University of Michigan Learning Analytics Task Force[1]

Course Evaluations at Michigan: What Do We Know?

Based on an analysis of Winter 2005–Winter 2013 course evaluations from the College of Engineering and LSA, we know the following:

- **There is virtually no correlation between grades and course evaluations.** (Pages 3–4 below.)
- **There are differences in evaluations between colleges and LSA divisions.** (Pages 3–4 below.)
- **There is no correlation between perceived workload (Q891) and Q1-4.** (Page 5 below.)
- **There is little evidence of bias on the basis of gender, race, ethnicity, or citizenship status when the data is aggregated at the college or divisional level.** This does not mean that there aren't instances of bias or that the instructor social identity is irrelevant. It simply means that, on average, these factors are not reflected in the quantitative measures. (Pages 6–8 below.)
- **The decline in response rates** since UM began electronic evaluations in F2008 is not uniform and **had no statistically significant impact on the evaluations**. (Pages 9–11 below.)
- **Course size (*not* response rate) is the most important factor in increasing the reliability of the evaluations**. Therefore, inferences should be drawn more cautiously for instructors who teach small courses than for instructors for whom there are evaluations from large courses. (Pages 12–14 below.)

[1] The Learning Analytics Task Force was created in 2013 for a three-year period by then Provost Phil Hanlon at the request of SACUA's Academic Affairs Advisory Committee. It was chaired by Professor Tim McKay and consisted of faculty members from several schools and colleges.

Data description

Analysis:	Mika LaVaque-Manty (Arthur F. Thurnau Professor, Associate Professor, Department of Political Science, Learning Analytics Task Force), David Cottrell (Graduate Student Research Assistant, Department of Political Science)
E&E data:	The Office of the Registrar
Course data:	College Resources Access System, UM Data Warehouse
Data:	**E&Es for all courses in LSA and the College of Engineering, from Winter 2005 to Winter 2013**
Courses, *n*:	107,000 (including DIS and LAB)[2]
Instructors, *n*:	13,801 (including GSIs)
E&E questions:	Limited to the university-wide Q1–4 and Q891.

Q1: Overall, this was an excellent course.
Q2: Overall, the instructor was an excellent teacher.
Q3: I learned a great deal from this course.
Q4: I had a strong desire to take this course.
Q891: The workload for this course was (SA=LIGHT...SD=HEAVY).

The teaching evaluation project's quantitative data analysis was limited to the two largest colleges at the University of Michigan: The College of Letters, Sciences and the Arts, and the College of Engineering. Because of the way E&Es are used in some schools and colleges, there were too many differences for us to include them in this summary analysis. As a result, inferences to other schools and colleges should be made with caution.

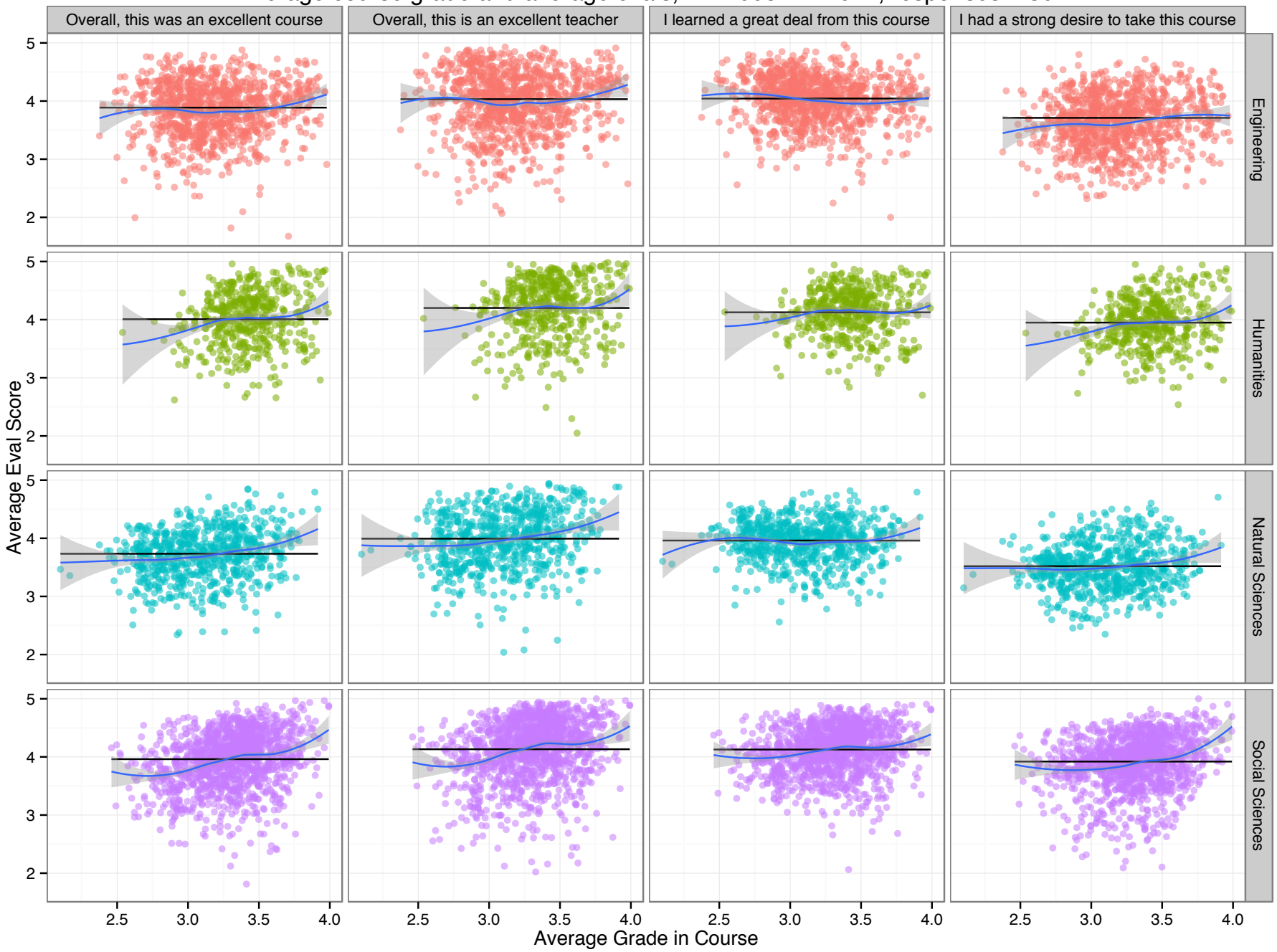
[2] The course components are defined by the Registrar's Office at http://ro.umich.edu/schedule/key.php.

