

# UC Berkeley

## UC Berkeley Electronic Theses and Dissertations

### Title

The ABC's of Mathematics Perceptions

### Permalink

<https://escholarship.org/uc/item/3wv9t6wd>

### Author

Jansen, Rachel

### Publication Date

2021

Peer reviewed|Thesis/dissertation

The ABC's of Mathematics Perceptions

by

Rachel Jansen

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anna Rafferty, Co-chair  
Professor Mahesh Srinivasan, Co-chair  
Professor Tom Griffiths  
Professor Marcia Linn  
Professor Steve Piantadosi

Fall 2021

The ABC's of Mathematics Perceptions

Copyright 2021  
by  
Rachel Jansen

Abstract

The ABC's of Mathematics Perceptions

by

Rachel Jansen

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Anna Rafferty, Co-chair

Professor Mahesh Srinivasan, Co-chair

What does the word “math” evoke? It is in many ways a fraught term eliciting negative reactions and unpleasant memories. In this dissertation, I explore the variety of ways we humans perceive this concept, starting with the internal (perceptions of *ability*), then the personal journey (perceptions of *belonging*), and finally the external (perceptions of *conception*). I employ a mixture of experiments and computational modeling in order to develop a more holistic understanding of how people perceive math and reinforce human studies with data collected from naturalistic settings - specifically, the internet, to explore word usage and how discussions of math seem to differ from mentions of other concepts.

To Dad

For showing me how to wonder at math.

# Contents

<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>viii</b>
<b>Acknowledgements</b>	<b>x</b>
<b>-1</b>	<b>1</b>
<b>0</b>	<b>5</b>
0.1 Introduction . . . . .	6
0.2 Methods . . . . .	7
0.3 Results . . . . .	9
0.4 Discussion . . . . .	11
0.5 Conclusion . . . . .	12
<b>1</b>	<b>13</b>
1.1 Self-Assessment . . . . .	13
1.2 Modeling self-assessment . . . . .	15
1.3 Testing the model: The impact of immediate feedback . . . . .	20
1.4 The Dunning-Kruger effect . . . . .	24
1.5 Sequential self-assessment . . . . .	33
1.6 General discussion . . . . .	47
<b>2</b>	<b>49</b>
2.1 Introduction . . . . .	49
2.2 Related literature . . . . .	51
2.3 Study 1: Math narratives in a large sample . . . . .	53
2.4 Study 2: Math attitudes across time . . . . .	57
2.5 Discussion . . . . .	76
2.6 Conclusion . . . . .	78
<b>3</b>	<b>79</b>
3.1 Introduction . . . . .	79

3.2	Stimuli design . . . . .	81
3.3	Study 1: Math conceptions of adults and children . . . . .	82
3.4	Study 2: Intervention on middle school students in India . . . . .	92
3.5	Study 3: Intervention on adults in a science museum . . . . .	97
3.6	General discussion . . . . .	101
<b>4</b>		<b>102</b>
4.1	Research into practice . . . . .	103
4.2	Final thoughts . . . . .	104
	<b>Bibliography</b>	<b>106</b>
<b>A</b>		<b>115</b>
A.1	Participants . . . . .	115
A.2	Hollingshead (SES) . . . . .	116
A.3	ANOVA tables . . . . .	117
A.4	Mean ratings . . . . .	119

# List of Figures

0.1	Google trends patterns of searching for the topics “mathematics,” “language,” and “art.” . . . . .	9
1.1	Graphical representation of the model: each observed item $X_{p,i}$ is influenced by latent variables $\beta_i$ (difficulty of problem $i$ ) and $\theta_p$ (perceived ability of person $p$ ) as well as a constant $\epsilon$ (ability to determine correctness), where the observed $X_{p,i}$ refers to a person’s beliefs about correctness on an item. Difficulty $\beta_i$ is drawn from a normal distribution with mean $\mu_\beta$ and standard deviation $\sigma_\beta$ and ability $\theta_p$ is drawn from a normal distribution with mean $\mu_\theta$ and standard deviation $\sigma_\theta$ .	18
1.2	Model predictions in a toy example where participants solve 10 problems in the baseline model ( $\mu_\theta, \mu_\beta = 0, \sigma_\theta, \sigma_\beta = 1$ , and $\epsilon = 0$ ). Each point shows the average of the posterior distribution on $\theta$ , corresponding to estimated ability. Overlaid are histograms of MCMC estimates of the posterior distribution on $\theta$ for three scores. The identity line represents completely accurate estimation, where a participant’s estimated score is equal to their true score. The guessing parameter $g$ is equal to 0.2, which represents a task with five multiple-choice solutions per question. . .	20
1.3	Model predictions in a toy example where participants solve 10 problems (a) when the mean on ability ( $\theta_p$ ) is adjusted ( $\mu_\theta = 0.5$ or $-0.5$ ), (b) when the parameter $\epsilon$ is adjusted ( $\epsilon = 0.2$ or $0.4$ ), (c) demonstrating the Dunning-Kruger effect. Note that $x$ in the equations in panel (c) refer to normalized score (score divided by maximum score). Again, we set $g = 0.2$ . . . . .	21
1.4	Mean estimates of score (out of 20) by quartile of actual performance in (a) no feedback and (b) with feedback conditions of Study 1. Error bars show 95% confidence intervals. . . . .	23
1.5	Participants’ estimates of score (out of 20) by true score as compared to the best model predictions in (a) no feedback and (b) with feedback conditions of Experiment 1. . . . .	23
1.6	Histograms of scores achieved in (a) the grammar study and (b) the logical reasoning study. . . . .	29



- 1.7 Results and best fitting models for the grammar study displayed by (a) score and (b) quartile of performance. The model where estimation accuracy is independent of score, or the ‘Bayesian inference’ model is parameterized by  $\epsilon = 0.4$  and  $\mu_\theta = 0.05$  while the performance-dependent model is parameterized by  $\epsilon_0 = 0.45$ ,  $\mu_\theta = 0.05$ , and  $\alpha = 0.1$ . Error bars represent 95% confidence intervals. . . . . 30
- 1.8 Results and best fitting models for the logical reasoning study displayed (a) by true score and (b) by quartile of performance. The simpler Bayesian inference model is parameterized by  $\epsilon = 0.45$  and  $\mu_\theta = -0.1$  while the performance-dependent model has  $\epsilon_0 = 0.5$ ,  $\mu_\theta = -0.15$ , and  $\alpha = 0.15$ . Error bars represent 95% confidence intervals. . . . . 32
- 1.9 Graphical representation of the model: each observed item  $X_{p,t}$  is influenced by latent variables  $\beta_t$  (difficulty of problem at time  $t$ ) and  $\theta_{p,t}$  (perceived ability of person  $p$  at time  $t$ ) as well as a constant  $\epsilon$  (ability to determine correctness). The difficulties of all problems  $\beta_t$  and the prior over perceived ability  $\theta_{p,1}$  are drawn from normal distributions with means  $\mu_\beta$ ,  $\mu_\theta$  and standard deviations  $\sigma_\beta$ ,  $\sigma_\theta$ . Each subsequent  $\theta_{p,t>1}$  is drawn from a  $N(\theta_{p,t-1}, \sigma_{dyn})$ . . . . . 35
- 1.10 Model predictions in a toy example where participants solve 10 problems in a baseline model ( $\mu_\theta$ ,  $\mu_\beta = 0$ ,  $\sigma_\theta$ ,  $\sigma_\beta = 1$ , and  $\epsilon = 0$ ). Each point shows the weighted average of the posterior distribution on each  $\theta_t$ , corresponding to confidence judgments at each time point. The first judgment made before the first problem is simply based on the prior over ability. The shaded areas demarcate problems solved incorrectly, while the unshaded regions correspond to correctly solved problems. . . . . 37
- 1.11 Model predictions on the same toy example as in Figure 1.10 for (a) when the mean on ability ( $\theta_{pt}$ ) is adjusted ( $\mu_\theta = 1$  or  $-1$ ), (b) when the error parameter  $\epsilon$  is adjusted ( $\epsilon = 0.2$  or  $0.4$ ), and (c) when ( $\sigma_{dyn}$ ) is zero or nonzero ( $\sigma_{dyn} = 2$ ). In all three plots, the parameters we are not adjusting are equal to zero. . . . . 38
- 1.12 Model predictions for four example participants from the algebra multiple-choice task (Study 2) where shaded regions signify incorrect responses: (a) the best fit static model was parametrized by  $\epsilon = 0.25$  and  $\mu_\theta = 1$  and the best fit dynamic model had parameters  $\epsilon = 0.2$ ,  $\mu_\theta = 1$ , and  $\sigma_{dyn} = 0.6$ ; (b) was fit by  $\epsilon = 0.05$  and  $\mu_\theta = -1.7$  while the dynamic model is fit by  $\epsilon = 0.1$ ,  $\mu_\theta = -1.3$ , and  $\sigma_{dyn} = 2$ ; (c) the best fit static model for this participant was parametrized by  $\epsilon = 0.3$  and  $\mu_\theta = -0.3$  while the dynamic model is fit by  $\epsilon = 0.45$ ,  $\mu_\theta = -0.1$ , and  $\sigma_{dyn} = 1.6$ ; (d) was best fit by  $\epsilon = 0.25$  and  $\mu_\theta = -1.9$  for the static model and  $\epsilon = 0.35$ ,  $\mu_\theta = -2$ , and  $\sigma_{dyn} = 0.2$  for the dynamic model. The dynamic model fit the data significantly better than the static model in (a) and (b). In (c) and (d), the dynamic and static models fit the data equally well. . . . . 40
- 1.13 Model predictions on the same toy example as in Figure 1.11 with the same parameter values, but including a guessing parameter  $g = 0.25$ . . . . . 41

