Information Inequality

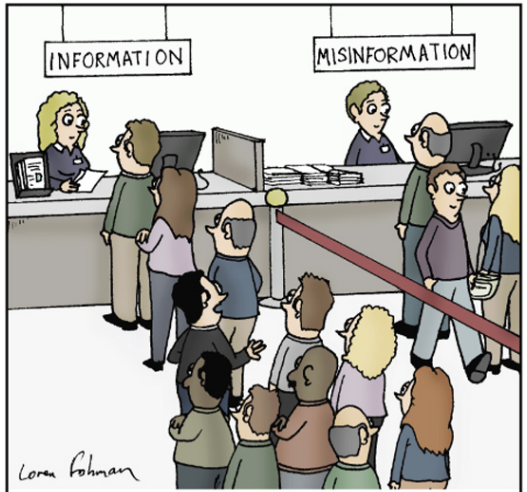
Gender Gaps in Knowledge

Molly M. King

Santa Clara University

August 11, 2019

1. Good afternoon, everyone! Thank you for being here.
2. My name is Molly King.
3. I am an Assistant Professor at Santa Clara University
4. Today I am going to be presenting a paper on gender disparities in factual knowledge.
5. I will be exploring these gaps using analyses of quantitative survey data
6. I will show that women's disadvantage in most areas of factual knowledge closely follows patterns of social norms, suggesting the power of socialization and gendered expectation in creating knowledge inequalities.
7. I will discuss how this reflects what knowledge is valued in our society and sheds light on important sites of the reproduction of inequality.

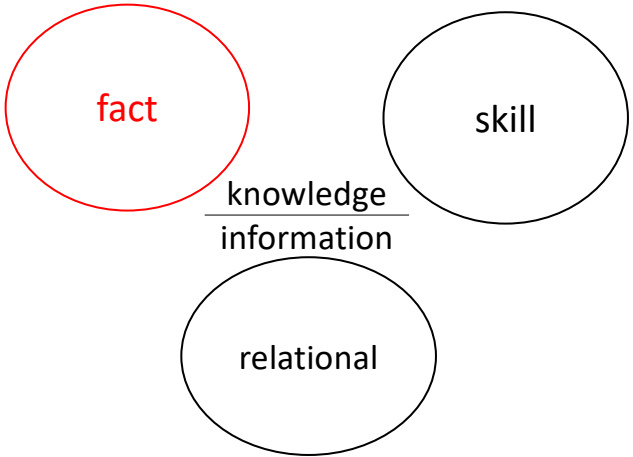


"Hey, that line's moving a lot faster!"

(Related theory: Hidalgo 2015, Simon 1971 | Cartoon credit: Loren Fishman | cartoonstock.com)

1. Never before has such a wealth of information been so immediately accessible to so many and yet the filtering demands so high. Topics many believe to be objective and resolved after years of scientific consensus are now resurfacing as topics of factual debate. What peopl know, what we think we know, and what we think others know all have dramatic consequences for the future of the U.S. and the world. Information truly is power; and who possesses it and wields it most effectively has profound consequences for inequality and human welfar
2. People disagree about facts, and this is shaping our public narrative more than ever.
3. But not only are we dealing with disagreements about facts. People als have differential access to the facts.

Types of Knowledge



So why should we care about inequalities in this?

1. Epistemology commonly distinguishes between three types of knowledge:
2. – knowledge of a person, or relational knowledge;
3. – knowledge of how to do something, or skills-based knowledge;
4. – and knowledge that some fact is true.
5. The category of factual knowledge, or “knowledge-that”, is what interest me here.

Why care about knowledge inequality?

- A virtue in itself

(Crisp 2015)

1. First, I argue that differences in information itself are, by definition, a dimension of 'inequality.'
2. Much like many people care about health or education as goods in and of themselves, not just for the ends they can help achieve, we may also think that knowledge is of value for its own sake, or for its ability to help live the good life.
3. Here I want to note that I do NOT mean to equate knowledge with intelligence. Knowledge of facts is only one aspect of knowledge, after all, and IQ measures something more akin to analytic abilities.

Why care about knowledge inequality?

- A virtue in itself
- Access to resources, outcomes, capabilities

(Weber [1922]1968, DiMaggio 1987, Ridgeway 2014, Lareau 2002, Lareau 2003, Sen 1992, Nussbaum 2011)

1. Second, information is a potential cause of later inequality in outcomes, resources, and capabilities.
2. Information discrepancies enable differential access to resources and institutional positions, thereby causing later inequality as well.
3. Furthermore, people need not only skills to navigate bureaucracies in order to get access to resources - they also need concrete knowledge about the types of resources they are trying to access.

Why care about knowledge inequality?

- A virtue in itself
- Access to resources, outcomes, capabilities
- Outcome of unequal social status

1. Finally, I argue that a concept I am calling knowledge inequality is important as an outcome of social inequality.
2. In my dissertation, I focus on this motivation – that social status causes knowledge inequality.
3. The field of sociology has long studied the production of knowledge; science inequalities in knowledge careers; and information diffusion and its consequences. Many studies have evaluated information seeking behaviors and needs. But the tendency has either been to study knowledge in one specific domain (e.g., health) or to reduce knowledge across all domains to a single test score – and hence we know shockingly little about the everyday knowledge stock of Americans.

Research Questions

Is there a gender gap in knowledge?

Does the gender gap in knowledge vary by domain?

What social process(es) explain(s) this gap?

1. So I wanted to perform a wide scan analysis of knowledge inequality, first finding out the answer to the most fundamental question: is there a gender gap in knowledge in different domains?
2. Then I wanted to look at who has and does not have knowledge in different domains, and how those inequalities might compare to each other.
3. I also sought to explore what social processes might explain the variance in knowledge by domain.

Datasets Gathered

Data Source	No. Datasets	No. Questions
Chicago Survey of Amer. Public Opinion	1	2
General Social Survey	5	40
Global Views American Public Opinion	1	2
Health Information National Trends Survey	7	125
Integrated Health Interview Series	1	12
Kaiser Family Foundation	1	7
National Financial Capability Studies	3	16
National Politics Study	1	5
Outlook on Life Survey	1	6
Pew Research Center	20	222
Rand American Life Panel	2	24
USC's Understanding America Study	3	48
21st Century Americanism survey	1	4
Total	48	513

1. in order to explore these questions, I collected data from 48 nationally representative data sets from between the years 2005 and 2015, each including at least one knowledge question.
2. I collected over 500 questions from public survey repositories, like the General Social Survey, Pew Research Center, and ICPSR.
3. These are true/false, yes/no, or multiple-choice questions that asked things like: 'Antibiotics kill viruses as well as bacteria. Is that true or false?" The answer is false.
4. I am happy to talk more about my data or methods in Q & A if anyone is interested.