There is minimal relationship between grades and eval scores

The plots on the next page map the correlation between **mean course grades** and **mean course evaluations** for Q1-Q4 in the College of Engineering and the three divisions of LSA.

- Each dot is a course-term, that is, an iteration of a course during a semester. In most cases, it tracks an individual instructor, but some courses have several sections during the same term, in which case it is an average.
- The black horizontal line is the evaluation average. This shows that there are differences between CoE and LSA divisions.
- The blue line is a “loess” curve and the gray shaded area its confidence interval. The loess curve is, roughly, a localized regression analysis: how much does a change in one variable in that area correlate with a change in the other variable. We chose the loess over a linear regression to show that the already very minimal relationship between grades and evals is driven by things on the far margins. But even here, it would be a mistake to infer that very low grades cost you in the evals or very high grades buy you good ones.

Only courses in which more than 30 evaluations were returned are included. See p. 12 below.

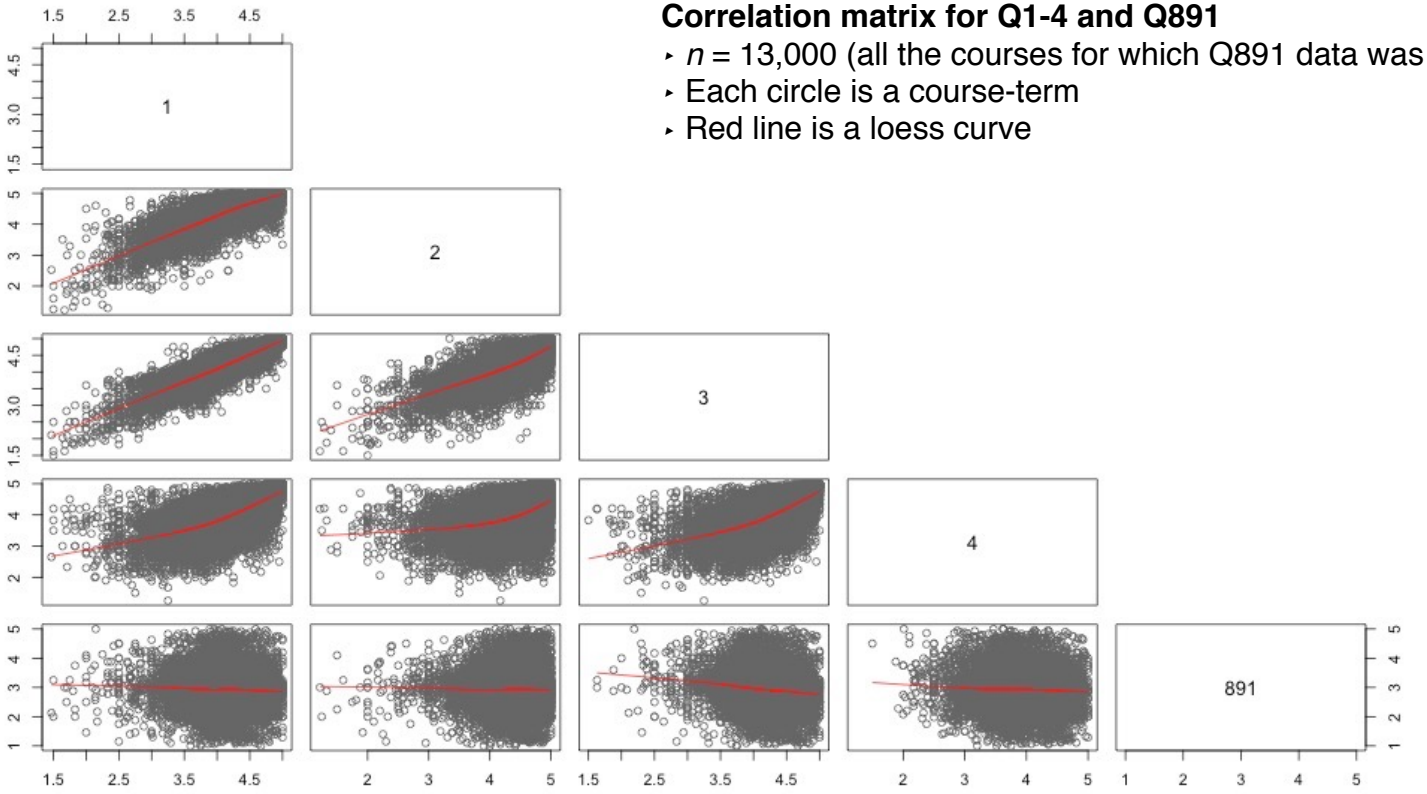

Average course grade and average evals, WN2005-WN2012, responses > 30
Overall, this was an excellent course
Overall, this is an excellent teacher
I learned a great deal from this course
I had a strong desire to take this course
Engineering
Humanities