1.14	Histograms depicting number of participants in each study who were best fit by each possible $\mu_\theta$ , $\epsilon$ and $\sigma_{dyn}$ value for each study (College, Nobel, Algebra Multiple-choice, Fractions multiple-choice, Tutor Algebra, Tutor Fractions, India Math, and India Reading). Note that when $\sigma_{dyn} = 0$ this indicates that the static model fit the data better for that individual. . . . .	48
2.1	(a) Screenshot of the Quanta magazine survey and (b) mean attitude ratings (from 1, meaning “hate,” to 5, meaning “love”) by gender and time when attitudes were formed. Female respondents (green) gave significantly lower ratings for grades 6-8 and grades 9-12 than other groups. Error bars reflect one standard error. . . .	55
2.2	Distribution of 30 most frequent word roots in Study 1 for (a) men, (b) women, (c) those with positive feelings about math and (d) those with negative feelings about math. All plots exclude words with root “math.” One observation is that the top words for all groups include “teacher” and “school” but “beauti” and “love,” as expected, appear in all groups except the negative attitude group which instead includes “hate,” “feel,” and “never.” . . . .	56
2.3	Diagram depicting open-ended survey questions for each section of the survey along with associated computational text analyses. Using the first general written response, we ran four classifiers distinguishing word usage between groups (positive and negative attitudes, high and low math anxiety, men and women, high and low SES), and one topic model. To explore narratives about individual time periods, we ran four classifiers separating those with positive and negative attitudes over each pair of responses for each time period. We then ran one topic model over these responses. . . . .	62
2.4	Histograms of ratings of (a) general feelings about math ( $M = 5.54$ ), (b) feelings about doing math in school ( $M = 4.97$ ), and (c) feelings about doing math at work ( $M = 5.52$ ). . . . .	63
2.5	Differences in overall attitude ratings by gender and time when attitudes were reportedly formed. Error bars indicate standard deviations. . . . .	66
2.6	Frequency of 40 most frequent words from the general open-ended responses across all participants. Note that these words are generated after pre-processing of the text, including eliminating stopwords, stemming, and lemmatizing. . . . .	67
2.7	(a) Word clouds of top keywords for each topic. The size of the words is proportional to the weight and (b) topic weights over the entire set. Specifically, the sum of the actual weight contribution of each topic to respective documents. . .	71
2.8	Histogram of attitude ratings from (a) Grades 1–5 ( $M = 6.30$ ), (b) Grades 6–8 ( $M = 5.42$ ), (c) Grades 9–12 ( $M = 4.76$ ) and (d) college ( $M = 5.35$ ). . . . .	73
2.9	Word clouds of top keywords for each topic over the time-point-specific responses. The size of the words is proportional to their respective weights. These can be viewed alongside exact weight values in Table A.15 in the Appendix. . . . .	74

3.1	A screenshot of the math conception measure. Participants dragged each item into the “Math” box if they believed it could involve math. . . . .	81
3.2	Number of adult participants in Study 1 who labeled each activity as involving math (a) and in a replication with more items (b). . . . .	83
3.3	Plot of linear regression line showing relationship between breadth of conception and math anxiety in Study 1a, controlling for education ( $\alpha = 6.93$ , $\beta = -0.18$ , $p < .05$ ) and in a replication with the SIMA (b) and the AMARS (c) measures of math anxiety. . . . .	85
3.4	Number of participants in Study 1b who answered “yes” when asked whether each kid was doing math. . . . .	87
3.5	Example sorting tasks about math (a) and art (b). The items are separated into two clusters, one of which contains the activities participants believed involved math in the first round of sorting, while the second cluster includes activities the participant believed <i>could</i> involve math in a second round of sorting. . . . .	89
3.6	Histogram of (a) number of participants and (b) number of parents who categorized each activity as math and art. Items are sorted from most math sorts to least in each figure. . . . .	90
3.7	Histogram of number of activities categorized by children as (a) math and (b) art.	90
3.8	Histogram of number of activities categorized by parents as (a) math and (b) art.	91
3.9	(a) Number of participants in Study 2 who answered “yes” when asked whether each item <i>could involve math</i> and (b) linear regression showing relationship between breadth of conception and math anxiety in Study 2 ( $\alpha = 2.10$ , $\beta = -0.02$ , $p < .05$ ). . . . .	95
3.10	(a) Boxplot of anxiety scores by condition. (b) Boxplot of conception score by condition. . . . .	96
3.11	Number of participants who categorized each item as involving math in Study 3.	99
3.12	(a) Histograms of conception scores and (b) AMARS math anxiety scores. . . .	99
A.1	Differences in ratings by gender and time when attitudes were formed (a) for attitudes about math in school and (b) at work. . . . .	119
A.2	Differences in ratings by gender and time when attitudes were formed (a) for attitudes in Grades 1–5, (b) in grades 6–8, (c) in grades 9–12, and (d) in college.	120

# List of Tables

0.1	Number of documents for each corpus. . . . .	10
1.1	Mean scores and perceived scores by condition (standard deviations in parentheses). . . . .	22
1.2	Pearson correlations between pre- and post- self-assessments and between self-assessments and actual score in both conditions. . . . .	22
1.3	Summary table (number of participants, mean score, mean post score estimate, mean time, and Pearson correlation between scores and post score estimates). . . . .	42
1.4	Average model parameters separating those with $\epsilon = 0.5$ and those with $\epsilon < 0.5$ . . . . .	46
2.1	Number of people in Study 1 who said they formed their opinions in each time-period as well as mean ratings of those individuals and standard deviations. . . . .	55
2.2	Number and percentage of female and male respondents who decided their feelings in each time-period as well as mean ratings in each bracket about math and standard deviations. . . . .	56
2.3	Number of people who said they formed their feelings about math in each time-period as well as mean ratings and standard deviations for those groups. Notably, relatively few made their decisions in college or after (83% of participants formed their feelings at some point in primary or secondary school, especially in high school). The 25 unaccounted for said they did not recall when, also a markedly small proportion. . . . .	63
2.4	Mean attitude ratings and standard deviations for women and men. . . . .	65
2.5	Example uses of informative features from attitude rating classifier. . . . .	68
2.6	Example uses of informative features from math anxiety rating classifier. . . . .	69
2.7	Example uses of informative features from gender classifier. . . . .	69
2.8	Mean ratings standard deviations made in each time-period subsection of the survey (e.g., “in grades 1-5, how did you feel about math” (on a scale from 0 to 10). . . . .	72
2.9	For responses from each time-point, proportion for which each topic was dominant (the topic a response is about). . . . .	74

