Benjamin M. Schmidt

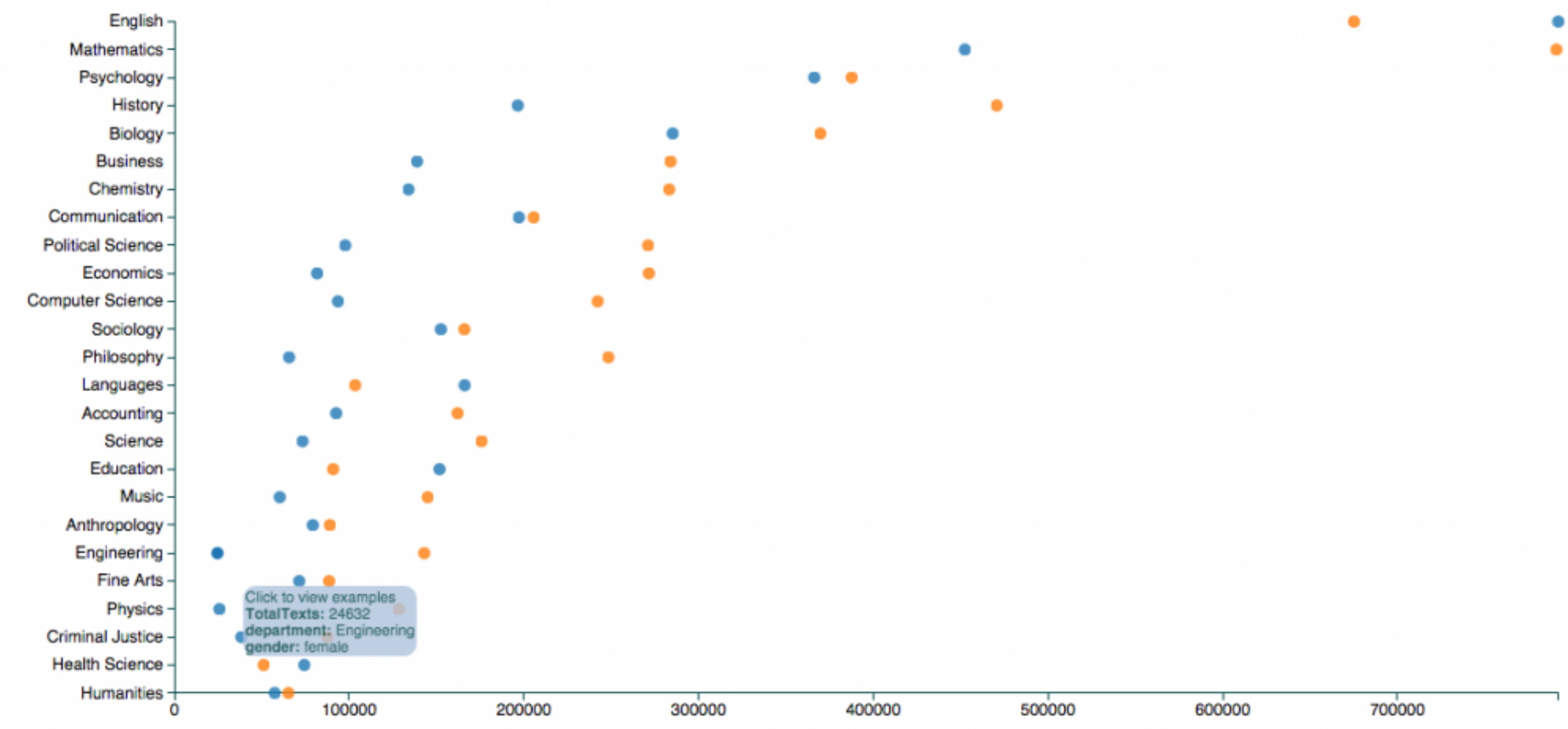
Rate My Professor

Just some quick FAQs on my professor evaluations visualization: adding new ones to the front, so start with 1 if you want the important ones.

-3 (addition): The largest and in many ways most interesting confound on this data is the gender of the *reviewer*. This is not available in the set, and there is strong reason to think that men tend to have more men in their classes and women more women. A lot of this effect is solved by breaking down by discipline, where faculty and student gender breakdowns are probably similar; but even within disciplines, I think the effect exists. (Because more women teach at women's colleges, because men teach subjects like military history than male students tend to overtake, etc). Some results may be entirely due to this phenomenon, (for instance, the overuse of "the" in reviews of male professors). But even if it were possible to adjust for this, it would only be partially justified. If women are reviewed differently because a different sort of student takes their courses, the fact of the difference in their evaluations remains.