Natural Sciences
Social Sciences
Average Eval Score
Average Grade in Course
2
3
4
5
2.5
3.0
3.5
4.0

Student perception of workload does not correlate with Q1–4

Question 891 asks students about their perception of the course workload:

> *"The workload for this course was (SA=LIGHT…SD=HEAVY)"*

The correlation matrix plot below shows that there is virtually no relationship between student perception of the workload and the other summative measures. There is very small (but statistically insignificant) relationship between workload and Q3 ("I learned a great deal in this course."): the higher the workload, the less students say they learned.



Correlation matrix for Q1-4 and Q891

- *n* = 13,000 (all the courses for which Q891 data was collected)
- Each circle is a course-term
- Red line is a loess curve

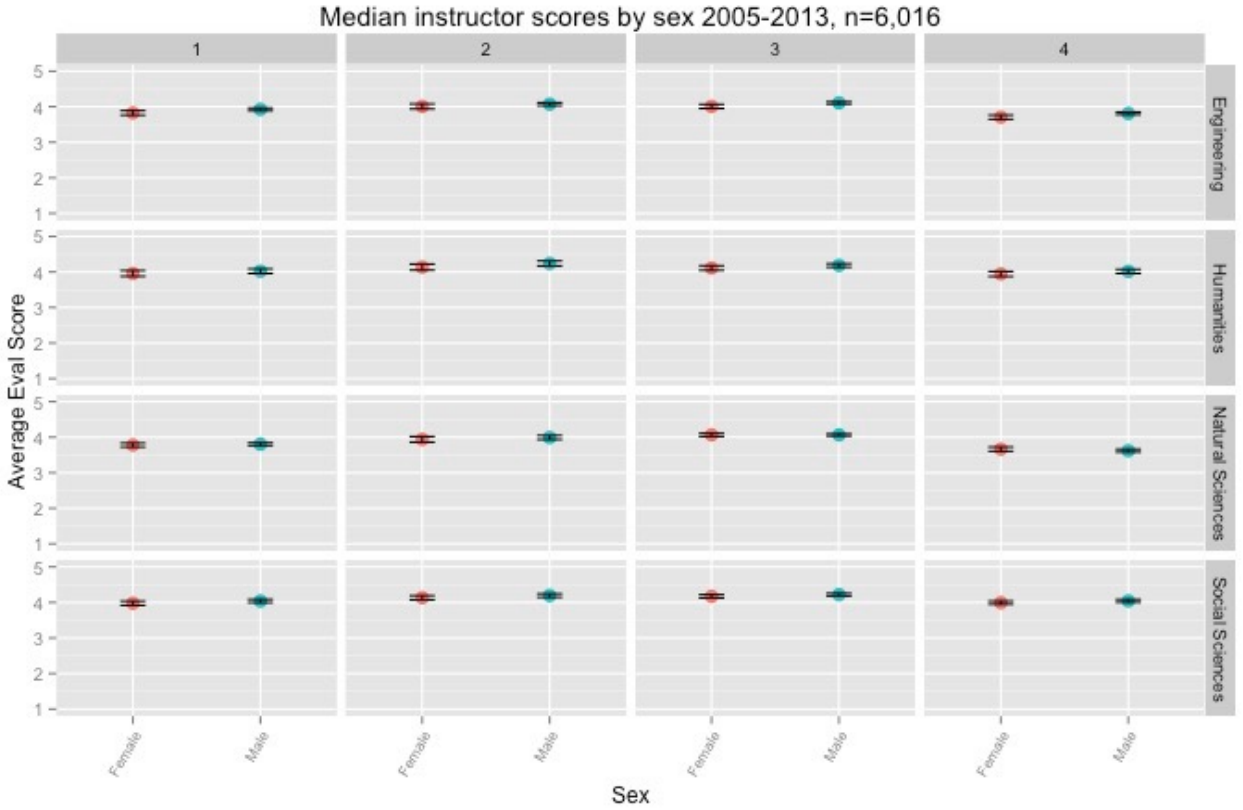
There is little evidence of bias in the quantitative measures

Instructors' own experiences as well as preliminary analyses of the open-ended written comments in E&Es tell us that instructor social identity *is* salient to students. However, this salience does not appear to translate into bias on the basis of gender (measured as instructor sex), race, ethnicity, or national origin, as the plots below suggest.

It is important to note that that these results are at a relatively high level of aggregation. There may be differences at the departmental level. Unfortunately, at that level, the data becomes unreliable and risks identifying individuals.