3.1	Descriptive statistics for four blocks in Study 1. ‘Items Categorized as Math’ is out of a total of 32, and was analyzed as a proxy for the breadth of participants’ math conceptions. ‘Math Anxiety’ is on a 10-point self-report scale. ‘Self-Assessed Skill’ represents the mean skill rating on a 5-point Likert scale, across all items for all participants. ‘Math Mindset’ is coded to be on a 5-point scale indexing how fixed individuals believe math ability to be, with larger values indicating more fixed mindsets. . . . .	84
3.2	Mean number of activities categorized as math or art by children and parents (all out of 18). . . . .	91
3.3	Mean concept and anxiety scores and standard deviations for each condition. . .	95
3.4	Means and standard deviations for math anxiety and math conception scores by condition. . . . .	100
A.1	ANOVA results predicting general attitudes about math using Mother’s Education as the measure of SES on the left (n=879) and using Hollingshead on the right (n=799). Because of the Bonferonni correction, we note with an asterisk (*) the instances when $p < 0.0036$ . . . . .	118
A.2	ANOVA results predicting attitudes about math in school. . . . .	121
A.3	ANOVA results predicting attitudes about math at work. . . . .	122
A.4	ANOVA results predicting attitudes about math in grades 1–5. . . . .	123
A.5	ANOVA results predicting attitudes about math in grades 6–8. . . . .	124
A.6	ANOVA results predicting attitudes about math in grades 9–12. . . . .	125
A.7	ANOVA results predicting attitudes about math in college. . . . .	126
A.8	Classifier results for responses of positive vs. negative ratings (accuracy on the test set: 77%). . . . .	127
A.9	Confusion matrix for Naïve Bayes Classifier for attitude ratings on the test set. It shows 34 negative responses and 94 positive responses were correctly predicted.	127
A.10	Classifier results for responses of high vs. low anxiety ratings (accuracy on the test set: 69%). . . . .	128
A.11	Confusion matrix for Naïve Bayes Classifier for anxiety ratings on the test set. It shows 43 high anxiety responses and 65 low anxiety responses were correctly predicted. . . . .	128
A.12	Classifier results for responses of men vs. women (accuracy on the test set: 60%).	129
A.13	Confusion matrix for Naïve Bayes Classifier for gender on the test set. It shows 37 responses by women and 75 responses by men were correctly predicted. . . .	129
A.14	Topic modeling output with 5 topics. This table shows the 10 words (or word roots) with highest probability under each topic and each word’s associated probability under that topic. . . . .	130
A.15	Topic modeling output over all time-points with 5 topics. This table shows the 10 words (or word roots) with highest probability under each topic and each word’s associated probability under that topic. . . . .	131

## Acknowledgments

I most pressingly need to recognize the outstanding mentorship I have received throughout my time as a grad student. Tom Griffiths and Anna Rafferty made me feel enormously supported through great personal and global transitions and crises. The sheer number and breadth of drastic events throughout my time in grad school has been remarkable: the move of the lab, the California fires and power outages, an international pandemic, not to mention numerous transitions in my personal life. It feels unlikely that I could have completed this dissertation throughout all of this, and yet, by virtue of being so fortunate in my pairing of mentors, I have successfully completed a great deal of work and learned an undeniably massive amount. Though adherence to “rules” forced Tom to lose his official status as co-chair of my dissertation committee, he remains, along with Anna, responsible for the completion of my degree. I thank them for setting the bar incredibly high – I hope to pay forward what I’ve been so fortunate to learn from them both. Gigantic thanks also to Mahesh Srinivasan and Ann Kring for having my back; Marcia Linn and Alex Paxton for their much-needed extra mentorship; Steve Piantadosi for adding his wisdom to my dissertation committee; and those from my labs who have given me community, most especially Sara Gottlieb-Cohen, Stephan Meylan, Jordan Suchow, David Bourgin, Paul Krueger, Jess Hamrick, M Pacer, Ellie Kon, Aida Nematzadeh, Josh Peterson, Thomas Langlois, Daniel Reichman, Monica Ellwood-Lowe, Ariel Starr, Beth McBride, Emily Harrison, Laura Armstrong, and my multiple crack teams of research assistants.

I also thank my family – firstly my moms, Barbara and Chris, who have kept a safe space for me in their home and helped me maintain some sense of self along with Charlie, the most neurotic of dogs. I am also grateful to those who preceded me in academia, namely my Uncle Art, a professor of pharmacy who is always brimming with advice and stories, as well as my grandfather Leo, reportedly the last Jew in Germany (in ‘38) to receive a PhD. And Rachel Hachlili, a professor of Archeology who was always one I could count on to gripe about sexism in Academia with. Finding my way to this path is very much thanks to this specific familial privilege. I also would not be where I am without all my influential math teachers and professors, including Alexandre (1st grade), Ms. Labreque (6th grade), Mme. Planchon, M. Macharinow (high school), and Profs. Laura Stevens, Ron Evans, Daniel Wulbert, and David Meyer as well as my other math role models: Valerie, Anila, Natalie, and Katie of the UCSD math club, the staff from my time at MoMath, and of course my father.

It’s very important that I acknowledge my bookclub (specifically our inimitable leader Jen Pearlstein) and my local bookstores (Mrs. Dalloway’s, European Books and Media, Moe’s, and Pegasus). We united as a gathering of compulsive readers and I am so grateful to this group of women for sharing ideas so freely. Enormous thanks also to my housemate Chris Wong who single-handedly prevented my descent into complete chaos during the pandemic and final writing stages. I also thank my personal support network, Rebecca and baby Owen, Katy and Ellen, Jessica, John, Josephine, Colin, India, Madeline, Natalie and Ivan, Darby, Brian and the rest of my Alzheimer’s support group, Sabrina, and my dear departed Gad. Of

course I wouldn't have come this far without my therapist who describes me as "articulate, but vague" – I sincerely hope this dissertation is only one of those.

Finally, and with greatest emotion, there is Ruthe. If ever I regretted beginning this journey, the feeling dissipated as soon as I realized I would not have otherwise met my dear Ruthe. What started out as a small collaboration unifying our two research foci became one of the most crucial friendships I have and will ever embark on. To be in her extraordinary presence makes me feel extraordinary.

*You are so good, you made the mathmos vomit.*

Barbarella

**-1**

## Introduction

EVERYTHING WE BELIEVE is brought upon by propaganda. It may be unwittingly so, but the majority of the information we absorb and generate is biased, whether it asserts that proper hygiene means owning many types of soap (Hamblin, 2020), that intelligence is a fixed construct (deBoer, 2020), that all forest fires are bad for the environment (Stephens et al., 2020), or that standards of beauty are exclusively based on evolutionarily advantageous traits (Wolf, 2002). The prevalence of misinformation causes us to truly believe (and have difficulty unbelieving) that natural diamonds are rare and valuable (Epstein, 1982), that sugar specifically causes hyperactivity (Wolraich et al., 1995), that we need to drink eight glasses of water per day (Vreeman & Carroll, 2007), and many more myths. The inception of such ideas occurs both through the deliberate spread of misinformation (as humanity has been reckoning with especially recently), but also via more subconscious and often unintentional means – what Lewis Thomas calls “spontaneous, undirected, societal *propaganda*.” One illuminating example is that of crossword puzzles. They may seem innocuous, but a recent analysis revealed that more published puzzles are composed by men than women and the clues and their answers most frequently reference people who are white and/or male.<sup>1</sup> This particular messaging is subtle, but even so prevents many from feeling able to complete crossword puzzles. And as the article states, “crosswords tell us something about what we think is worth knowing.” The researchers demonstrate just how straightforward it is to adjust the clues to be more inclusive. The simple presentation of these data is an intervention on anyone who has internalized that crossword puzzles are not for them. More generally, through exposure to the origins of ideas and belief systems, we can slowly undo the power certain problematic ideas have over our society.

A particularly insidious set of views relates to how numerous people perceive that they have an inability to engage with mathematics. This is in large part driven by the host of ways math is negatively portrayed and discussed. Certain turns of phrase likely come to mind easily, such as the concept of being a “math person,” and are referred to regularly as tropes in fictional contexts (see references to the “Everybody Hates Math” TV trope<sup>2</sup>).

---

<sup>1</sup> “Who’s in the Crossword?”