Knowledge Domains (Categories)

biological science
current events
domestic politics
economics
finance
foreign politics
geography
health
language
math
natural world
physical science
religion
social science
technology
war

1. For each question, I marked for each individual whether they got the question correct or incorrect.
2. I curated these data and categorized them by domain.
3. This resulted in 16 topical domains.
4. Now of course, people have studied knowledge gaps in many of these domains before. What is unique to my study is gathering these domains all together in one comparative framework, allowing me to look at the structured acquisition of knowledge.

Logistic Regression (x 513)

Factors
Gender
Income
Race / Ethnicity
Age + age^2
Education

$=$

Outcome
Probability question correct

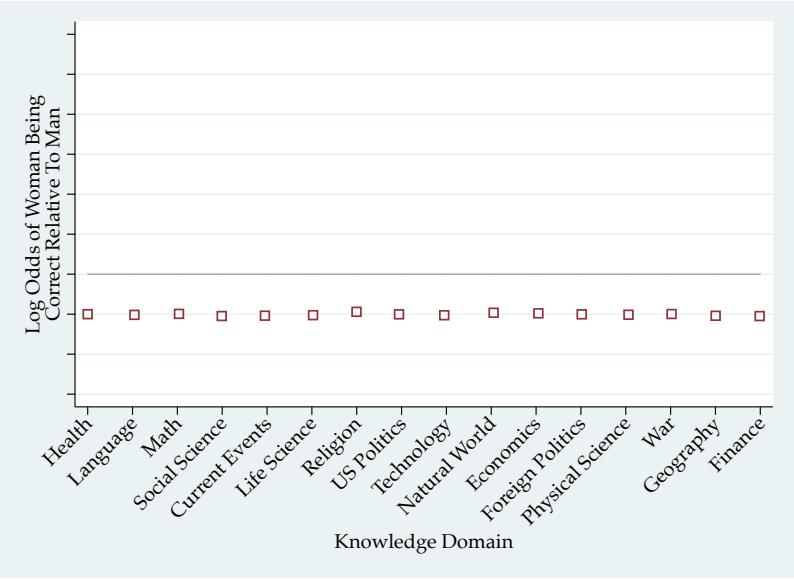
1. I also gathered many demographic characteristics about the individuals answering these factual knowledge questions.
2. For each question, I then use logistic regression to predict the probabilit of getting the question correct. I run this analysis for each of my 513 questions.
3. I also correct for multiple comparisons, meaning I can be very confident my results would replicate using another sample.
4. I'm happy to discuss any details in the Q&A.
5. In order to develop hypotheses about how my results might turn out...

Hypothesis: Question Maker Bias

(Photo credit: Wellcome Images)

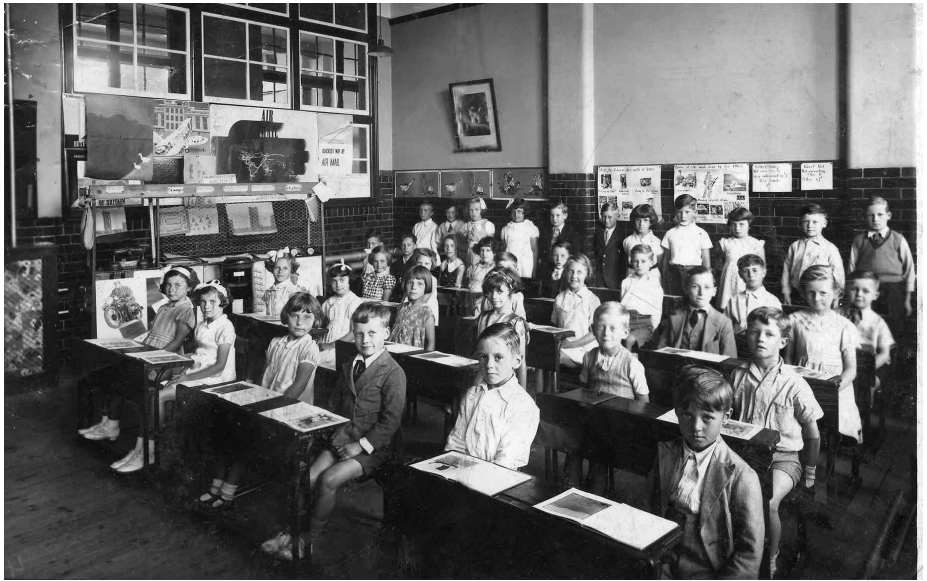
1. I first consider how the questions themselves are produced.
2. In a way, “everything is information”, so it can be very difficult to narrow what should be put into surveys to assess the knowledge base of the populace.
3. Politics, power, professions, personal relationships, and historical contingency all play a role in determining which knowledge items warrant special rewards from society. Some of this bias will inevitably creep into surveys.
4. Even though current advisory boards are much more gender balanced than my image here, many of the items still being asked are holdovers from an earlier time.

Hypothesis: Question Maker Bias



1. This leads me to my first hypothesis: the question maker bias hypothesis.
2. Under this, questions in all domains are biased to be more likely to be answered correctly by men - or to be answered by men at all - because they are more likely to be written by men in the first place.
3. This is indicated here with predicted log odds in favor of men knowing more. The log odds of a woman knowing the correct answer relative to man is presented along the y axis. Under this hypothesis, men would b expected to know more in all domains.
4. In other words, if the squares are under the line, men would know more if they are above the line, women would know more. This is the first hypothesis.