-2 (addition): This no peer review, and I wouldn't describe this as a "study" in anything other than the most colloquial sense of the word. (It won't be going on my CV, for instance.) A much more rigorous study of gender bias was recently published out of NCSU. Statistical significance is a somewhat dicey proposition in this set; given that I downloaded all of the ratings I could find, almost any queries that show visual results on the charts are "true" as statements of the form "women are described as x more than men are on rateMyProfessor.com." But given the many, many peculiarities of that web site, there's no way to generalize from it to student evaluations as used inside universities. (Unless, God forbid, there's a school that actually looks at RMP during T&P evaluations.) I would be pleased if it shook loose some further study by people in the field.

-1. (addition): The scores are normalized by gender and field. But some people have reasonably asked what the overall breakdown of the numbers is. Here's a chart. The largest fields are about 750,000 reviews apiece for female English and male math professors. (Blue is female here and orange male–those are the defaults from alphabetical order, which I switched for the overall visualization). The smallest numbers on the chart, which you should trust the least, are about 25,000 reviews for female engineering and physics professors.



0. (addition): RateMyProfessor excludes certain words from reviews: including, as far as I can tell, "bitch," "alcoholic," "racist," and "sexist." (Plus all the four letter words you might expect.) Sometimes you'll still find those words typing them into the chart. That's because RMP's filters seem not to be case-sensitive, so "Sexist" sails through, while "sexist" doesn't appear once in the database. For anything particularly toxic, check the X axis to make sure it's used at a reasonable level. For four letter words, students occasionally type asterisks, so you can get some larger numbers by typing, for example, "sh *" instead of "shit."

1. I've been holding it for a while because I've been planning to write up a longer analysis for somewhere, and just haven't got around to it. Hopefully I'll do this soon: one of the reasons I put it up is to see what other people look for.

2. The reviews were scraped from ratemyprofessor.com slowly over a couple months this spring, in accordance with their robots.txt protocol. I'm not now redistributing any of the underlying text. So unfortunately I don't feel comfortable sharing it with anyone else in raw form.

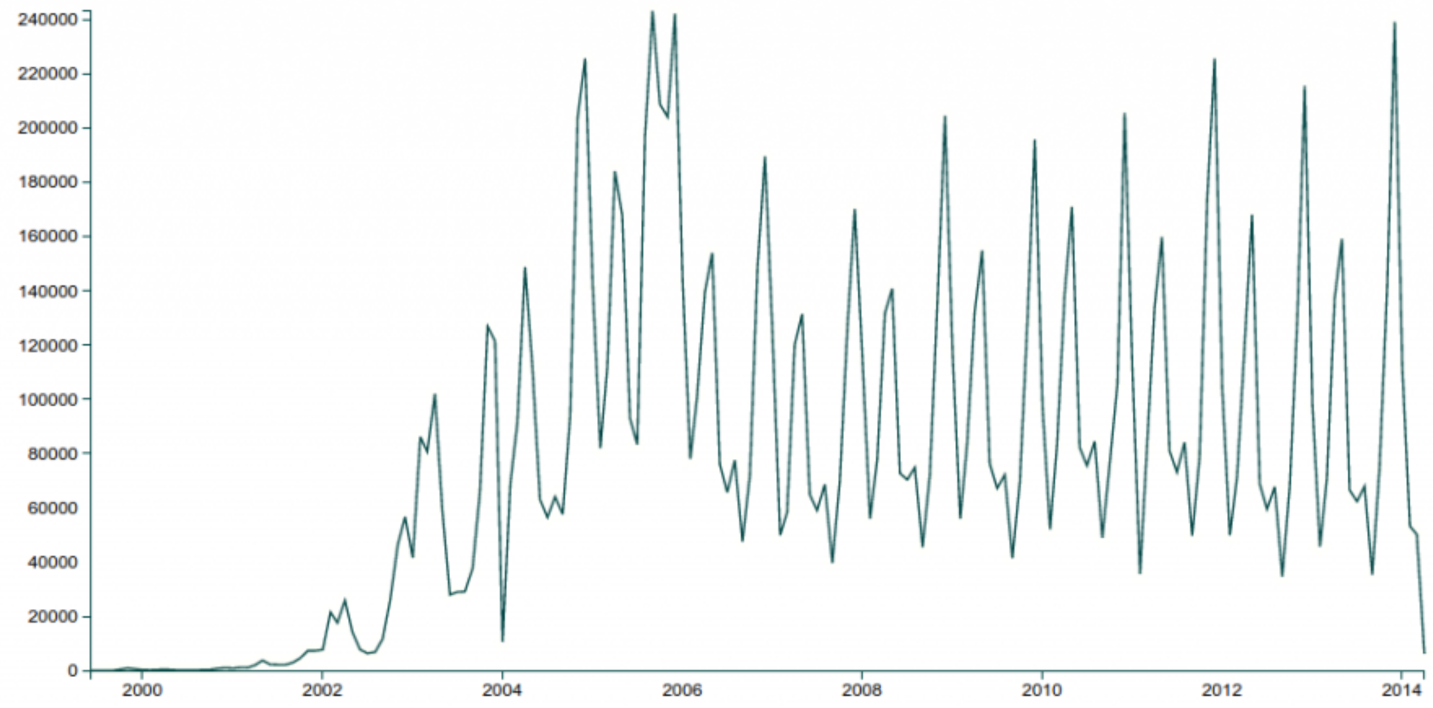
3. Gender was auto-assigned using Lincoln Mullen's gender package. There are plenty of mistakes–probably one in sixty people are tagged with the wrong gender because they're a man named "Ashley," or something.