*Each plot aggregates **by instructor**, not by course, for all courses the instructor has taught and for which more than thirty evaluations were turned in. The plots do **not** include GSIs.*

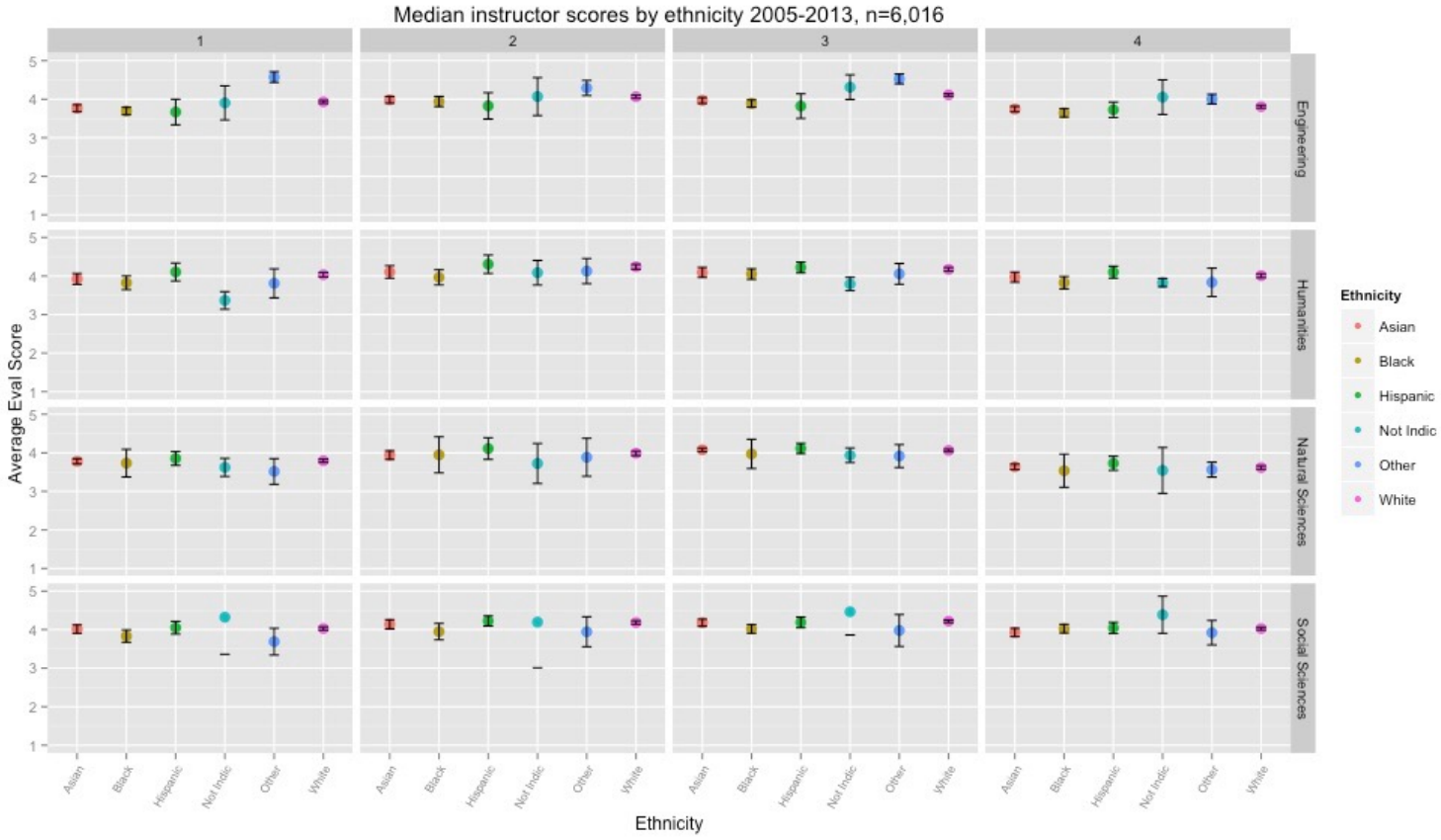
Sex



The error bars (black horizontal lines) in the plot indicate the 95% confidence interval that the number is the true mean. The smaller the population evaluated, the larger the interval.