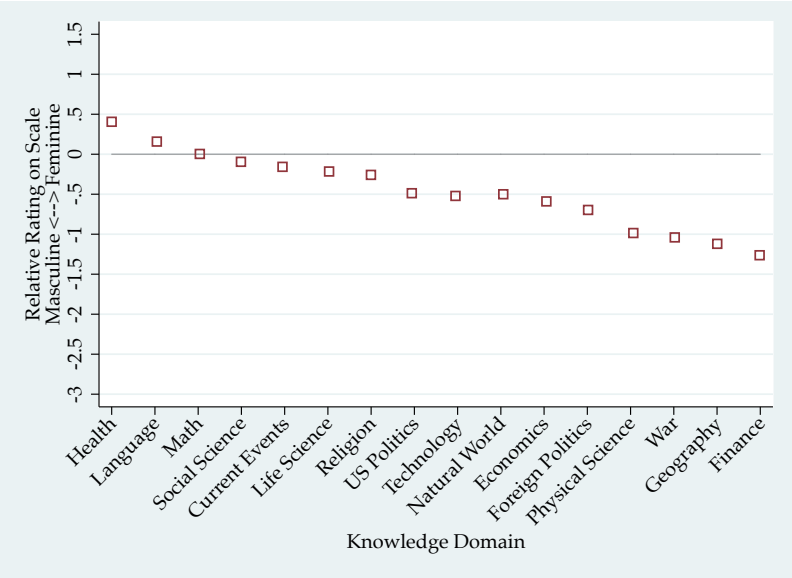
Hypothesis: Socialized Essentialism



(Related theory: Ridgeway 2006, West 1987, Cech 2013, Charles 2009 | Photo credit: Medium.com)

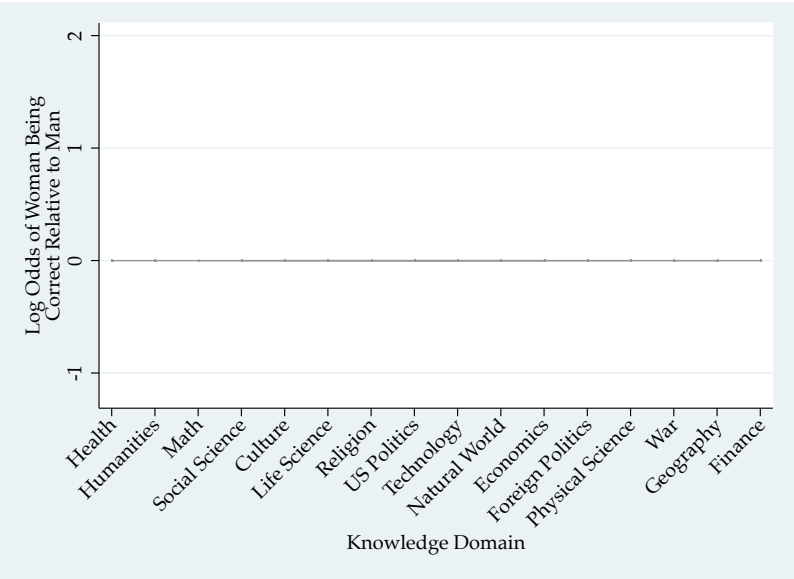
1. Alternatively, socialization may cause specialized knowledge by gender
2. Gender may have a direct impimpact on the types of knowledge acquired via education, occupations, and social networks.
3. Gender may also have an indirect impact via social norms, gaps in leisure time, or other socially shaped expectations of gendered knowledge interests.

Hypothesis: Socialized Essentialism



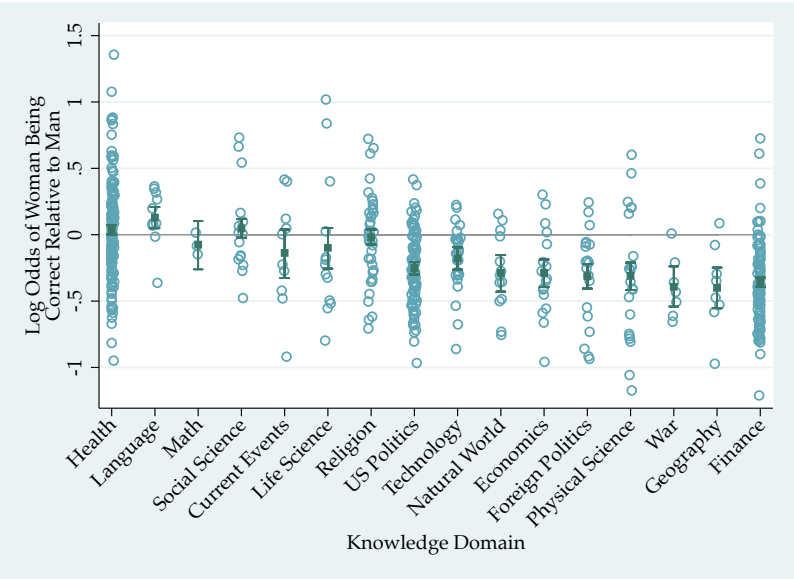
1. These interrelated mechanisms lead me to the Socialized Essentialism hypothesis.
2. Under this hypothesis, men and women specialize in knowledge domains.
3. But how do we know what gendered specialization in knowledge look like?
4. My estimates for the relative specialization of men and women in different domains in this hypothesis come from a survey I fielded asking about respondents' perceptions of each knowledge question.
5. I don't have time to discuss the details of this survey here, but I am happy to talk more about it in Q&A.

Domains Across X-Axis



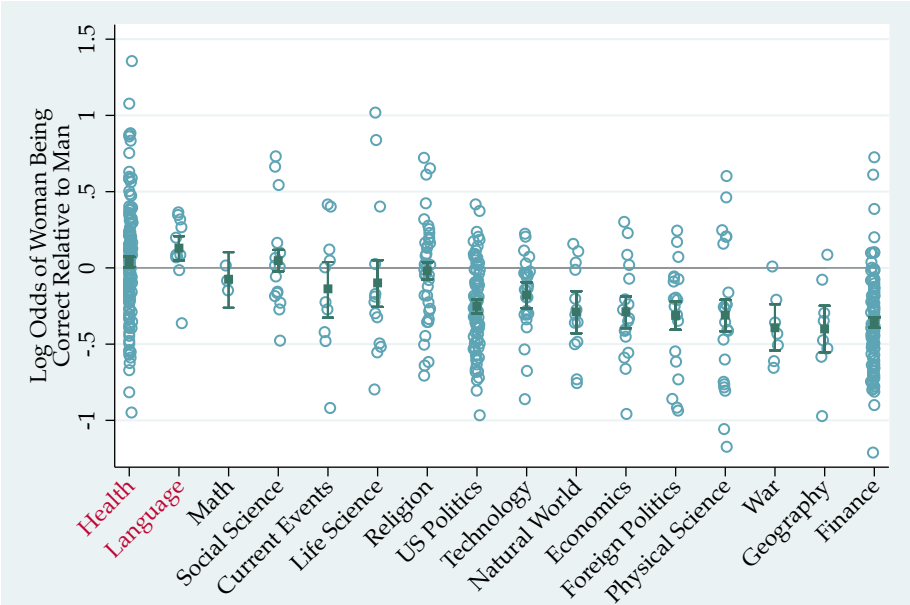
1. So here I take the 500 questions and divide them into these 16 domains across the x-axis.
2. The domains are lined up on the x-axis according to their perceived gender from the survey I fielded, with the most 'feminine' domains on your left and the most 'masculine' domains on your right.
3. I tested whether gender had a significant effect on knowledge within each entire domain.
4. If there were no difference in the average difference between men and women, we would see the confidence interval bar across the 0 line.