4. 14 million is the number of reviews in the database, it probably overstates the actual number in this visualization. There are a lot of departments outside the top 20 I have here.

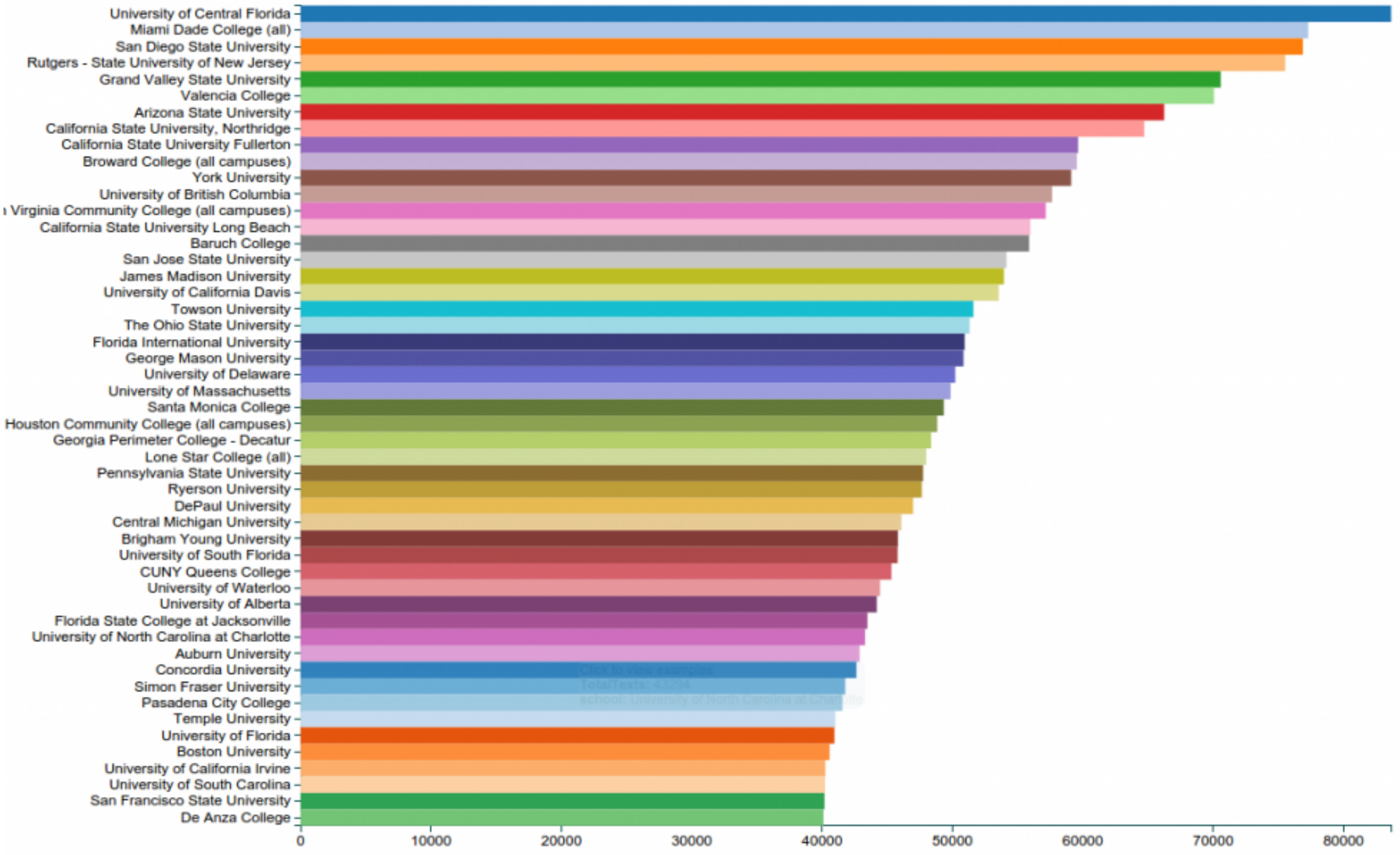
5. There are other ways of looking at the data other than this simple visualization: I've talked a little bit at conferences and elsewhere about, for example, using Dunning Log-Likelihood to pull out useful comparisons (for instance, here, of negative and positive words in history and comp. sci. reviews.) without needing to brainstorm terms.

6. Topic models on this dataset using vanilla sets are remarkably uninformative.

7.People still use RateMyProfessor, though usage has dropped since its peak in 2005. Here's a chart of reviews by month. (It's intensely periodic around the end of the semester.