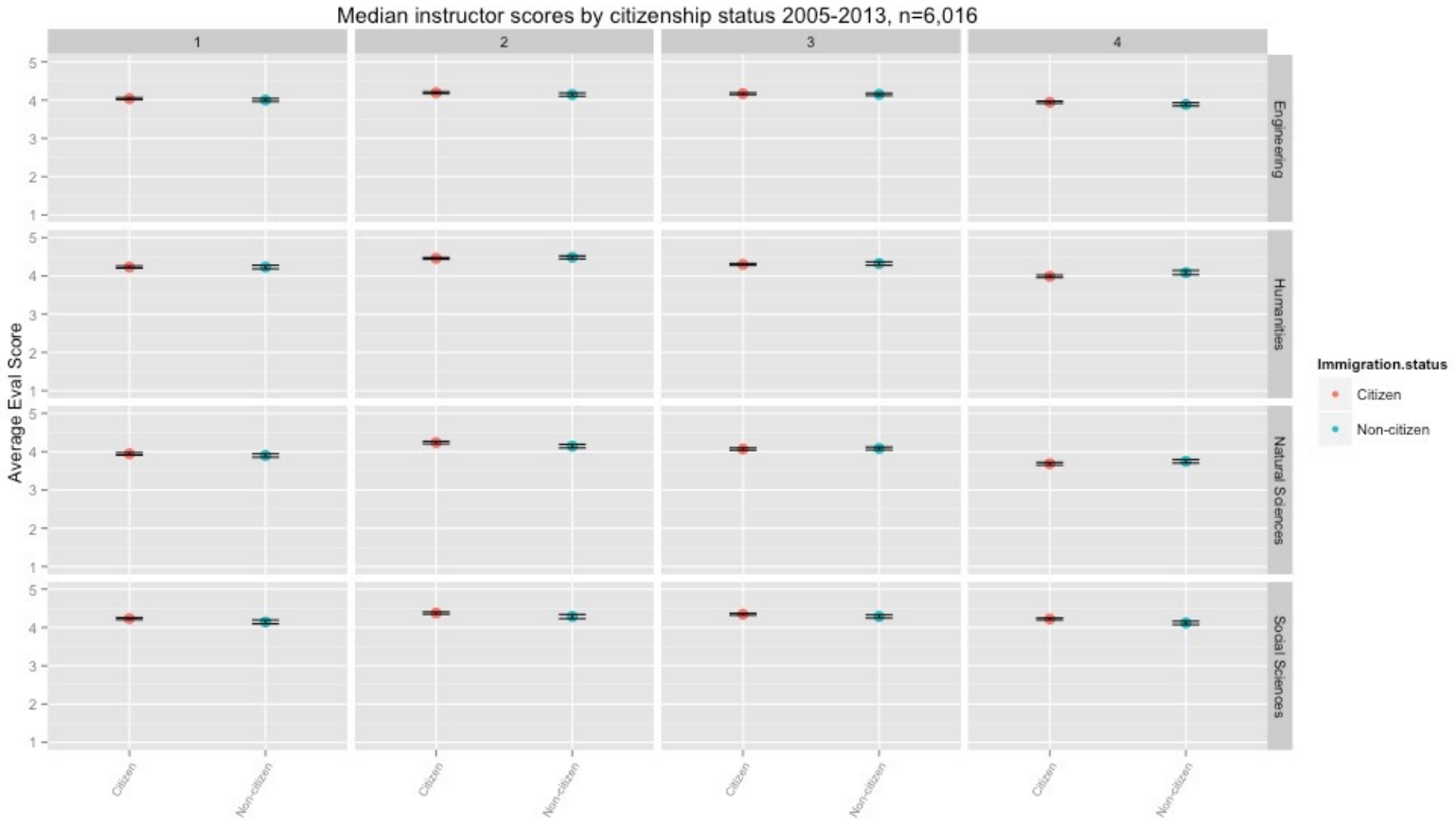
Race and ethnicity

Note: Despite the large *n* of more than 6,000 individual instructors, disaggregating the results into four divisions risks revealing the identities of some instructors from ethnic groups with few representatives at UM. Therefore, in the plot below, we have made the problematic but necessary decision to aggregate Native Americans, Hawaiians, and instructors identifying with “2 or more” ethnic groups into the single category “Other.”



Citizenship

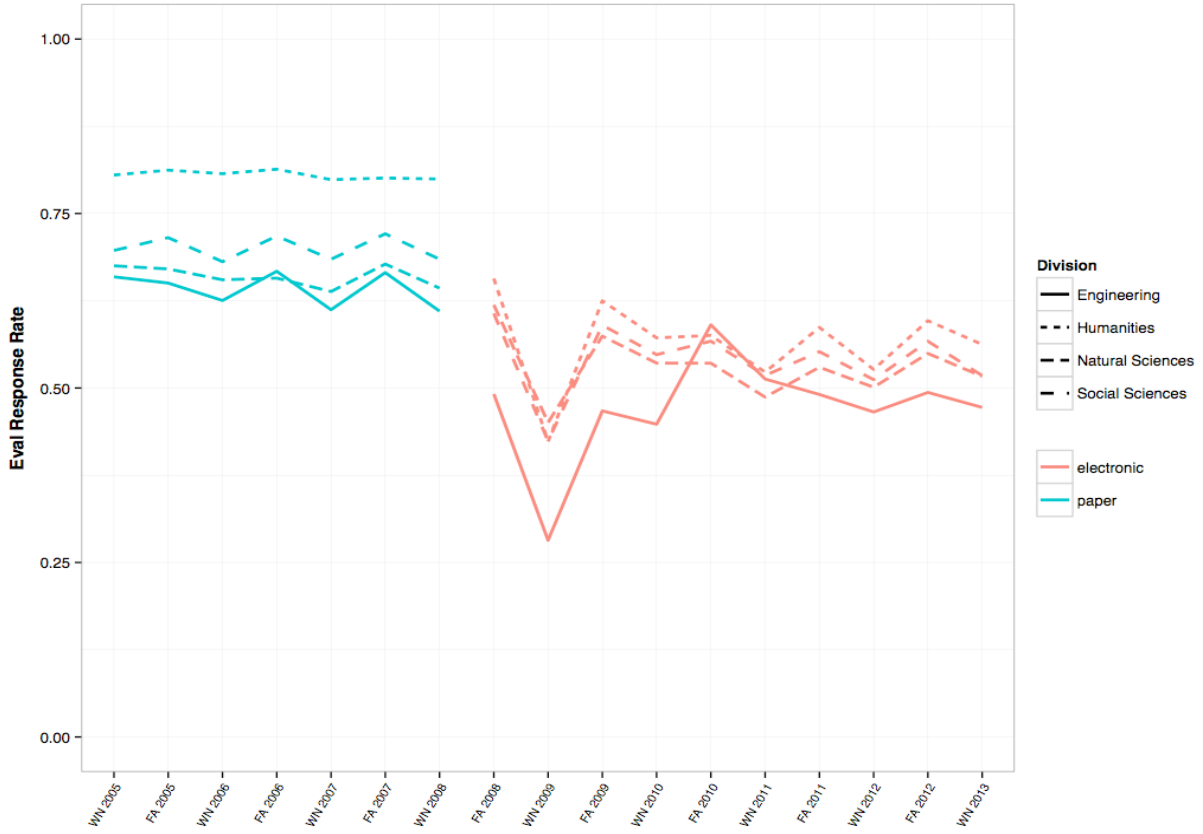
Some research on teaching evaluations suggests instructor national origin (or perception of the instructor's national origin) as well as whether the instructor is a native English speaker can be a source of bias. We have no direct measure of either. The closest, though far from perfect, proxy is the instructor's U.S. citizenship status.