Each Circle is a Question



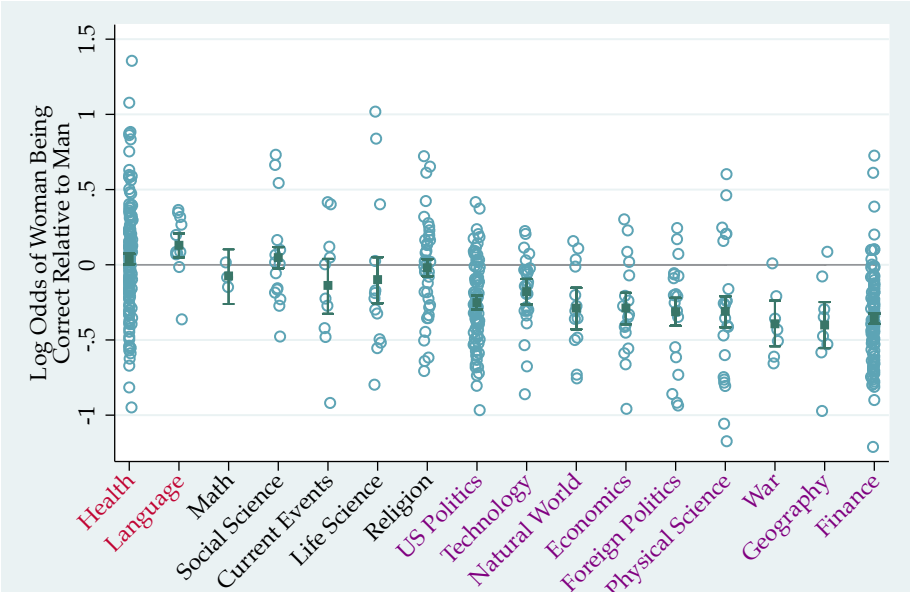
1. Each blue circle represents an individual question.
2. The green square is the mean knowledge level within each domain, and the confidence interval represents 95% simulated certainty around that mean for each domain.
3. For each domain, the simulated mean and confidence intervals show whether there is a significant difference between the 2 gender groups and the direction of that difference.

No mean gender difference in 5/16 domains



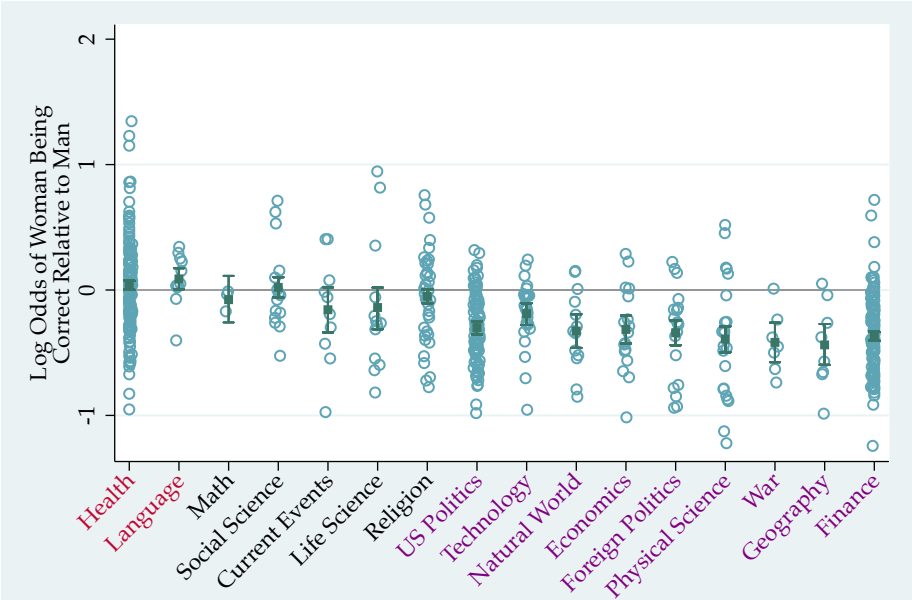
1. Women have greater average knowledge in the domains of health and language.
2. There is no average gender difference in 5 of the 16 domains – specifically math, social science, current events, life science, and religio – because the confidence interval crosses the 0 line.

Men have greater knowledge in 9/16 domains



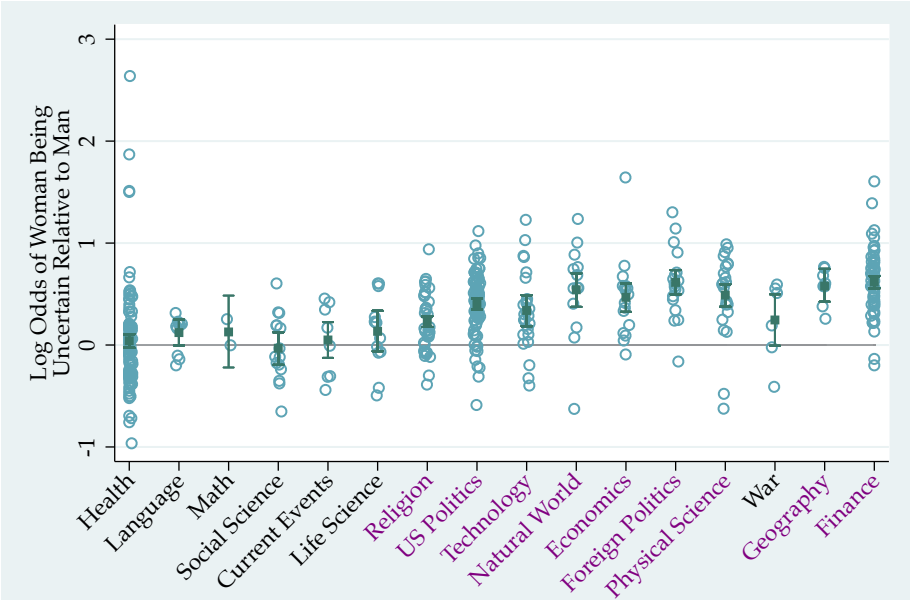
1. And men have greater average knowledge in 9 domains.
2. These are all domains where the questions were rated more masculine by respondents in my survey where I asked people to rate the gendered nature of these questions.