8. This includes many different types of schools, but is particularly heavy on masters and community colleges in the most represented schools. Here's a bar chart of the top 50 or so institutions:



This entry was posted in Uncategorized on February 6, 2015 [http://benschmidt.org/2015/02/06/rate-my-professor/] .

47 thoughts on “Rate My Professor”

Asif

February 6, 2015 at 10:01 pm

Excellent work! Thanks for sharing the details.

J Z

February 7, 2015 at 3:04 am

How do you avoid negative expressions like “not a genius”?

ben Post author
February 7, 2015 at 2:24 pm

I don’t–and this will certainly increase the noise on the thing. I do have the full text of the reviews myself, which I use for sanity checks–but it’s worth keeping in mind. (And entering terms like “not boring,” to see that it’s used at about 2% the rate of “boring”).

ycd
February 7, 2015 at 3:30 am

Fantastic! It would also be great to have an option to split by the gender of the reviewer!

ben Post author
February 7, 2015 at 2:25 pm

I really wish this were possible–it’s a major confound, because there are probably more women in classes taught by women. (We know there are more more female profs in fields that have more women majors).

majining
February 7, 2015 at 2:41 pm

Though it would be imperfect (for a number of reasons), you could do a search for words that indicate sexual attraction that are as gender neutral as possible. (Test for gender neutrality on another data-set where the genders are reported.) Make adjustments based on the size of the LGBQT population, and leave plenty of margin for error. Differences observed by field in the responses should correspond inversely to the gender of the respondents (again, with plenty of margin for error), giving a rough approximation of the gender divide among the reviewers, by field. The results for “hot” are interesting, here, because there are large gaps between male and female, in both directions, and they do seem to reflect my (non-expert) understanding of the student demographics in those fields (e.g. engineering, business, and education).

ben Post author
February 7, 2015 at 3:14 pm

Can you think of a set with full two-way gender reporting, discussion of attraction, and full text? I’d sure love to see one, it would make for something really interesting on its own.

But I think this problem is so difficult on this set, and the results would be so provisional, that it’s not worth doing. If I were going to build a linear model on this, I guess I’d be regressing against the ICOADS-reported gender breakdown in each major. (This is imperfect, because major counts don’t map onto numbers in the classroom– for instance, math majors are incredibly male, but lots of people take intro math classes). But there’s yet another confound I have no way to test for: who’s more likely to head off to RateMyProfessor.com, a man or a woman? By how much? Are heterosexual women more likely to describe women as attractive than hetero men about men? We know that male faculty tend to be significantly older than female faculty, and more in some fields than others… does that need to be controlled for in some way? Women are more likely to teach online courses, in which attractiveness can’t be called into play.

Most importantly of all, though: I didn’t scrape the chili peppers from the RMP site. So I’m missing the data I’d really need for this.

Of course many of these questions apply to the

Wayne Lobb
February 7, 2015 at 1:56 pm

Interesting work.

But do you normalize results for the numbers/counts of male professors versus female professors? If not, you should. As far as I can tell, currently in the US males account for about 58%, females 42%.

Did anyone sample the raw data for the word “genius”, for example, to see what percent of the time the word was used to describe the professor as opposed to describing, say, Mozart? I ask this particular question because the Music category in the NY Times article had the highest apparent disparity per gender.

Why do I get such different return counts for “his, he” [without quotes, with the comma] versus “his he” [no comma]? How can one explain the difference between results for “his he” and “her she”? And what’s going on with “helpful” versus “unhelpful”, which together seem to say that female professors are both more helpful than males and more unhelpful than males? Are you searching only on whole words? Or on partial words too? I would have to think whole words and phrases.

Regarding the instructions for the interactive chart, do you search case-sensitively (as stated) or do you not? I get the exact same return results for “genius”, “Genius” and “GENIUS”.

About use of pronouns, isn’t it true that a review of a female professor will use “she” and “her” etc while a review for a male would use “he” and “his”? So of course there’d be a huge gender split in pronouns. Or do I misunderstand your instructions?

I’m a mathematician by training, recently retired from a long career in developing business- and life-critical software. In my experience, Big Data in any form is extraordinarily difficult to slice and dice for true and useful information. One would have to work for days or weeks with multiple combinations of search terms to pry truth and utility from your tool – and from nearly all such tools. One HAS to sample statistically significant subsets of returned results, one by one, to rule out or account for complex usages, such as “the genius of composers such as Mozart and Beethoven” versus “this professor is a genius” or “this professor thinks he’s a genius but he’s absolutely not.”

ben Post author
February 7, 2015 at 2:59 pm

So many questions!

Yes, I’ve looked at hundreds of these individual reviews to get a sense of what’s in there. We’ve put a lot of care into designing a general purpose tool that includes features like random sampling of uses for spot-checking, rather than just rank-ordered search results. I don’t feel comfortable (legally or ethically) sharing reviews of individual professors myself in the web tool; but I just spot checked “genius” in 100 reviews of male music teachers, and could only find one that wasn’t in reference to the professor. (Along the lines of: “Don’t take this class if you’re not a genius.”)