<sup>2</sup> <https://tvtropes.org/pmwiki/pmwiki.php/Main/EverybodyHatesMathematics>



Because the language used (Maier & Abdel Rahman, 2018) and the frequency of referencing a concept (W.-C. Wang et al., 2018) are deeply connected to how people perceive it, perceptions of math no doubt are shaped by these prevalent negative portrayals of the subject (Eccles & Jacobs, 1986). The vilification of math is further advanced by the ubiquity of unnecessary contrasts. For example, many identify as either good at math *or* good at reading, which is partially influenced by opposing stereotypes about the domains: girls are perceived to be superior in reading ability while boys hold the superior stereotype for math (Plante et al., 2019). As these perspectives become more commonplace, people end up making unnecessary choices in their pursuit of knowledge (Breda & Napp, 2019). Removing language that perpetuates stereotypes and negativity about math from everyday speech may allow children to discover that they are interested in math. In addition to media representations and colloquialisms, research about math perceptions has also largely been limited to a narrow swath of ways it can be perceived, namely through the lens of math anxiety (Foley et al., 2017), which in turn limits possible schemas of math perceptions. If researchers think more broadly about math attitudes (as with those who study “enchantment” in literary contexts, such as in Prezioso, 2020) and disseminate their findings, news outlets will report about broader views of math, and the population may in turn consider math more generously, which will then give researchers a more comprehensive array of attitudes to explore.

As I describe further in the next section, this dissertation is comprised of an investigation of how math is publicly discussed followed by three sections that experimentally address (1) perceptions of *ability*, (2) perceptions of *belonging*, and (3) perceptions of *conception*. That is, (1) how people perceive their ability in math, (2) autobiographical accounts of experiences with math, and (3) what people think the term “math” refers to. In all chapters, I explore contrasts between math and other domains (e.g., reading, science, or art) and also identify various origins of these perceptions. I begin to delve into possible ways of intervening as I explore the connections between different types of perceptions and take a developmental approach by conducting research with children of various ages alongside adult participants.

## The present research

You’re coming of age in the 21st century, a century in which, I promise you, mathematics is going to play a starring role.

---

President Josiah Bartlet, *The West Wing*

Math achievement is indisputably crucial for student success. Regardless of whether expertise in all domains of math is truly necessary for many careers, it constitutes a hurdle for admission into high school, college, and graduate school (Lee, 2012). The 2012 PISA

report found that “proficiency in mathematics is a strong predictor of positive outcomes for young adults, influencing their ability to participate in post-secondary education and their expected future earnings” (OECD, 2013). Yet math anxiety and negative stereotypes are prevalent and cause many to avoid math-related coursework which in turn prevents them from pursuing math-related careers (M. T. Wang & Degol, 2013). Attitudes about math tend to form early (Bian et al., 2017) and impact math achievement significantly (Seaton et al., 2014). Students’ sense of belonging in mathematics predicts learning of the subject (Barbieri & Miller-Cotto, 2021). And in order to intervene to ameliorate negative feelings towards math, we need to fully understand how individuals perceive math.

Here, I explore mathematics perceptions via a broad range of research methodologies in order to further extend existing work. The studies in each section either make use of practices from different research contexts or academic disciplines or apply new study designs. With methods not typically used in the study of math perceptions, I show patterns that align with earlier work and demonstrate that we can glean more information from alternative measures. I pair these experiments with computational modeling methods, namely Bayesian models of cognition and computational text analyses to demonstrate the usefulness of computational modeling for studying affective constructs in more naturalistic settings. I begin with online experiments, but scale to more ecologically valid paradigms throughout each chapter.

In a preliminary chapter (Chapter 0), I use a data-driven approach to demonstrate that math is in fact discussed negatively compared to other domains in public discourse. This investigation serves to verify many of the claims put forth in this introduction and to justify the remainder of the studies in this dissertation.

The first chapter about math perceptions (Chapter 1) focuses on perceptions of *ability* by drawing on research about self-assessment (the act of estimating one’s performance on a task) and developing a Bayesian model to assist with interpreting the meaning behind possible choices underlying self-assessment judgements. Using this methodology across a variety of domains and in educational contexts shows the wide applicability of self-assessment as a tool for assessing perceived ability when paired with an interpretable modeling approach. I show that this model produces explainable parameters that vary as we might expect, and that the model’s versatility allows for testing hypotheses and adapting it to fit multiple forms of self-assessment data. I observe that there is a substantial amount of both domain differences (lower perceptions of ability in math compared to trivia, for example), and individual variability in explanations for poor or superior calibration to true performance.

In the next chapter (Chapter 2), I investigate perceptions of *belonging* through open-ended narratives written by adults about their experiences learning math. By leveraging computational text analysis methods, I identify systematic patterns across narratives and thus capture a fuller and more accurate picture of attitudes and how experiences vary. These results align with earlier findings (e.g, women use more negative language than men when writing about their experiences with math) and contain new insights that can be further explored (e.g., those with low math anxiety highlight the practicality of math in their narratives, while those with positive feelings about math frequently mention enjoyable contexts for engaging with math).

The final chapter (Chapter 3) explores perceptions of *conceptions* of mathematics, or what people think math is. My collaborator and I developed a method for assessing how broadly individuals think about math and adapted it for various ages and contexts, including preschoolers, adults online, children in India, and adults in a museum. I compare responses on this measure with more direct measures of attitudes to begin to link the different types of perceptions explored through my dissertation. Most interestingly, ideas of what *math* refers to vary considerably within age groups and there may be important contextual influences.

I conclude this dissertation with a discussion of ways to extend these findings to practice through interventions and a shift in research modus operandi. I also consider how the ideas explored here extend beyond the domain of math: a thorough investigation of how humans perceive anything can make us better equipped to reverse the effects of problematic propaganda.

## A note on social constructs

Because this woman perceived her story to be true, it was true in its consequences.

---

Pauline Boss, *Ambiguous Loss*

When studying math ability and attitudes, researchers often have often chosen to analyze differences across gender groups (Benbow & Stanley, 1980; Hyde, 2005). With something as pervasive/delicate/wide-spread as gender differences, it is crucial to avoid making unverified biological claims when there are an indefinite amount of social and cultural influences behind our reality and statistics. There are enormous gender data gaps in most contexts (Perez, 2019) and yet much research has been devoted to demonstrating women's inferiority in mathematics (Benbow & Stanley, 1980) and other domains that are perceived to involve innate brilliance (Leslie et al., 2015). This surplus of gender-disaggregated data has perhaps given a misleading impression of what is worth focusing on. You may wonder why I say so little about gender in this dissertation, other than Chapter 2: I have chosen to focus only on gender differences in explicit recollections of experiences with math because any other hint of a gender difference would be a disservice to women. I acknowledge that women have lived in a different world from men – that the differences in lived experiences are what matter, rather than possible differences in performance, metacognitive abilities, or ideas about what math is at this stage in human history. Of course, these are all interconnected concepts, as I will highlight throughout this document, but specifically referencing gender differences too frequently goes against my main point: the way we communicate about math can have great impact and to spend too much time on gender differences in this context would lend more credence to the existence of inherent differences. Until environments for different social groups are considered identical to one another, we cannot actually draw any firm conclusions about ability differences.

*If such a slight change in temperature was all it took to transform the life of a public square, why should we think the course of human history any less susceptible?*

Amor Towles, *A Gentleman in Moscow*

0

## Math Discourse

### *How do people talk about math?*

WHAT POINT ARE WE MAKING when we contrast math with other topics? In studies of school performance, attitudes, and stereotypical beliefs, math is most frequently compared to language abilities and occasionally artistic qualities. Most studies about these topics administer assessments and closed-form surveys to make sense of how math ability or beliefs are different from similar constructs in other educational domains. In an analysis of Google search terms using Google Trends, “math” occurs in search queries far more frequently than “language” or “art” and—unlike searches about the other topics—the prevalence of “math”-related searches shifts in conjunction with the academic year, likely alluding to the relevance of math in daily life, though especially in educational environments. This chapter’s goals are to (1) sample from diverse naturalistic text-based datasets to expose how math is referred to in non-experimental settings and (2) identify similarities and differences between math and the domains most frequently used as contrasts. I perform computational analyses on text derived from naturalistic sources written across a variety of different registers, from a journalistic source (*New York Times*) and a social media website (Twitter) to referential sources containing basic definitions (Merriam-Webster) and more informal descriptions (Urban Dictionary). We see that, across data sources, queries related to “math” refer more frequently to education-related themes and incorporate more disparaging terminology compared to content related to “language” or “art.” This knowledge can inform and empower future researchers and practitioners interested in changing the discussion around math.

---

This chapter is derived from Jansen and Foushee (2020).

## 0.1 Introduction

Math is frequently discussed in a derogatory way: “Everybody hates math” is a regularly referenced trope in popular media in the United States, with hundreds of instances across television sitcoms, comics, and movies.<sup>1</sup> In addition, when researchers investigate attitudes about math, they tend to focus their efforts on the study of math *anxiety* (E. Carey et al., 2016; Foley et al., 2017), as opposed to positive feelings. These ways of portraying math, both in the media and in research and education appear unique to math. When researchers measure stereotypical beliefs (Chestnut & Markman, 2018) or attitudes about another school subject (Gunderson et al., 2017), they are typically included as a contrast to math.