Result holds when controlling for education



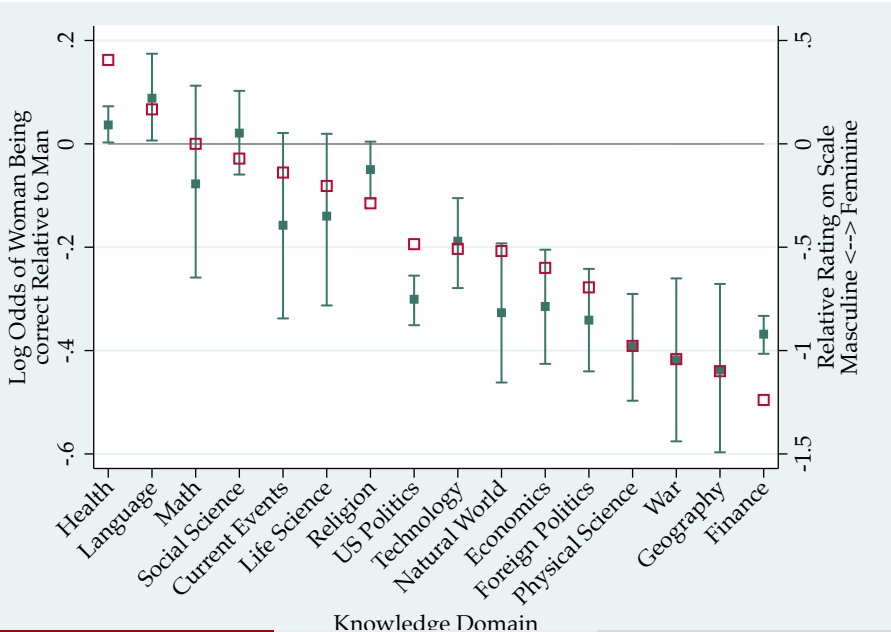
1. You might think this is driven by differences in education.
2. However, controlling for education increases the number of questions that men answer correctly significantly more often and decreases the number of questions that women answer correctly. This is because women are getting more of their knowledge advantage from their educational advantage relative to men - especially in the sciences.

Women answer more questions “don’t know”



1. One possible explanation for this trend is that women are more likely to be uncertain about these questions that are viewed as more ‘masculine
2. The same knowledge domains where men are more likely to answer correct on average are the domains where women are more likely to answer “don’t know”.
3. However, this does not tell us if this uncertainty gap results because women truly are less likely to know the information or are less willing to guess.

Results: Socialized Essentialism



1. Revisiting my original theory, I found my results do not match the question maker bias hypothesis.
2. I find my results do closely match the socialized essentialism hypothesis
3. Remember, the socialized essentialism hypothesis – represented here by red open squares – predicts that the respondents on the nationally representative surveys with the factual knowledge questions will perform in correlation to the gender each item is perceived to be. In other words, women answer more questions incorrectly in those domains that are perceived to be more "masculine," and vice versa.
4. I do find high degree of correlation between what people think women and men are likely to know more about and the actual factual knowledg respondents report on nationally representative surveys.
5. Remember, my classifications into domains had nothing to do with thes gendered ratings of the questions themselves. So the respondents' performance on the knowledge items and the ratings of how gendered each item is perceived to be are as independent as possible.
6. This is strong evidence that gendered socialization mechanisms are at work here in determining the domains of knowledge men and women specialize in.
7. In this way, our expectations shape our reality. Social expectations and socialization shape the knowledge acquired by men and women, with dramatic consequences for inequality. I'm happy to talk more about

Thank You

Molly M. King
mmking@scu.edu

1.

Supplemental Slides

- Theory: Status Differences

Methods

- Data Search & Selection
- Data Sources
- Data Structure
- Logit Details
- Multiple Comparisons
- Confidence Interval Simulations
- My Survey on Gendered Perception of Knowledge Questions
- Model for Don't Know Responses by Gender
- Adjusting for the Guessing Gap

Results

- Mean significance vs. Question significance (3 slides)
- Proportion of Don't Know Responses by Gender
- Proportion of Population Correct by Domain
- Proportion of Population Don't Know by Domain

Other Research

- Gender & Self Citation: Proportions
- Gender & Self Citation: Ratio

1.

Information & Status Differences

(Back)

1. Weber's three interrelated bases for inequalities in modern industrial societies are resources, power and status. Status differences between individuals may be created by virtue of information differences. Social status is defined as inequality grounded in differences in respect, esteem, and honor. "People use culture to make connections with one another"; information that is important to one's social contacts may lead those important others to publicly acknowledge that person and contribute to their sense of beir valued. When aggregated at the group level, this control of information may be "transformed into more essentialized differences among 'types' of people, status beliefs that fuel social perceptions of difference." At the macro-level, I argue these differences in information, originally created on the basis of gender, give a force to status beliefs that ca reproduce material inequalities independent of initial differences in power or resources.

Data Search & Selection

Databases searched:

- ICPSR – 4,581 surveys reviewed
- Data.gov – 1,117 surveys reviewed

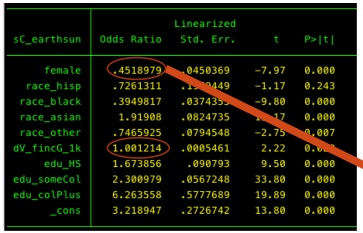
Inclusion criteria:

- survey search includes the term “knowledge”
- years 2005 – 2015
- U.S. nationally representative on race, gender, age

(Back)

1. Additional surveys were identified by searching ICPSR using the term “knowledge,” limiting the search to the time period 2005 to 2015 in the United States. All (4,581) surveys returned from the search were assessed for relevance and inclusion in the study. I performed the same search for the term “knowledge” on the Data.gov database. Criteria for selection included whether the surveys were generally nationally representative and were limited to the time period 2005 to 2015. I reviewed the resultant 1,117 surveys from Data.gov for relevance.
2. This returned several additional surveys:
3. Annenberg National Health Communication Survey;
4. the National Financial Capability Studies
5. the 21st Century Americanism survey
6. the Global Views American Public Opinion and Foreign Policy Survey;
7. the Outlook on Life Survey;
8. the State of the First Amendment surveys;
9. and the Chicago Council Survey of American Public Opinion and U.S. Foreign Policy.