Some pronouns (I, me, our, we) are gender insensitive. You can also compare the union of "he,she" to each other.

What's going on with "helpful/unhelpful," I think is something actually quite interesting: that reviews of women *start off* with the presumption that they should be helpful, and then proceed to assess them on that. Just like men are more likely to be a "genius" or and "idiot." But that will require me reading some moe of these to substantiate.

And yes, everything is normalized; if it weren't, you'd see massive biases in one direction or the other for every field/gender.

I was in error describing it as case-sensitive–that's now corrected.

Barb

February 7, 2015 at 3:21 pm

Can you give any sense of how large a difference must be to be substantively significant? I'm assuming given the N's that even small differences are statistically significant.

ben Post author

February 7, 2015 at 7:08 pm

This is a really complicated question, actually, because there are a lot of lurking variables. I have about 25,000 reviews for the smallest classes, but in many cases that will include 20 or 30 reviews of the same professor.

If you want a standard test of significance on gender as a whole, one way to think about it is to view each discipline as an independent sample and how many disciplines cut in the same direction. If anything fewer than 8 of the fields are in the wrong direction, you'd technically be passing a 95% significance test on a two-sided binomial distribution (if I'm doing my math right).

But there are so many sources of non-random error, and the overall size of the set is so big, that I'd be reluctant to use a test like that to claim significance.

Stas K
May 10, 2015 at 7:51 am

Barb, there is no sense in invoking the concept of statistical significance because the samples are self-selected. And, of course, as Ben said, they are not homogeneous and would require an enormous amount of adjustment on the variables that he does not have.

John MUccigrosso
February 7, 2015 at 6:34 pm

I'do be curious to see the prevalence of categories of words (like "physical descriptors") to see whether these tend to differ by sex generally. That is, do students tend to describe female instructors' physical appearance (etc.) more then they do males'?

(Punctuation seems to be excluded. I had to search on doesnt not doesn't. Also putting two spaces between words in one search put the markers to the left of the axis.)

ben Post author
February 7, 2015 at 7:10 pm

You can jam together a bunch of lists of words together with commas, which I've done; it's hard to come up with a comprehensive list, though. I'm going to go at some point through the most frequent queries and class them like this, because I think in the case of physical appearance/clothing/etc. it shows something very surprising to most people; that men have their appearance described as often as women. It's in evaluative language that the real differences arise.

Lauren
February 8, 2015 at 6:45 am

I'm assuming the x-axis (unlabeled) is percent of total rather than number of reviews, but if you could make that more clear that would be awesome!

ben Post author
February 8, 2015 at 3:47 pm

It's uses per million words of text for that combination of gender and major: I'm trying to make this more clear on the chart.

Moin Syed
February 9, 2015 at 1:27 am

it seems like people are using more descriptors for men in general (positive and negative). is this true, and have you made any adjustments?

ben Post author
February 9, 2015 at 4:27 am

It's uses per million words overall in that set, so if men are getting more adjectives, women must be getting something else. There are certainly some classes of things that men have more of, but I'm not sure if adjectives as a whole is the case.

Bruno Verbeek
February 8, 2015 at 10:12 am

Thanks for this! Most interesting! I do hope you get to write up a more detailed analysis. I would look forward to that. Question: do the data allow for spotting trends? E.g., the use of "brilliant" in reference to a female teacher since 2000?

ben Post author
February 8, 2015 at 3:39 pm

I can do trend lines, but they tend to not be especially illuminating as far as I can tell; since RMP's user base has changed (it was originally mostly California, where it started; now it's heavily in Canada) the trend lines tend to just be expressing other patterns underneath.

Frances Woolley
February 8, 2015 at 2:09 pm

What percent of the profs with non-Western names do you think you've captured? It would be interesting to try to flag those profs and see what kind of comments they get.

A lot of the gender stuff is kind of predictable, but I found it interesting to see what the ratemyprofessors comments say about differences across disciplines in teaching methodology and teaching style. E.g. try things like textbook, memorize, discussion, lecture, think, bias.

ben Post author
February 8, 2015 at 3:47 pm

Interesting idea. I was unable to find a conclusive gender based on the Social Security Administrations list of names for about one in 5 professors in the set. I experimented a bit with parsing the last names against census data to use as proxies for ethnicity, but with only mild success. It's often hard to tell foreign names apart from rare names, though, and even mildly common foreign names will show up in the census sets.

Stas K

May 10, 2015 at 7:56 am

“accent” seems to be an issue for females in hard sciences.

Comparing the **scale** of the horizontal axis on the default “funny” vs. “explain”, let alone “example”, says a lot about the student expectations of what is supposed to happen in the classroom.