In public discourse, this differential treatment of math compared to other domains is perpetuated by *neuromyths*, or false ideas about the brain, such as the idea that “some of us are ‘left-brained’ and some are ‘right-brained’ and this helps explain differences in how we learn” (Macdonald et al., 2017). Implying a natural contrast between math ability and art or language ability does a disservice to students, current and former: it encourages the belief that if we are “good” at one thing, we cannot be “good” at another. Beliefs about innate brilliance further amplify the folk distinction between math and other subjects. Math is perceived as requiring the most brilliance out of any STEM discipline, and significantly more than all art and language-related fields, including English Literature, Art History, Linguistics, and Music Composition. Such essentialist beliefs about domain-specific ability are bolstered by parallel sex disparities, with more “brilliant” fields like math including significantly fewer women (Leslie et al., 2015).

Here, I capitalize on naturally occurring data where people discuss math, and compare parallel discourse about other domains. I specifically analyze communications related to language and art, as these are frequently compared to math. I detail my rationale for each comparison domain in the next section.

### Comparison domains

In research contexts, language very frequently serves as a comparison domain for math. Math is compared to language—and most often reading and writing skills—in research on ability (Guiso et al., 2008), stereotyped beliefs (Chestnut & Markman, 2018), and theories of intelligence (Gunderson et al., 2017). O’Dea et al. (2018) contrasts STEM performance (math and science) with non-STEM (language, humanities, and social science) and finds no actual performance difference nor evidence for gender differences in variability in academic grades.

On the other hand, art is much less studied in relation to math, in part because it is far more difficult to design art assessments than math or language assessments, and judging art ability is perceived as subjective. Some work has contrasted creative ability in math and art (Jeon et al., 2011) and explored gender differences in stereotypical beliefs across the

---

<sup>1</sup><https://tvtropes.org/pmwiki/pmwiki.php/Quotes/EverybodyHatesMathematics>

two areas (Steele & Ambady, 2006). More frequently, when conducting research on a task that involves some artistic expression in research about math ability, the idea of “art” goes unmentioned, as with drawing (Fan, 2015) or any study of spatial ability (Barner et al., 2016b).

In studies comparing math ability or perceptions to parallel constructs in another domain, no justification is given for the choice of alternative domain. It may seem obvious to us that reading ability is the most direct contrast to math ability, but this is something that could stand to be reassessed. The assumptions about concepts and distinctions among them present themselves in our communications. Therefore, investigating naturally occurring data may provide us with the justification we need for domain comparison choices across different contexts. I assert that though art is only sporadically studied in relation to math, there are many reasons why it may serve researchers as an appropriate foil. For example, math is a required course throughout schooling and necessary for attending college (Lee, 2012), while when faced with budget shortages, art classes are the first to be cut from curricula. Though these subjects are treated very differently in educational settings, professional mathematicians regularly enjoy drawing comparisons between artistic and mathematical abilities (Chaitin, 2002; Lockhart, 2009).

## Measuring math talk

Human attitudes are typically explored via closed-form surveys in addition to open-ended prompts and semi-structured interviews. But sampling bias as well as the wording of the questions can impact responses. I thus use naturally occurring datasets to supplement existing research about math attitudes and as a guide for developing new theories and experimental paradigms (Paxton & Griffiths, 2017).

The goals of this chapter are (1) to source data from non-experimental contexts to examine naturalistic discourse surrounding math and its comparison domains and (2) to identify ways in which math is discussed that differ from related domains in similar contexts so as to make an argument for why this dissertation is primarily devoted to mathematics perceptions.

## 0.2 Methods

I started by identifying a variety of online sources with freely accessible APIs (Application Programming Interfaces). First, I used Google Trends as a measure of frequency of term usage. Next I collected a selection of articles from the *New York Times*, tweets from Twitter, and definitions from the Merriam-Webster dictionary and Urban Dictionary that related to the search terms “math,” “language,” and “art.” I recognize that these terms, particularly “language” and “art,” may be used in contexts beyond the abstract domain I am considering here. This chapter is intended as a preliminary exploration of these terms used broadly in the hopes of refining search terms to more accurately contrast with one another.

## Data sources

Though Google does not provide access to search history data, the company built an online interface, Google Trends,<sup>2</sup> for observing both fluctuations of searches for specific keywords or topics over time and across locations (Stephens-Davidowitz & Varian, 2014). I focused my observations exclusively on the US. This source provides a very general sense of how these topics are thought about differently. I next explore actual word usage in multiple other sources, namely the *New York Times*, Twitter, and two different online references (Merriam-Webster and Urban Dictionary). Two of these may be seen as relatively objective (Times and Merriam-Webster), though a computational analysis of word usage will show whether this is truly the case.

I used the “Article Search API” from the *New York Times* (NYT)<sup>3</sup> to collect all articles that include the terms “math,” “language,” and “art.” The NYT API provides the headline, keywords, date, word count, and lead paragraph for all articles that come up for a specific search term. The NYT Article Search API yielded 441,773 results for “math”, 367,707 for “language” and 1,276,036 for “art.” Due to restrictions with the API, I was allowed to sample 2,010 results for each search term.

Because of its dual role as a social media site and a host to public conversations, Twitter provides easy access to build datasets of user postings, in contrast to some other social media companies like Facebook.<sup>4</sup> In order to align results with the data I obtained from the *New York Times*, I used the `twitter` package for Python<sup>5</sup> to load the most recent 2,000 tweets per search term on March 4, 2020.

Merriam-Webster additionally offers easy access to their definitions.<sup>6</sup> This data mining returned just the definitions for each term of interest, but it is meaningful that “math” has three definitions, while “language” and “art” each have ten. From Urban Dictionary, I downloaded all existing results for each term, which was 856 for math, 262 for language, and 876 for art (see Table 0.1 for total documents used in analyses for each term and each data source).

## Text Analyses

With each data source, I created Naïve Bayes classifiers to contrast word usage for documents about math, language, and art.<sup>7</sup> Details about this algorithm and reasons for choosing it are included in Chapter 2 where computational text analyses are employed in greater depth. Prior to text analyses, I ran a series of standard text pre-processing techniques: a) removing stopwords b) removing punctuation and c) reducing words to their roots (stemming and

---

<sup>2</sup><https://trends.google.com/trends/?geo=US>

<sup>3</sup><https://developer.nytimes.com/>

<sup>4</sup><https://developer.twitter.com/>

<sup>5</sup><https://pypi.org/project/twitter/>

<sup>6</sup><https://dictionaryapi.com/>

<sup>7</sup>We employ the `NaiveBayesClassifier` function from Python’s Natural Language Toolkit (`nltk` version 3.2.2) package <https://www.nltk.org/>

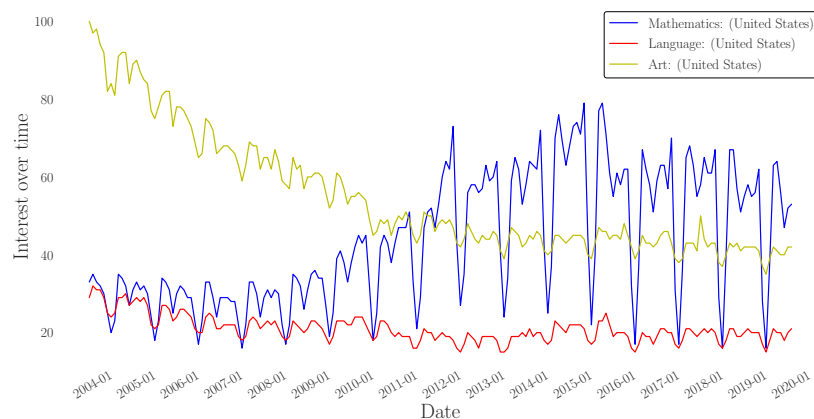


Figure 0.1: Google trends patterns of searching for the topics “mathematics,” “language,” and “art.”

lemmatizing). To test the accuracy of each classifier, I shuffled the data and separated it into a training set consisting of 80% of the data and a test set comprising the remaining 20%. I trained each classifier on the training set, then report the classifier’s accuracy predicting responses on the test set, alongside a subset of informative features (words that are more representative of one specific subgroup).