Data Structure



sC_earthsun	Odds Ratio	Linearized Std. Err.	t	P>|t|
female	.4518979	.0450369	-7.97	0.000
race_hisp	.7261311	.1[illegible]0449	-1.17	0.243
race_black	.3949817	.0374335	-9.80	0.000
race_asian	1.91908	.0824735	[illegible]17	0.000
race_other	.7465925	.0794548	-2.75	0.007
dV_fincG_1k	1.001214	.0005461	2.22	0.0[illegible]
edu_HS	1.673856	.090793	9.50	0.000
edu_someCol	2.300979	.0567248	33.80	0.000
edu_colPlus	6.263558	.5777689	19.89	0.000
_cons	3.218947	.2726742	13.80	0.000

DOMAIN	Literature		Science		
STATUS	L1	L2	S1	S2	S3
Gender (G)	GL1	GL2	GS1	GS2	GS3
Class (C)	CL1	CL2	CS1	CS2	CS3
Race (R)	RL1	RL2	RS1	RS2	RS3

(Back)

Logit Details

Factors
Gender
Income
Race / Ethnicity
Age + age^2
Education

=

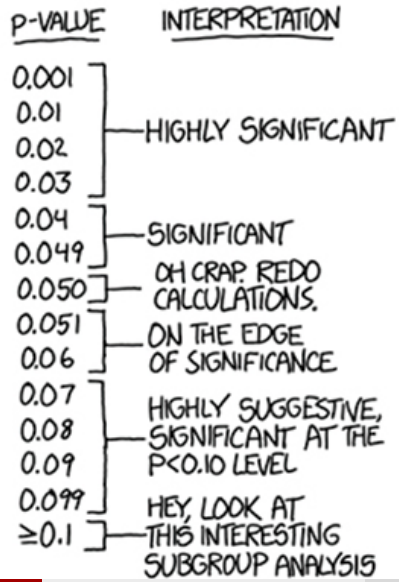
x 513

Outcome
Probability that you get the question correct

(Back)

1. I regressed the outcome of correct answer on the independent variables income, gender, race and ethnicity categories, and controls education categories, and age and age squared.
2. I regressed the outcome of correct answer on the independent variables gender and control variables.
3. I run this analysis for each of my 513 questions.
4. Because I am dealing with over 500 separate logistic regressions, I also adjust for multiple comparisons.
5. I mention this because the implication is that I can be extremely confident that these results would replicate using a different sample.
6. Finally, I created 95% confidence intervals using simulation methods, resampling from the standard error of each coefficient and averaging those across each domain.
7. Although the ideal measure of class here would have been parental occupation, I had to use respondents' family or household income because parental education was not available in most surveys.

Adjusting for Multiple Comparisons



1. Because I am dealing with over 400 separate regressions, I also adjust for multiple comparisons using the Holland method.
2. I mention this because the implication is that I can be extremely confident that these results would replicate using a different sample.

Confidence Interval Simulations

Draw 1000 repetitions from a standard normal distribution, generate a distribution of 1000 possible coefficients for each knowledge question:

$$\hat{\beta}_i = \beta + (SE * x_i),$$

where $x_i \sim N(0, 1)$.

Order all $\hat{\beta}_i$ such that

$$\hat{\beta}_i \leq \hat{\beta_{i+1}}$$

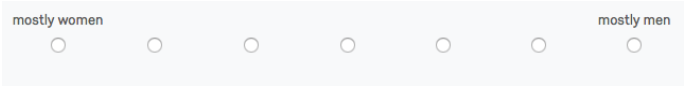
Find value of $\hat{\beta}_i$ at the 2.5th percentile and the 97.5th percentile.

(Back)

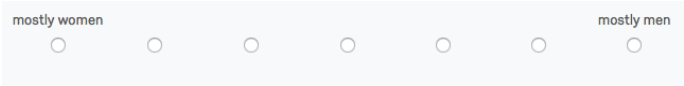
1. I take each coefficient from each of my logistic regressions, and "resample" from a possible distribution of these coefficients adjusted by each coefficient's standard error multiplied by a random draw from a normal distribution.
2. Drawing 1000 repetitions from a standard normal distribution, I generate a distribution of 1000 possible coefficients for each knowledge question. In other words, I repeat this 1000 times for each of the knowledge questions (each of the 513 logistic regressions).
3. Then, within each knowledge domain, I line these resampled simulated coefficients up in order and use the 2.5th percentile and the 97.5th percentile as my upper and lower confidence bounds for that particular knowledge domain. These values are the lower and upper bounds of th 95 percent confidence interval for each information domain within each demographic subgroup.

My Survey: Gendered Perception of Knowledge Questions

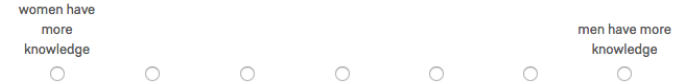
- Who mostly watches news about X?



- Who mostly talks to their friends about X?



- Who do most people in our society think has more knowledge about X?



(Back)

1. To get these ratings, I fielded a survey using these 3 items, where X wa a shortened version of each knowledge question, like "who mostly talks to their friends about Antibiotics?"
2. I used these questions to create a scale representing the socially perceived masculine or feminine nature of a given knowledge question.
3. I then use this perceived gender of each item to develop an overall scal for each domain.

Adjusting for the Guessing Gap

Upper Bound:

$$ceiling_W = C_W + (\frac{C_W}{C_W + W_W})D_W$$

which simplifies to

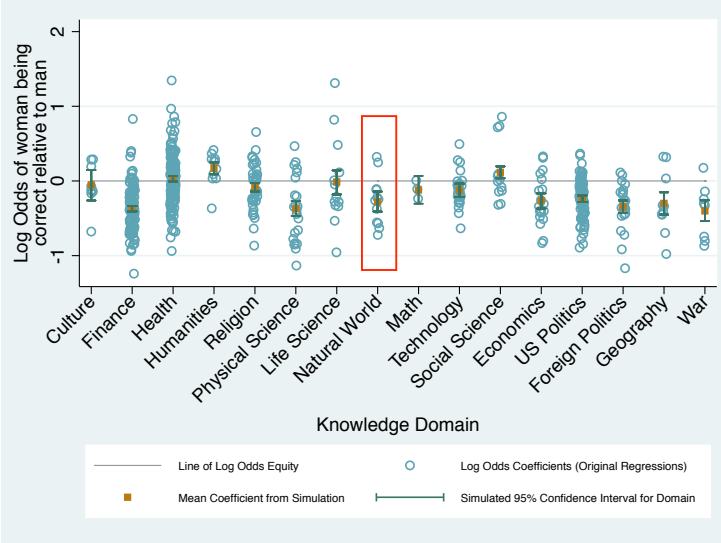
$$= C_W + (C_W * D_W).$$

Lower Bound:

$$floor_W = C_W + (\frac{1}{K_Q})D_W$$

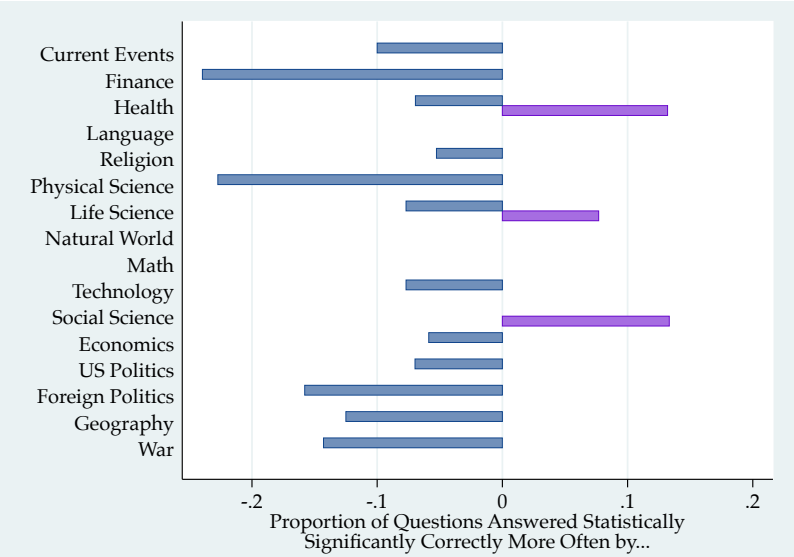
(Back)

Mean significance vs. Question significance



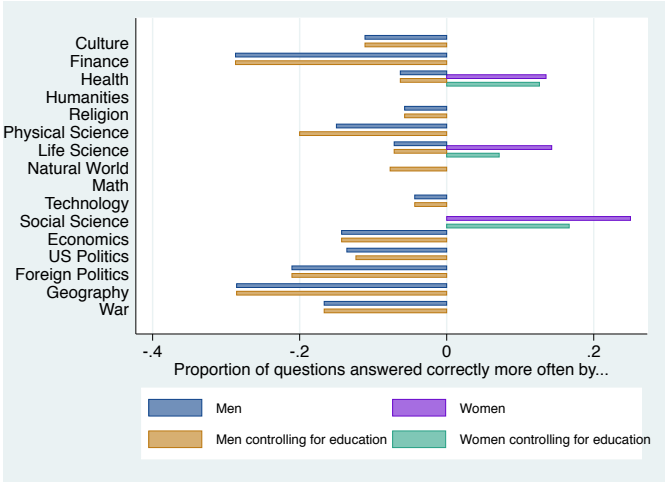
1. Note here how in the natural world domain women have .8 odds smalle of knowing the correct answer in the domain, on average.
2. But this first figure does not indicate how many of these questions show knowledge differences that are statistically significant.
3. So for each of these 500 questions, I tested whether the gender difference in knowledge was significant.

Men answer greater proportion of questions correctly in 65% of domains



1. However, not all of those coefficients are statistically significant. In this figure, each bar represents the proportion of all questions in a given domain that men or women were more likely to answer correctly such that these differences were statistically significant.
2. So in the example of the natural world, the average gap between men and women across all items in the natural world domain is significantly different from 0, but there was no single question where men or women were more likely to be correct.
3. This can also go the other way, in theory, where the average gap is not significantly different from 0 but individual questions are. Though there are no examples of this here, social science does come close.
4. I also find that women answer a greater proportion of questions correctl in the domain of health, though this is a bit more complicated. Both me and women have some questions that they are significantly more likely than the other group to answer correctly, so there are proportions in bot directions on the graph.
5. One reason we might consider this to be notable is that men's poor health outcomes are typically explained by behavioral differences. Here differences in health outcomes might also be explained by a disparity in health knowledge.
6. You might assume this is driven by gender differences and education.

Controlling for education, men answer greater proportion of questions correctly in 63% domains



(Back)

1. But overall the trend holds.
2. Men are particularly advantaged in the domains of finance, foreign politics, and geography.
3. Overall, Men answer a greater proportion of questions statistically significantly correctly in 3/5ths of the domains. Women answer a greate proportion of questions correctly in 1/5th of the domains.
4. Essentially, this residual effect of gender cannot be explained by differences in education level or income.

“Don’t Know” Model

Factors
Gender
Income
Race / Ethnicity
Age + age^2
(Education)

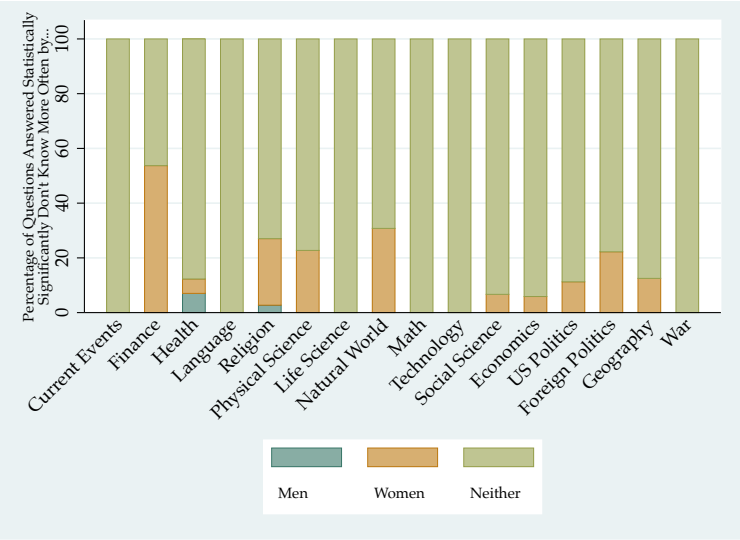
=

Outcome
Probability that you answer “don’t know”

(Back)

1. In a slightly different model, for each question, I then use logistic regression to predict the probability of answering ‘don’t know’ for all questions where that was an option.
2. I regressed the outcome of uncertainty on the independent variables gender and control variables.
3. I have also estimated linear probability models and the results are much the same.