Ben

February 8, 2015 at 3:46 pm

Why do you think men are much more likely to be viewed as arrogant? Could the combination of arrogance and brilliant reveals something about how men and women teach in the classroom. I wonder if the same professors viewed as brilliant by some students are viewed as arrogant by others.

Questioner

February 8, 2015 at 4:46 pm

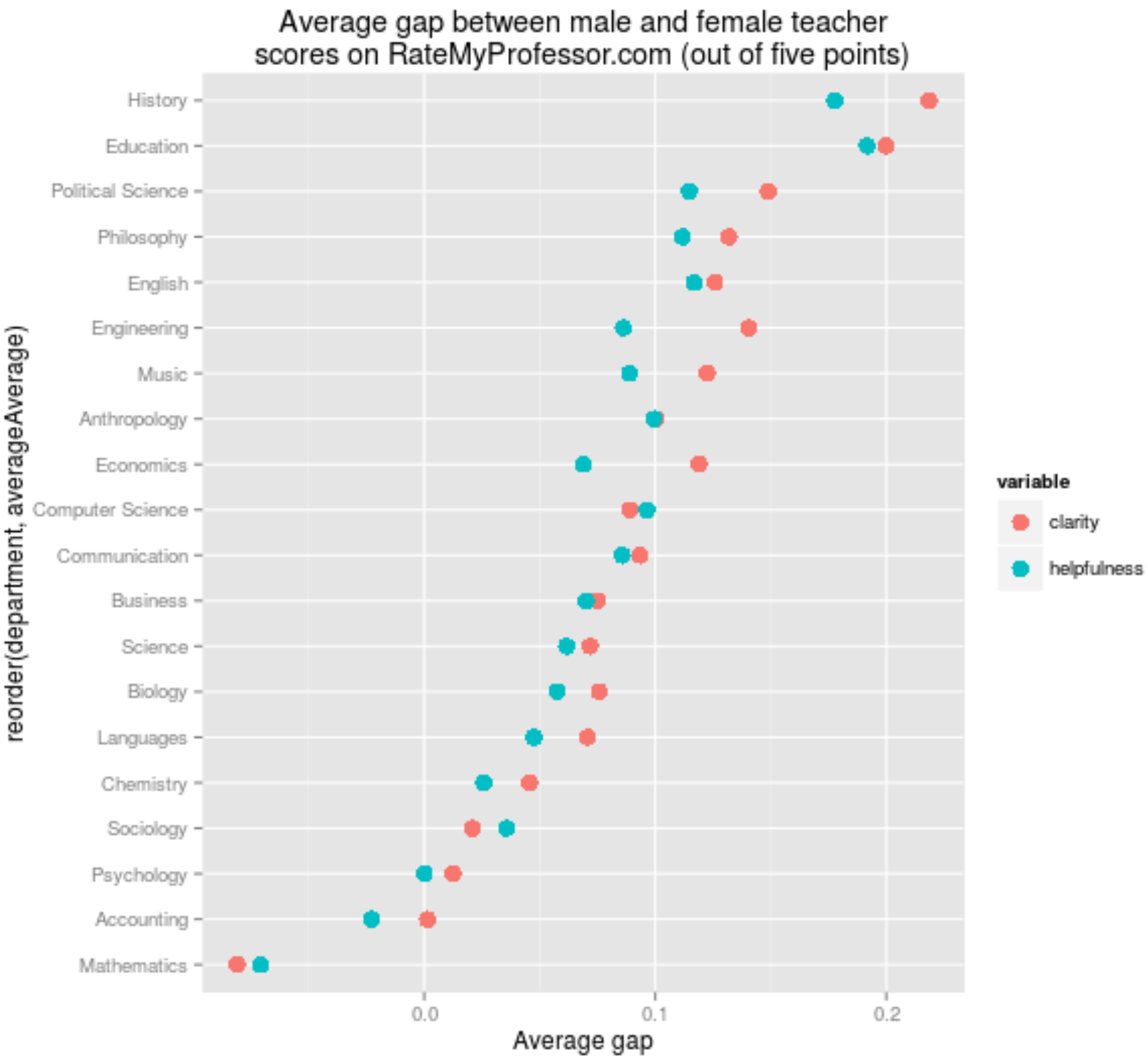
Is it possible to use the numerical ratings on the site to determine which gender is rated higher on average?

---

ben Post author

February 8, 2015 at 4:52 pm

Yes, I actually ran this a couple months ago to show the gaps. On average, men tend to score about one-tenth of a point higher (out of five):



Lisa Renee Pomerantz
February 8, 2015 at 5:33 pm

Do we know whether the use of different terms could be correlated with different teaching styles used by men or women? If you type in typically feminine traits such as “caring” or “concerned” women dominate.

ben Post author
February 9, 2015 at 4:24 am

This is certainly reflecting some real differences in the classroom, yes. But as several people have observed, in a lot of cases women dominate both for those traits bot also their opposites: which suggests that women may be assessed for their caring, whether their teaching style deploys it or not.