## 0.3 Results

### Google Trends

From Fig 0.1, created from data generated in Google Trends to compare searches pertaining to the topics “math,” “language,” and “art,”<sup>8</sup> it is clear that, compared to the other terms, math is generally searched for at a higher rate, though it takes steep dives in the summer months when school is no longer in session. This suggests that the term “math” is much more associated with formal education than these contrasting terms, an idea I explore in more detail in the other data sources.

### Journalistic source

I excluded a set of “math” searches from the initial 2,000 to ensure that the results would not be overly skewed. Specifically, 270 hits contained a daily math challenge and the lead paragraph began with “Test your math skills with today’s question” and an additional 40

<sup>8</sup><https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fm%2F04rjg,%2Fm%2F04g7d,%2Fm%2F0jjw>



	MATH	LANGUAGE	ART	TOTAL
NY TIMES	1,541	1,946	1,835	5,322
TWITTER	2,000	2,000	2,000	6,000
M. WEBSTER	3	10	10	23
URBAN DICT.	856	262	876	1,994

Table 0.1: Number of documents for each corpus.

started with “Our weekly math problems are written by teachers at Math for America.” For “art,” I excluded 157 whose lead paragraphs began with “Our guide to new art shows and some that will be closing soon.” There did not appear to be anything so consistent for searches relate to “language.” Search results with blank lead paragraphs (159 math; 18 art; 64 language) were excluded from the training data. First, I analyzed the distribution of keywords. Of the 1541 math queries, 323 contained “school” (21%) compared to 81 of 1946 language queries (4%) and 14 of the 1835 art queries (0.8%). There was a similar pattern for the keyword “test,” included in 149 math queries, compared to 6 and 1 for language and art, respectively.

I next looked at the text from each lead paragraph. I used all three sets of nonempty lead paragraphs for each topic to construct the classifier which included a total of 5,322 texts (1541 math; 1946 language; 1835 art). After removing all terms with roots “math,” “language,” or “art,” the classifier achieved 70% accuracy on the test set. The most informative features for math included “test,” “score,” “grader,” “improv,” “educ,” “competit,” and “teacher” (e.g., “Growing up, I thought math class was something to be endured, not enjoyed. I disliked memorizing formulas and taking tests, all for the dull goal of getting a good grade”). The most predictive terms for an art-related hit contained “galleri,” “sculptur,” “paint,” and “noteworthi.” For language-related queries, “speak,” “translat,” “dictionari,” and “writer” were most informative. Though there are no apparent emotive terms, math arises much more frequently in documents related to school than does art or language in this source.

## Social media

The average word count for tweets corresponding to each term was approximately equal (18 for math, 19 for language, and 18 for art), likely due to platform word count restrictions. In the classifier (accuracy: 62%), informative features for tweets about math included words very similar to those from the *NYT*, such as “test,” “fail,” “wrong,” “class,” and “science.” For language, we saw “tiktok,” “english,” “speak,” “video,” and “utter.” The set of most informative features for art contained “draw,” “style,” “anim,” “design,” and “cute.” Here, we see many domain-specific similarities to the *NYT* data, but with the addition of terms conveying negative emotions occurring in texts referencing math, such as “wrong,” “fail,” and “hard,” which might speak to the greater subjectivity of the text source. The language- and art-related searches also appear to encompass more popular culture references. There

was an interesting pattern of math being more ubiquitous in tweets relating to current events such as the election (note that this data was gathered in the lead-up to the 2020 democratic primary): ‘berni,’ ‘vote,’ and ‘warren’ (e.g., “Math says that Warren has a path”) and the very beginning of the coronavirus outbreak: “million,” (e.g., “It is simple math. The flu infects millions a year”).

## Reference materials

The Merriam-Webster dictionary produced few results, but the primary definitions themselves serve as a baseline set of relevant objective terms. Mathematics is defined as “the science of numbers and their operations,” language as “the words, their pronunciation, and the methods of combining them used and understood by a community,” and art as “skill acquired by experience, study, or observation.” By their very definitions, math is a science (rather than an art) and art is referred to as a “skill” rather than requiring innate ability.

The Urban Dictionary API yielded 856 definitions of “math,” 262 of “language,” and 876 of “art.” The mean length of the math definitions was 36 words, 49 for language, and 58 for art (similar to the *NYT* results). Because this corpus was not evenly distributed across topics, I ran separate classifiers between each pair of topics, rather than over all 1,994 total definitions. For the math/art classifier ( $n = 1732$ ), accuracy on the test set was 79% and the most informative words for math definitions were almost all negative: “abus,” “number,” “mental,” “stress,” “tortur,” and “bore.” For art on the other hand, informative features included “style,” “emot,” “draw,” “amaz,” “visual,” and “color.” The classifier comparing math to language yielded an accuracy of 85%<sup>9</sup> and primarily identified words that were informative of the language texts, as they represented a smaller proportion of our dataset. These included “speak,” “talk,” and “special,” while terms more indicative of a math entry were “abuse,” “mental,” “human,” and “bore.”

Finally, the classifier comparing art and language arrived at an accuracy of 85% and words from art-specific definitions included “music,” “best,” and “style,” while definitions of language included terms “sign,” “french,” “wrong,” and “number,” which potentially likens math to language more than to art.

Analyses of the example sentences included for each definition from Urban Dictionary produced comparable results, with notably more derogatory and profane terminology used to describe math than the comparison domains.

## 0.4 Discussion

In a preliminary analysis of naturally occurring data sources, we observe that math is more frequently written about and discussed in relation to education compared to language and

---

<sup>9</sup>Chance would be 77% because 77% of definitions are math ones. So though this classifier appears to be performing better than the math/art classifier, it was an easier task to classify math and language-related documents.

art. Or more accurately, the words “math,” “language,” and “art” co-occur with different kinds of words. In contexts where valenced language use is common (Twitter and Urban Dictionary), math is discussed using notably more unflattering terminology. Each data source explored yielded different frequencies at which the three chosen topics were mentioned. The Google Trends analysis revealed that math is searched for more frequently than language or art. However, though the *NYT* provided fewer “language” articles than “math” ones, there were more than double the number of hits related to “art” (owing to “arts and leisure” having its own section in the newspaper). In the references, “math” had fewer entries in Merriam-Webster compared to language and to art, but in Urban Dictionary, there were a comparable number of entries for “math” and “art,” and this was more than triple the number of “language” entries.<sup>10</sup> Based solely on these simple search counts, we can identify important differences in how these topics are thought about: “math” appears to be defined more narrowly than the other domains (based on Merriam-Webster definition counts and shorter text lengths in the *NYT* and Urban Dictionary data) while emotions surrounding “math” and “art” are stronger than for “language” (based on the relative number of Urban Dictionary results). “Math” is also much more associated with education, a claim supported by the keywords from the *NYT*, our classifiers’ informative features from the *NYT*, Twitter, and Urban Dictionary data, and from the cyclical nature of Google searches for “math.”

## 0.5 Conclusion

Using large-scale datasets of naturally occurring text, this chapter presents an initial exploration of how math is discussed, compared to its most frequent comparison domains. These data confirm that the term “math” is generally referenced in more limited ways (e.g., to educational contexts) and also more negatively valenced texts. Previous work has shown that familiarity with an idea increases belief in that idea (W.-C. Wang et al., 2018), which means that the restricted and unflattering ways in which we generally discuss math may progressively degrade public opinion about the topic. If—through the media and other sources—people continue to hear (or read) about math as a narrowly defined concept associated with negative emotions, this perception will continue to thrive, and be unwittingly transmitted to future generations (Maloney et al., 2015). I take a moment here to make a plea for everyone to limit unnecessary disparaging references to math in mass communication. Now, with some understanding of the public discourse surrounding math, I move forward with studies designed to explore individuals’ perceptions of math. Contrasts to other domains are drawn in each chapter, including reading/writing (Chapter 1), science (Chapter 2), and art (Chapter 3).

---

<sup>10</sup>I was not able to acquire total hit numbers from Twitter.