Controlling for education, women answer greater proportion of questions “don’t know”



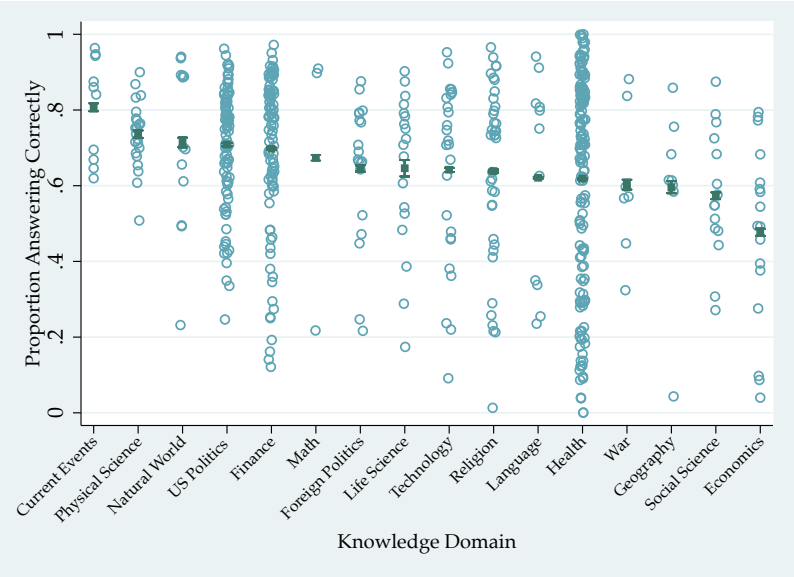
1. In all but 16 domains, women are more likely, on average, to answer don’t know across all questions as well, even after controlling for education.
2. However, I do not yet know if this is because women truly are less likely to know the information or are less willing to guess.
3. I am currently working on a process to estimate each group’s certainty threshold. I can talk this in q&a if you are interested.

Research Questions

How much factual knowledge do U.S. adults know and say they ‘don’t know’?

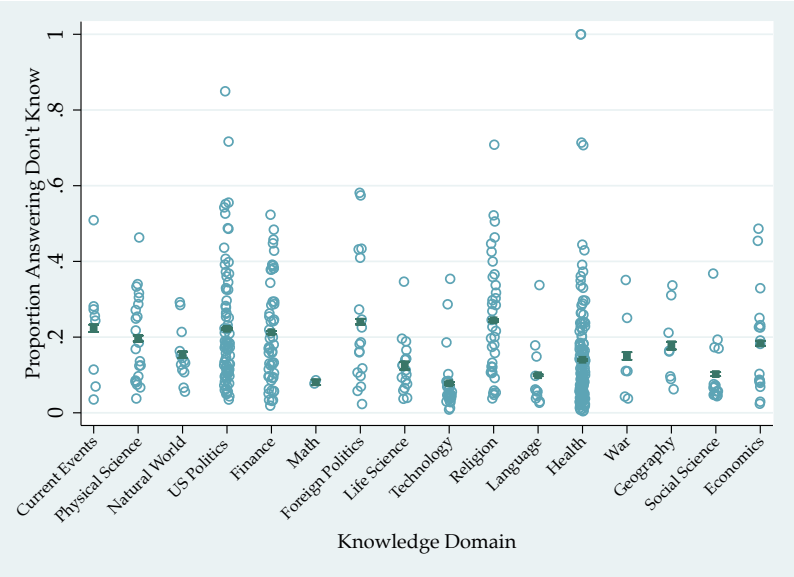
1. My first paper looks at the factual knowledge landscape of U.S. adults.

Proportion of population answering correctly



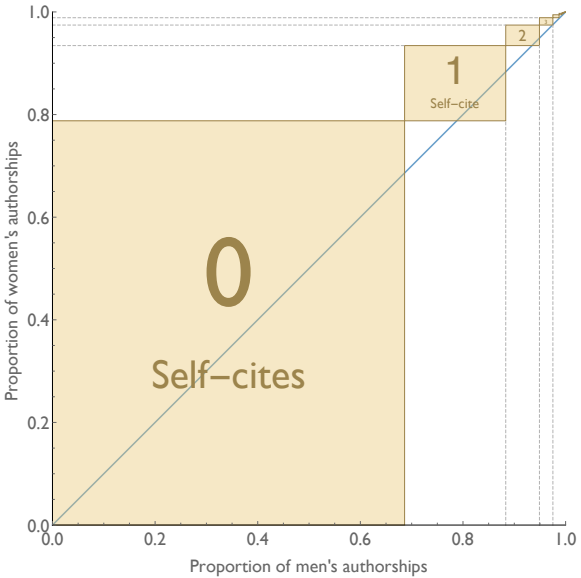
1.

Proportion of population answering ‘don’t know’



1.

Gender & Self Citation: Proportion with Self Citation



1. shows members of self-citations grouped by proportions of menâĂŹs and womenâĂŹs authorships. We show the proportion of men with a certain number of self-citations on the x-axis and the corresponding proportion of women on the y-axis. If men and woman behaved similarl in their approaches to self-citation, the corners of the boxes would trace the x-y diagonal. Instead, wherever there is a difference in the proportio of men and women citing themselves a certain number of times, the corners of the boxes deviate from the diagonal.
2. relative to menâĂŹs authorships, womenâĂŹs authorships are more likely to feature zero self-citations. Women cite themselves one or more times in their papers less often than men do. In other words, compared with men, women are overrepresented in the zero self-citations categor and underrepresented in terms of citing their papers at all. For example if in a paper you never cite another paper of your own, you are among the majority of men (68.6 percent) and women (78.8 percent) who do no cite themselves.
3. whenever a box is wider than it is tall, there is a greater proportion of men authorships in that category of self-citations. If you have one self-citation, you are in the 68th to 88th percentile range for men (representing 20 percent of menâĂŹs authorships) but the 78th to 93rd percentile for women (representing only 15 percent of womenâĂŹs authorships). With four self-citations in a single paper, a woman is in th

Gender & Self Citation: Ratio

1. shows the self-citation ratio for each year. If men and women cited themselves at equal rates, the ratio shown would be 1.0. A value of 1.5 means that men cite themselves 50 percent more than women in paper published during that year. Shaded intervals represent 95 percent bootstrap confidence limits.
2. In the 1950s, the relative rate15 of menâĂŹs self-citations relative to womenâĂŹs self-citations was 1.23. However, during the 1950s, the bootstrapped 95 percent confidence intervals of the annual ratios overlap with an equality ratio of 1.0, indicating that we cannot reject the null hypothesis of gender equality in self-citation rate during this decade. However, beginning in the 1960s, the ratio of menâĂŹs to womenâĂŹs self-citations per authorship remains steadily significantly above 1.0. In the 2000s, the relative rate was 1.71. There is no evidence that that the gender gap is decreasing over time.