Mark Crovella

February 10, 2015 at 8:42 pm

The popular press (NYT) has presented this work as mainly showing evidence of negative attitudes toward female teachers. But it's not so clear to me – for each of the following words I get more hits for women than men:

disorganized
organized
confusing
clear
helpful
unhelpful
wonderful
awful
friendly
unfriendly

Comments?

ben Post author

February 11, 2015 at 2:35 pm

Yeah, both this and the previous comment raise an important point. When it comes to word choice, gender bias is zero sum game–for everything that's more female, something else is more male.

And as you say, I don't think you can draw the conclusion that the split is simply "positive for men, negative for women."

With the paired lists of opposites you give, and some of the other ones you can find ("warm/cold","caring/uncaring"–to some degree) for women and the ones for men ("genius"/"idiot","boring/interesting","funny/corny"), I think–and this is all provisional–you're actually seeing a difference the the things teachers are evaluated for. Men and women might be playing by a slightly different set of rules. (Or, of course, something real about their teaching styles might be bringing it out, though I doubt this is universally true).

So is it possible we just have separate but equivalent standards? That seems unlikely to me, although it take a lot of work to rule it out on this dataset alone. First, because other studies have shown a pro-male bias in cases where teacher gender can be masked. Second, because a few high-profile words don't show a split: best is male, worst is female. And last, because some negative female words are especially harmful just because of their bias–I've been getting many people searching for "shrill" and "bossy," neither of which are especially common but which do effectively remove authority from female figures more strongly than whatever the male counterparts ("arrogant","corny"?) might be.

Not every pair of opposites shows this trend.

Professor Nerd
February 11, 2015 at 1:16 pm

One of the most interesting things I have noticed is that people are reporting on this project as evidence of gender discrimination, but there are some phrases that might question this interpretation.

Some interesting ones: **jerk** has a big gap, as well as **asshole**, and try **stick, ass** (essentially looking for phrases related to having a stick up one's ass)

Professor Nerd
February 11, 2015 at 2:21 pm

More words that are more commonly used with male faculty: **jackass, idiot, moron, sexy, smells, fat, corny, unfunny, funny, old, bad breath**

Words that are more commonly used with female faculty: **Unqualified, retarded, ugly, dresses** (probably as in "she dresses…"), **discriminates** (interestingly large gaps by field with that one), **"white people"**

Other interesting entries: **aspergers, autistic, racist**

ben Post author
February 11, 2015 at 2:36 pm

See my comment on the post below.

eric palmer
February 11, 2015 at 8:25 pm

Such a beautiful tool! Intriguing … useful?

Since I don't know the relative proportion of courses taught by men vs. women in the different disciplines, but I am sure they vary greatly by discipline, I don't know too well what to say about results, without applying the relevant correction.

How to tease out that variable? If "is" is a word that occurs as frequently in reviews regardless of gender, then women receive about 0-10% more "is", plus or minus some, on average, in all disciplines than men do, with the single exception of criminal justice. That's possible, given the bottom-heavy distribution of women in academe, doing more of the teaching labor, but I don't think it's likely, all around. It suggests to me that the writing about women is lengthier than writing about men … and that's another factor to consider, when looking at those orange and blue spots.

"caring" … are the results of "not caring" screened off? I don't think so. "not caring" is uncommon … but it's influence is doubly confounding to the result, too.

"dressed" is decidedly a term used more for male gender names; "clothes" for female, and about the same frequency for both. What to make of that?

So much, so interesting, and in my case, so little familiarity with details of social science survey research. It would be helpful, for tenuring arguments, if this is done more exactly.

ben Post author
February 11, 2015 at 8:56 pm

The “uses per million words” is normalized by gender, discipline, and length of the reviews. (I also have the number for “percentages of reviews,” but I don’t display it for the reason you note–that reviews of men might be longer.) The increased usage of “is” in reviews of women is a “real” phenomenon, just as is the increased usage of “the” in reviews of men. I suspect, as I say above, that this is due to known corpus linguistic effects where some words are used more by male speakers and others used more by female speakers, and that there are relatively more women reviewing female professors.

As for social science survey research: this differs from the classic sets. Unlike, say, the American Community Survey (which I’ve worked with in the past) there is no larger population being randomly sampled from here; instead, we have a full population known to be different from the group we might like to compare it against. (In this case, all college students or all college evaluations.) I think it is safest to make conclusions about this particular site.

eric palmer
February 11, 2015 at 9:06 pm

Thanks for taking the time — this is helpful.

John
February 12, 2015 at 1:44 pm

Hi Ben, you mention that: “The “uses per million words” is normalized by gender, discipline, and length of the reviews.” I am just wondering is this done automatically by the underlying Bookwork engine somehow or did you pre-process the text to normalize it before feeding to Bookworm?

ben Post author
February 12, 2015 at 2:40 pm

It’s done automatically; for each query, the engine database calculates both how many total words are in each category (reviews of female education teachers, reviews of male music teachers), and how many times the search term(s) are used in each group. Then it divides the second list by the first, so you get a ratio, and finally, I multiply by a million because otherwise the numbers are difficult to read (ie, they are preceded by 4 or 5 zeroes.