The effect of the transition to electronic evaluations in 2008 was minimal

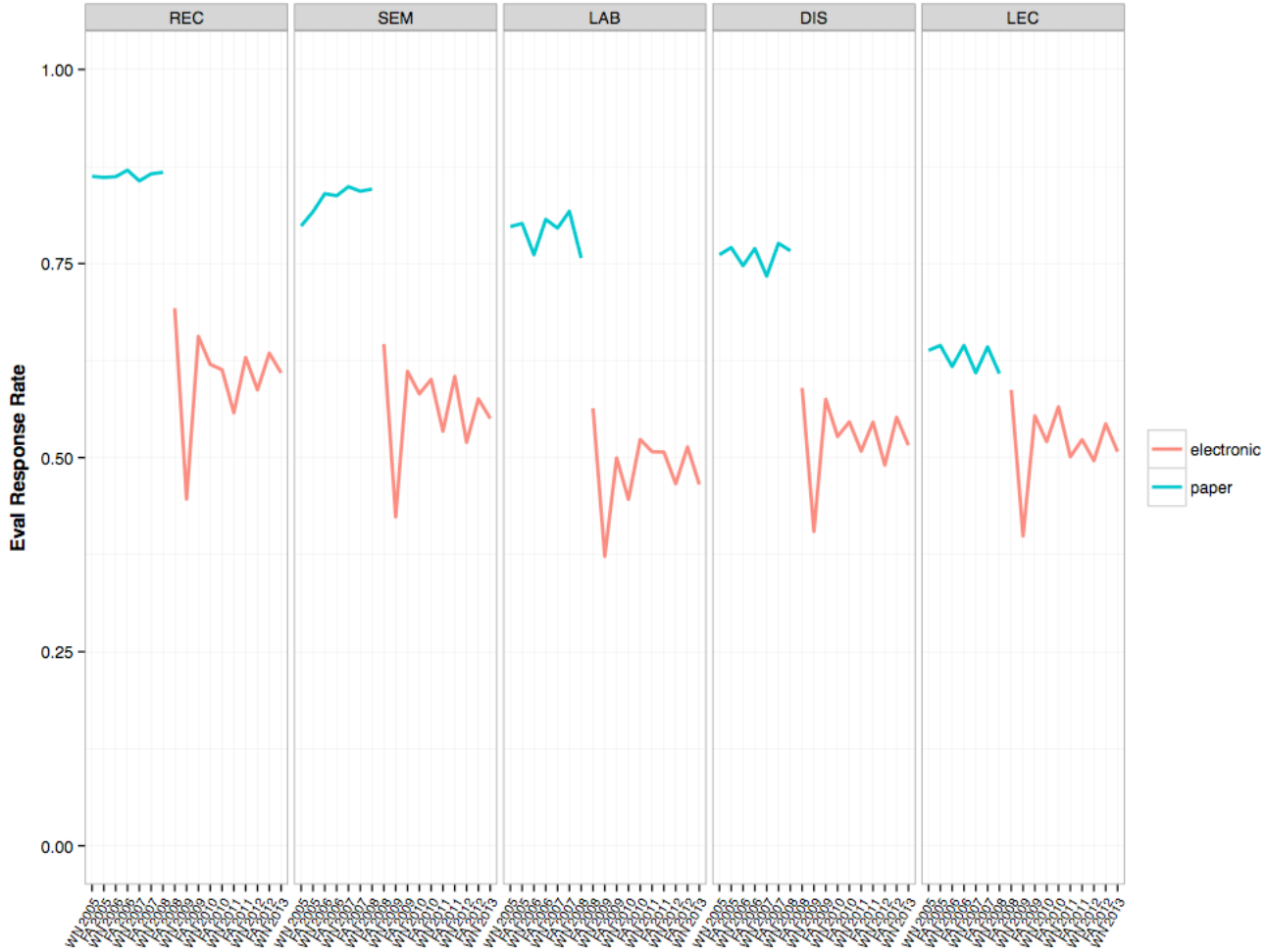
UM transitioned from in-class paper evaluations to electronic evals in Fall 2008. There is a widespread perception that this fundamentally changed things, exacerbated by the fact that in Winter 2009, the still new evaluation system crashed, resulting in a massive drop in response rate.