*Our senses don't deceive us: our judgment does.*

Johann Wolfgang von Goethe

# 1

## Perceptions of Ability

*How good are you at math?*

METACOGNITIVE ABILITY — evaluation and understanding of one’s cognitive processes — is regarded as an inventory of fundamental skills, yet people appear poorly calibrated to their abilities. It has been documented that targeted training in metacognitive skills can lead to improved student outcomes, including higher motivation and increased performance on assessments (Zepeda et al., 2015). But what does it mean to be poorly calibrated? Is accurate calibration always the most adaptive state for a learner? The goal of this chapter is to thoroughly examine perceptions of ability by developing a model that can disentangle self-assessment judgments into multiple component traits and test out theories of self-assessment as well as compare self-assessment in math to other domains.

### 1.1 Self-Assessment

In copious work studying adult metacognition, participants appear to be miscalibrated in their ability to judge their own performance across a large variety of domains (Dunning et al., 2004; Ehrlinger et al., 2008; Zell & Krizan, 2014). Though there are age-related improvements in metacognitive abilities whereby very young children overestimate their competence a great deal more than adults (Bjorklund & Green, 1992), as well as differences by domain (Jansen et al., 2017; Tyszka & Zielonka, 2002), on most tasks, researchers find that accuracy is low when making judgments about one’s performance, either when estimating one’s score or one’s standing relative to others (Zell & Krizan, 2014). There have been studies of specific domains such as weather forecasting (Tyszka & Zielonka, 2002) and particular ways of eliciting judgments (Nelson & Dunlosky, 1991) where participants do show much better calibration to their own abilities, but in most settings, accuracy is typically limited. In an influential paper, Kruger and Dunning (1999) conducted a series of studies that suggested

---

This chapter is derived from Jansen et al. (2021) as well as conference papers Jansen et al. (2017), Jansen et al. (2018), and Jansen et al. (2020)

that poorer performers tended to be less well-calibrated in their ability to judge their performance after completing a task than higher performers (Kruger & Dunning, 1999). They construed poor perceived performance by the lowest-scoring individuals as a metacognitive deficit: the worst performers lacked the skills needed to correctly do the task and also to judge their performance on the task. Commonly known as the “Dunning-Kruger effect,” this theory continues to be featured regularly in the media (Lopez, 2017), particularly for the purpose of rationalizing others’ seemingly irrational behavior (e.g., anti-vaxxers (Andrews, 2018) and government officials (Purtill, 2018)).

While discussions of the overconfidence of poor performers have focused on the idea that these people are less sensitive to their own errors, thinking about self-assessment from the perspective of a rational agent potentially offers a different account. If we imagine individuals as naive statisticians analyzing their own behavior, the rational Bayesian solution is to combine the evidence from experience with one’s prior beliefs. If those prior beliefs are that one will perform relatively well, this should lead to poor performers overestimating their ability and good performers underestimating their ability to at least some extent (Healy & Moore, 2007).

In this chapter, I present a rational model of self-assessment that allows for unpacking various reasons behind self-assessment judgments, namely the influence of prior beliefs about ability and error detection following each individual problem. Results are broken into three sections that fully explore the model’s potential, beginning with a) establishing that the parameters we set out to evaluate do in fact correspond to the respective traits we associate with them in Section 1.3, then b) a demonstration of how the model can be adjusted to account for different theories (specifically to evaluate the existence of the Dunning-Kruger effect) in Section 1.4, and finally c) adapting the model to predict individual parameters by using confidence judgments throughout a task in Section 1.5. Before getting to the data, I begin with a detailed explanation of the model used throughout, but in each subsequent section I explain how the model is adapted for that particular context.

“My dear Watson,” said he, “I cannot agree with those who rank modesty among the virtues. To the logician all things should be seen exactly as they are, and to underestimate one’s self is as much a departure from truth as to exaggerate one’s own powers. When I say, therefore, that Mycroft has better powers of observation than I, you may take it that I am speaking the exact and literal truth.”

---

Sir Arthur Conan Doyle, *The Memoirs of Sherlock Holmes*

## 1.2 Modeling self-assessment

There have been a variety of computational models of self-assessment, some of which have been focused on alternative explanations for the Dunning-Kruger effect while others have focused on making sense of various methods and types of self-assessments. In their general model of self-assessment, Fleming and Daw (2017) took into account confidence and error detection in order to unify various methods of measuring self-assessment. This model’s parameters represented a participant’s “sensitivity” and “bias,” where sensitivity is their ability to discriminate between correct and incorrect performance and bias is a penchant towards high confidence ratings. This model is based on a signal detection approach, and aims to develop a unified computational account of metacognition that accounts for both confidence and error-detection. Formally, it asserts second-order computation as the means by which humans judge their own confidence and assumes that confidence is determined not just based on a person’s actions but along with knowledge of the covariance between decisions and metacognitive states. Healy and Moore (2007) developed a formal model to contrast expected outcomes based on the type of self-assessments measured, specifically comparing overestimation of score and overplacement in comparison to others.

We specifically model absolute self-assessment to start (where participants estimate their total score after an assessment) and introduce parameters to adjust perceived prior ability in a domain, difficulty of the assessment, and competence at accurately concluding whether an individual problem was solved correctly or not. These factors are similar to those identified by other researchers as contributing to poor absolute self-assessment. For instance, poor self-assessment has been linked to lack of insight into one’s errors (Ehrlinger et al., 2008), similar to the idea of “sensitivity” in previous models (Fleming & Daw, 2017). In addition,

a claim has been made that a person’s “self-concept” forms the foundation for participants’ judgments about their performance (Ehrlinger & Dunning, 2003). This is akin to the “bias” parameter in previous models (Fleming & Daw, 2017). A separate research group (Dunning & Helzer, 2014) additionally labeled the relevant components of self-assessment as “bias” and “discrimination.” Here, we identify a computational approach that incorporates similar parameters, one corresponding to perceived ability in a domain and another to discrimination ability. We additionally integrate a difficulty parameter, motivated particularly by results indicating that self-assessment ability is also dependent on item difficulty (Burson et al., 2006; Jansen et al., 2017), though leave manipulating this parameter for future studies. This model makes predictions based on simple Bayesian inference, but in subsequent sections of this chapter, adjusted versions of this model are presented that account for other theories.

In this basic model of self-assessment, we assume that people’s inferences about their ability are based on three factors: (1) beliefs about the correctness of individual responses, (2) beliefs about their own ability, and (3) the difficulty of the task they are performing. We then conduct a rational analysis, in the spirit of Anderson (1990), considering how a rational agent should solve the problem of estimating their ability conditioned on the observed data and their prior beliefs. The notion of rational behavior used here is a little different from classical rationality as we condition on people’s beliefs without asking whether those beliefs are well-calibrated to the environment. This allows us to explain behavior in terms of these beliefs, in accordance with other rational models of cognition (Oaksford & Chater, 1994). The rational solution to estimating one’s ability is now to use Bayesian inference, modeling someone’s posterior beliefs about their ability following an assessment as a function of their beliefs about their ability before the assessment and about the difficulty of that assessment (the priors) and beliefs about their performance on each individual problem (the likelihood).

Because in this model the participant is making inferences based on their own beliefs, the likelihood is person  $p$ ’s belief about correctness on item  $i$ , where they either believe they are correct ( $X_{p,i} = 1$ ) or incorrect ( $X_{p,i} = 0$ ). The likelihood is dependent on the difficulty of the item  $i$  ( $\beta_i$ ) and the *perceived* ability of person  $p$  ( $\theta_p$ ). If a person has perfect knowledge of whether or not they answered correctly, then the 1-parameter Item Response Theory (IRT) model known as a Rasch model (Embretson & Reise, 2013) is a reasonable way for people to make inferences about their own ability. Rasch models are widely used for estimating student ability in the psychometrics literature and can also be seen as a simple logistic regression model. This track record of previous use and the simplicity of the model led us to favor an approach based on an IRT function:

$$P(X_{p,i} = 1|\theta_p, \beta_i) = \frac{1}{1 + e^{-(\theta_p - \beta_i)}}. \quad (1.1)$$

Note that given the perceived ability and difficulty parameters, the probability of believing one gave an incorrect response is one minus the probability of believing one gave a correct response:  $P(X_{p,i} = 0|\theta_p, \beta_i) = 1 - P(X_{p,i} = 1|\theta_p, \beta_i) = \frac{1}{1 + e^{(\theta_p - \beta_i)}}$ .