Amanda
February 16, 2015 at 5:00 pm

I'm sorry if this is a repeat comment, but I find this work very interesting. I'm unfortunately too impatient to read all the comments.

First I think this is a great step, and I like that you used frequency per million to account for differences in number of professors per field, and distribution of men and women in each field. I think one thing I would like to see would be rather than words per million using a binomial distribution to estimate the probability that a the word appears in a review for a given field and given gender. This might correct for issues with reviewers in each field writing longer reviews than others (e.g. English majors vs. engineering majors). Also, it would reduce impact of repeated use of a word (whether you think that's a good thing or not is up to you).

Additionally I think it would be really cool to look at co-occurences of words within a profile. For example I would expect that "smart" and "attractive" and more likely to occur in a male professors profile than a female professor, considering there is a quite a bit of research showing that perceptions of attractiveness are negatively correlated with perceptions of competency for women.

Again, very cool work!

ben Post author
February 18, 2015 at 10:34 pm

There are a lot of possible metrics–I agree that some sort of probabilistic calculation of how likely a word is to appear in the vocabulary of a reviewer might be somewhat useful if it seemed useful to reduce the impact of repeated use. But to express it as the chance that word appears in a *review* would actually be problematic, because it would actually create the problem you identify of varying review length. (Currently that's not an issue–normalizing by million words of texts means it doesn't matter if English majors tend to write novels. The only issue is if certain words are more likely to appear in long reviews than short ones–it might be that reviewers only get around to certain subtleties if they're in it for more than 100 words. This is something I may follow up on.

That's a great point on negatively disposed co-occurrence. I'll file it away for future investigation. It's unfortunately difficult to do in the current version of the software, but should be restored at some point in the future.

Ian
February 23, 2015 at 4:32 pm

Is there a word frequency list for these reviews? I'm interested in the different uses of certain words like "strict" and synonyms like "bossy" and "controlling". I'm also interested in the frequency of words that suggest the emotions or feelings of faculty. While "mean" shows up more frequently for female faculty so do words like "happy" and "caring." This may suggest that students perceive that their male professors are more emotionally neutral or disengaged than their female counterparts regardless of the emotion.

---

ben Post author
February 23, 2015 at 5:09 pm

What sort of list would be useful? I think most of this should be accessible through the chart.

Cathy Young
February 25, 2015 at 7:08 am

Hi Ben,

Fascinating stuff! I think it's being reported in a very simplistic way — while the results do indicate likely gender bias, it's a much more complex picture than the one being presented. I'm probably going to write about this.

Quick question (just one, for now). The "science" and "humanities" categories are aggregates for other fields on the list, right? Are all the other fields included in either one or the other category, or are some not included? (I'm not sure education counts as part of the humanities, for instance, and am unsure about the status of psychology and sociology.)

ben Post author
March 10, 2015 at 4:17 pm

Yes, I haven't gotten completely into the results yet. I think what is interesting about the size of the corpus is that it gives a slightly different way to talk about differences in gendered expression.

Sciences and humanities are not aggregate categories, actually; they are from schools where the students report professors as being just in one of those large aggregate fields, probably at something like a community college that doesn't have separate English and history departments.

Sarah Creel
February 26, 2015 at 8:02 pm

Ben, thanks for doing this! Is there a way to get overall frequency across the dataset for a particular word/word group, without getting the numbers in each of the 50 hovertexts and then averaging?

ben Post author
March 10, 2015 at 4:20 pm

Good question–I could build a quick site for this when I get back home, but at the moment there's no easy visual way to do it on this site. Averaging is probably not a good idea, because each of the samples here is a different size. You wouldn't want to weight the 25000 reviews of female engineering professors the same as the 250,000 (or whatever it is) reviews of male engineering professors.

rob hollander
March 9, 2015 at 10:40 pm

Great and important work!

The more frequent use of “genius” and “brilliant” for men is disturbing, but “bossy” is probably a red herring. Students use it for women because it’s the female-specific counterpart to “jerk,” “dick,” “schmuck,” “douche” and several others. “Jerk,” for example, is more weighted to men in the corpus than “bossy” is weighted to women in the corpus. And “jerk” is, from what I can tell from the chart, over thirty times more frequent in the corpus than “bossy.” So it doesn’t appear to be the case that students are more willing to accept authority from men — students object thirty times more to male authority than to women. Or maybe men in authority are more likely to be jerks, a hypothesis worth looking into. We grow up, after all, more privileged, pampered and so probably more narcissistic. That would explain both the “genius” and the “jerk” facts — being more focused in the classroom on ourselves, we are more likely to abuse our authority and inflate ourselves.