It is true that the response rate has dropped by 15-20 percentage points *on average, in the aggregate*. A more fine-grained look shows that there is significant variance by college and course type, and that in terms of the actual evaluations, the change has been relatively small.



The plot on the left shows the effect of the switch from paper to electronic evaluations on response rates (*y*-axis, 1=100%). There was a significant difference between Humanities courses and everything else before the switch, whereas afterward, the rates are roughly equal, with LSA divisions moving together, and Engineering being somewhat different.

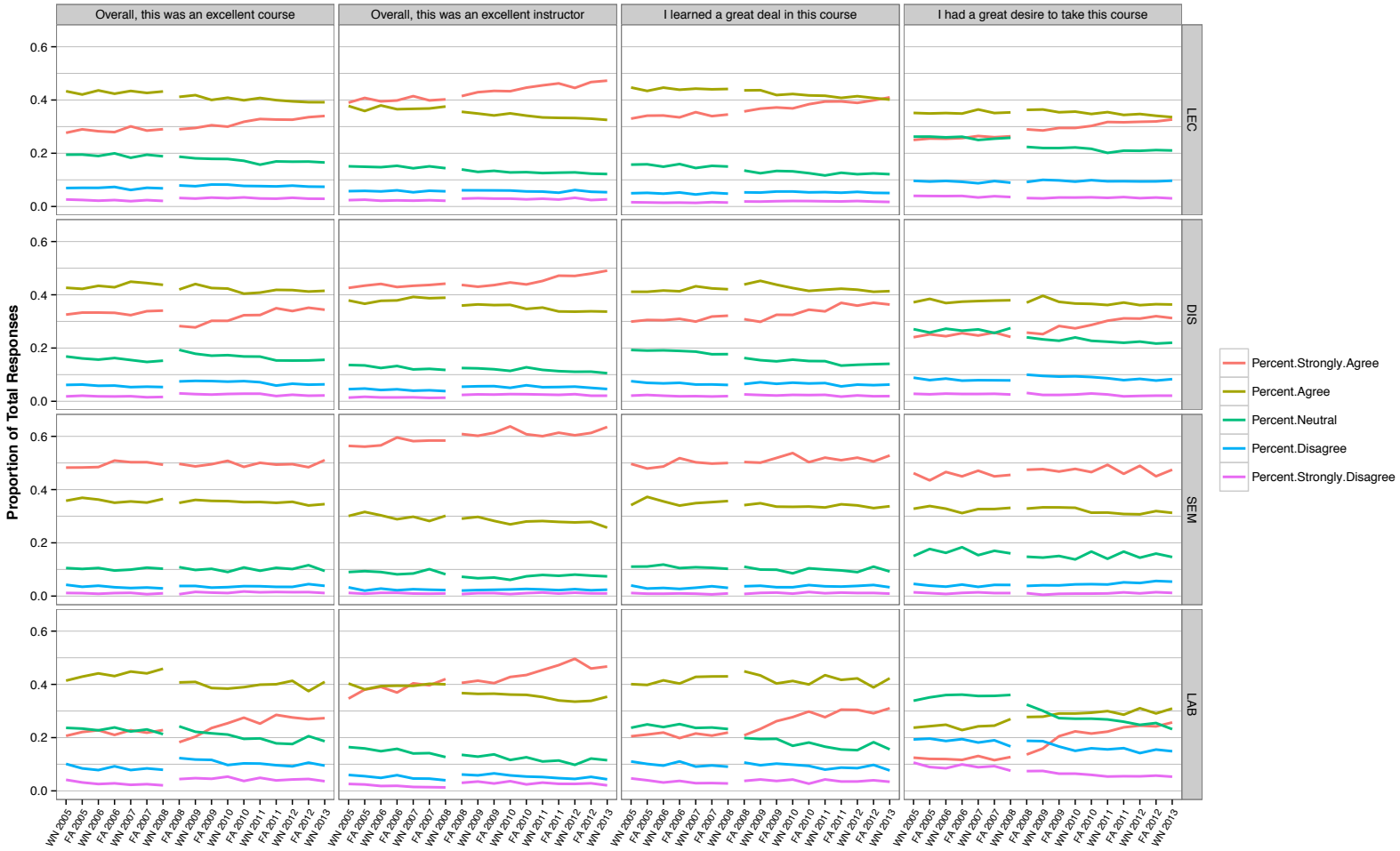
As the plot on the next page suggests, the difference between Humanities and the rest likely has to do with course types: Humanities has the smallest number of large lecture courses.



This plot shows that lecture courses never had a particularly high response rate. Those who showed up in class more or less continue to fill out the evaluations, even when it happens on their own time, outside the classroom. At the same time, recitations (smaller lecture courses, usually upper level) and seminars do continue to have higher response rates than other course types.

The response rate between labs and discussions — most of them usually taught by GSIs — have changed differently: labs have dropped more than discussions.

The plot below shows how the responses changed after the shift to the electronic evaluations



It's not the response rate as much as the number of respondents that matters.

