dec. 24th, KU ID *abc123* → 241223

The design of the custom ID serves two purposes. First, it is *almost unique*. In a class of 100 students, the probability of every single custom ID being unique is $e^{\frac{-100(100-1)}{2\times 36500}} = .87$. Second, it is *credibly anonymous*, i.e. it is not possible for a teacher to infer the identity of the student. The anonymity minimizes the risk of students opting out of the survey out of hesitance to provide negative evaluations. Future designs may, within feasibility constraints, incorporate more sophisticated anonymization methods such as hash functions.

The final data set contains a total of 1,218 evaluations of 8 readings from 317 students. I use multiple imputation to handle missing responses.

Replication data, as well as code to implement the procedure in other classes, is available at `github.com/fghjorth/tlhe17`.

Methods

I apply **k-means clustering**, a method for estimating distinct k (a number set by the researcher) 'types' of students based on ratings patterns. For simplicity, I set $k = 2$. In Figure 4 below, I illustrate the result of the clustering analysis, showing the ratings for two select texts from the class for the two estimated types of student.

I also estimate the coefficient of student type on rating in an OLS regression, i.e. the average rating difference between the two types. The result is presented in Figure 5. To put the result in context, Figure 5 presents the same quantity for the simulated data from the motivating example. If types' evaluations vary across texts (as in Figure 1), the coefficient will be zero; if they move in tandem (as in Figure 2) it will be positive.

Results

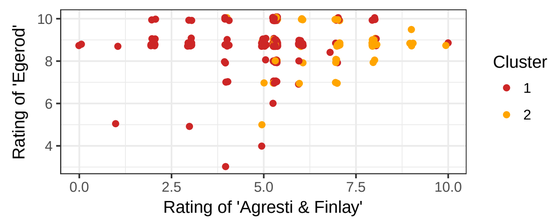

Figure 4: Illustration of result of clustering analysis for two readings.

Compared to cluster 1, students in cluster 2 give 'Agresti & Finlay', an American statistics textbook, much higher ratings. In contrast, for 'Egerod', a broadly popular Danish-language statistics explainer written by a DPS Ph.D. student, cluster 1 students assign marginally higher ratings.

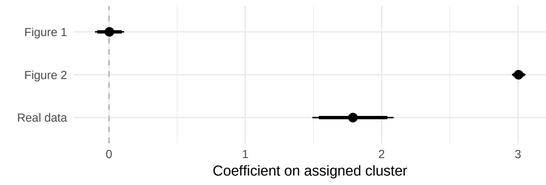

Figure 5: The coefficient of student type on rating in the simulated data in Figure 1 (top row) + 2 (middle row), and the actual data collected in this project (bottom row). Error bars are 90/95 pct. CIs.

The coefficient indicates that the ratings data are somewhere in between the two extremes illustrated in Figure 1+2, but resemble Figure 2 slightly more. In other words, types' evaluations differ somewhat across texts (cf. above), but generally tend to move in tandem.

Conclusion

The project demonstrates the basic feasibility of constructing evaluation panel data by using custom student ID's to link evaluations across classes. There is still room for improvement: most notably, considerable student nonresponse limits the value of the data's panel structure.

The subsequent analysis suggests that, in this particular case, variation in ratings mostly reflect differences across texts, but there is some evidence of student heterogeneity, i.e. distinct 'types' w.r.t. ratings patterns. I discuss the implications for teaching below.

Implications for teaching

The results provide some evidence for student heterogeneity w.r.t. English-language (vis-à-vis Danish) statistics readings. In other words, some students appear to place a large premium on Danish-language material. Future versions of the course may accommodate this heterogeneity by including more material in Danish, a consideration which was not prominent in initial course design.

References

[1] Alison Cook-Sather, Catherine Bovill, and Peter Felten.
Engaging Students as Partners in Learning and Teaching. A Guide for Faculty.
Jossey-Bass Publishers, 2014.

[2] Ronald A. Smith.
Formative Evaluation and the Scholarship of Teaching and Learning.
New Directions for Teaching and Learning, 2001(88):51–62, 2001.

[3] Lotte Rienecker, Peter Stray Jørgensen, Jens Dolin, and Gitte Holten Ingerslev.
University Teaching and Learning.
Forlaget Samfundslitteratur, Copenhagen, 2015.

[4] Clifton F. Conrad, Jason Johnson, and Divya Malik Gupta.
Teaching-for-Learning (TFL): A Model for Faculty to Advance Student Learning.
Innovative Higher Education, 32(3):153–165, jul 2007.

[5] Morten Nyboe Tabor.
Online course evaluation for active learning in mutual respect and appreciation.
In *Make a Difference - Teach and Learn with Technology*, Copenhagen, 2016.

Acknowledgements

Thanks to the students in 'Metode 2' spring 2017, already inundated with survey requests, for providing evaluation data.

Contact Information

- Web: fghjorth.github.io
- Email: fh@ifs.ku.dk

UNIVERSITY OF COPENHAGEN