Much of the literature on student ratings says that the ratings are valid if the response rate is about 70%. "Validity" is usually understood as inter-rater reliability: the variance of responses of two iterations of the same thing is the same. While we are agnostic on whether the evaluations used at Michigan are really *evaluations of teaching*, we believe they should be measuring *something* (whatever the something is) about either the *instructor* or the *course.* If they do, we would expect either two iterations of the same course or the same course by the same instructor to have a roughly the same responses. Of course, we hope that instructors improve and know that they frequently change strategies, so perfect correlations would be a problem. But, on average and in the aggregate, we would want either the instructor or the course the be relatively consistent.

We see the reliability of the evaluations to increase as the number of evaluations turned increases.

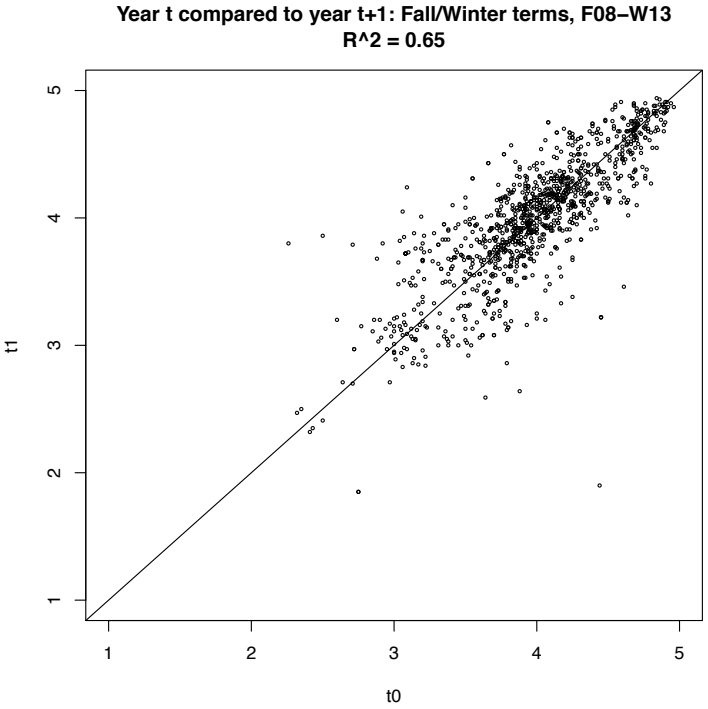
The general message of the following plots is that **in courses fewer than 50 students, *regardless of the response rate*, the scores for a single course are not a reliable reflection of the instructor in general. In particular, low scores should not be held against the instructor.** This does not mean that the numbers are meaningless, simply that the signal to noise ratio is lower.

All units, but particularly those whose instructors regularly teach small courses should consider additional ways of evaluating teaching. In reporting and analyzing the scores, the number of responses should be aggregated as much as possible, instead of a narrow focus on single courses.

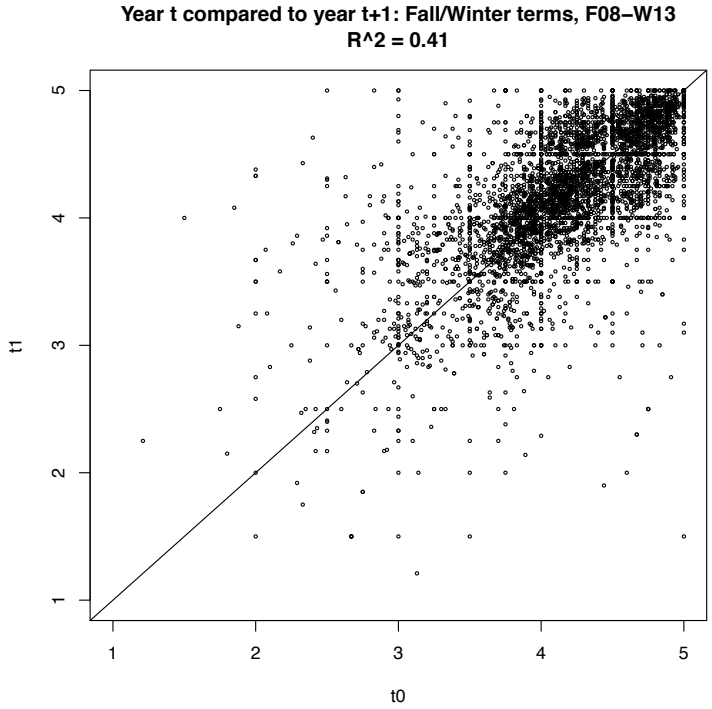
Because of this, in this report we have used **thirty responses per course** as the bottom limit of inclusion in our analysis. It is less than ideal, but increasing the cutoff to the even more reliable fifty would have ruled out too many courses and instructors.

Same instructor, same course

The plots below indicate how much the overall course evaluation (Q1) for one instructor teaching a course in year one (t0) correlate with the same instructor teaching the same course in year two (t1). On the left are courses the Registrar's Office regards as large: enrollment greater than 70. On the left are courses smaller than that. The answer is that there is a very significant difference, *depending on the size of the course*, in predicting the instructor's Q1 rating for the course.



Enrollment > 70



Enrollment < 70

Same instructor, over time

The plot below compares how well an instructor's past courses predict his or her current courses, for all four E&E items. The cutoff date (fall 2011) was chosen to ensure statistically significant sample size. As the plot suggests, there is a significantly greater correlation for courses with **more than 50 evaluations turned in** than with those with fewer. Again, these are *not* response rates, but absolute numbers.