Equation 1.1 assumes that people have perfect knowledge about whether they have answered a problem correctly, but in reality, people may not know the correctness of each of their responses with certainty. To account for this fact, we combine the Rasch model with uncertainty about whether the person is correct via an error parameter  $\epsilon$ . This term corresponds to the probability that a person is erroneous in their beliefs about whether or not they have correctly solved a problem. This additional parameter is similar to the “sensitivity” parameter described above (Fleming & Daw, 2017) or the idea of poor error detection (Ehrlinger et al., 2008). Thus the likelihood equation where we include an error parameter to adjust for uncertainty in people’s beliefs about their correctness becomes:

$$P(X_{p,i} = 1|\theta_p, \beta_i; \epsilon) = (1 - \epsilon) \cdot \frac{1}{1 + e^{-(\theta_p - \beta_i)}} + \epsilon \cdot \frac{1}{1 + e^{(\theta_p - \beta_i)}}. \quad (1.2)$$

The probability of the person believing they’ve responded correctly is then the probability of answering correctly and recognizing that one is correct plus the probability of answering incorrectly but erroneously believing one is correct.

On a multiple-choice assessment, participants might answer some questions correctly by guessing, even if they are unaware of which answer is correct or whether they are guessing correctly. Since people are likely to be aware that they are guessing some answers correctly at random, we add a guessing parameter,  $g$ , using a simple variant of the IRT model that is commonly applied to multiple-choice assessments to account for this additional source of uncertainty:

$$P(X_{p,i} = 1|\theta_p, \beta_i) = g + \frac{1 - g}{1 + e^{-(\theta_p - \beta_i)}}, \quad (1.3)$$

where  $g = \frac{1}{N}$  and  $N$  represents the number of possible answers. This can also be modified to incorporate imperfect knowledge as in Equation 1.2, yielding:

$$P(X_{p,i} = 1|\theta_p, \beta_i; \epsilon) = (1 - \epsilon) \cdot \left(g + \frac{1 - g}{1 + e^{-(\theta_p - \beta_i)}}\right) + \epsilon \cdot \left(1 - \left(g + \frac{1 - g}{1 + e^{-(\theta_p - \beta_i)}}\right)\right). \quad (1.4)$$

The priors are defined over the difficulty of an item  $i$  ( $\beta_i$ ) and the perceived ability of person  $p$  ( $\theta_p$ ). Here, we assume the priors are normally distributed, although the model can use any prior distribution, which would allow for making more complex predictions. Varying the skew of the prior distribution over perceived ability  $\theta_p$ , for example, would capture differing interpretations of successes and failures such as learners being more likely to attribute a failure to a lack of ability rather than the task being difficult or vice versa.

A graphical model depicting the dependencies among the variables is shown in Figure 1.9. To model people’s posterior beliefs about their ability after performing a task given both their prior beliefs and the accuracy of their judgments of correctness, we insert the likelihood and priors into Bayes’ rule:

$$P(\theta_p, \beta_i|X_{p,i} = 1) \propto P(X_{p,i} = 1|\theta_p, \beta_i; \epsilon) \cdot P(\theta_p) \cdot P(\beta_i). \quad (1.5)$$



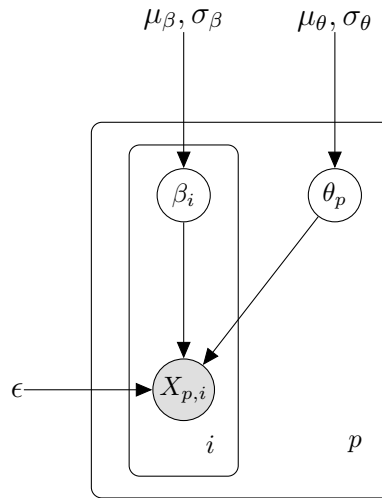


Figure 1.1: Graphical representation of the model: each observed item  $X_{p,i}$  is influenced by latent variables  $\beta_i$  (difficulty of problem  $i$ ) and  $\theta_p$  (perceived ability of person  $p$ ) as well as a constant  $\epsilon$  (ability to determine correctness), where the observed  $X_{p,i}$  refers to a person's beliefs about correctness on an item. Difficulty  $\beta_i$  is drawn from a normal distribution with mean  $\mu_\beta$  and standard deviation  $\sigma_\beta$  and ability  $\theta_p$  is drawn from a normal distribution with mean  $\mu_\theta$  and standard deviation  $\sigma_\theta$ .

Note that this equation is referring to a single item: to account for all items on a task, we multiply in each item's log likelihood. To calculate beliefs about performance from this posterior distribution, we first marginalize over  $\beta_i$ , which represents someone's posterior beliefs about the difficulty of the current assessment. We then transform the distribution to one about performance expectations rather than ability by calculating the probability of a correct response on an average difficulty question ( $\beta_i = 0$ ) using Equation 1.1 for each point in the distribution. This transformation allows for directly comparing human data to our model's output.

## Bayesian inference

Here, we demonstrate through simulations that this basic model predicts that people will not be fully accurate in their self assessments and that poor performers will tend to overestimate while high-scoring individuals underestimate, consistent with Bayesian inference shifting estimates towards the mean of the prior. Changing the mean of the prior over the ability parameter  $\theta_p$  adjusts the degree to which these patterns emerge. We begin by exploring a version of the model where people are always accurate in their judgments of correctness on individual problems, which is instantiated by setting  $\epsilon = 0$ , then show what happens when  $\epsilon$  is increased. In this model, we assume both constant mean of the prior distribution over  $\theta_p$  and constant  $\epsilon$  across individuals. We set the guessing parameter  $g = 0.2$  which assumes

there are five possible choices offered for each question.

We first consider what patterns in self-assessment occur when we assume people make perfectly accurate assumptions about their performance on each problem by setting  $\epsilon$  equal to 0. For these simulations, we assume  $\theta_p$  and each  $\beta_i$  are distributed normally such that their means  $\mu_\theta, \mu_\beta$  are 0 and their standard deviations  $\sigma_\theta, \sigma_\beta$  are 1. As shown in Figure 1.2, simulated participants from a toy example who perform on the low end tend to overestimate their performance while the highest performers slightly underestimate their score, demonstrating a pattern of results similar to those found by the original authors (Kruger & Dunning, 1999). To compute estimated performance given true score, we use Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1995) to calculate a posterior distribution over beliefs about ability (Equation 1.5) for each true score, and then transform the result into a distribution over beliefs about performance. To obtain the estimated total score from this distribution, we scale the probability correct on a single problem by the maximum score (here, ten) and compute the expectation.

We run MCMC with 10,000 iterations and remove the first 1,000 as a “burn-in,” following standard practice, before taking the mean predicted score estimate. These baseline model predictions demonstrate that self-assessment ability need not be dependent on people’s actual ability to obtain this pattern. Our rational model makes it straightforward to evaluate the consequences of changing people’s prior expectations about their ability (the prior on ability parameter,  $\theta_p$ ) or their skill at recognizing whether they are correct on each problem ( $\epsilon$ ). Changing these aspects of the model has direct consequences for the form of the function relating estimated ability to true score.

To retrieve patterns of results even more similar to those found in previous research on self-assessment, we adjust the model parameters. Varying the prior via the mean,  $\mu_\theta$ , of the ability parameter,  $\theta_p$ , changes the overall assessment of ability. As shown in Figure 1.3a, when the mean on  $\theta_p$  is high ( $\mu_\theta = 0.5$ ), reflecting optimism about ability, there is much more overestimation by all simulated learners. But when the mean is lowered ( $\mu_\theta = -0.5$ ), reflecting pessimism, we see the manifestation of the opposite pattern: except for all but the lowest performers, the model predicts underestimation rather than overestimation. The pattern of less underconfidence of the high performers compared to overconfidence of the low performers, the hallmark of the Dunning-Kruger effect, can be fit by this simple Bayesian inference model if we just increase the mean of the prior over ability  $\theta_p$ .

While changes to the prior affect the intercept of the line, changing  $\epsilon$  affects its slope. As shown in Figure 1.3b, as  $\epsilon$  increases, the slope of the line decreases. In other words, as inferences about correctness become more similar to guessing randomly (which would be captured by  $\epsilon = 0.5$ ), inferences about ability are predicted to become more and more similar to one another regardless of actual performance. Similar patterns of results are produced by manipulating the standard deviation of the mean on  $\theta_p$ ,  $\sigma_\theta$ , which are detailed in our supplementary materials. Both parameters affect the influence of performance on self-assessment. While we will keep  $\sigma_\theta$  fixed and vary  $\epsilon$ , the conclusions we draw as a result of this apply to this broader capacity for updating beliefs about ability based on performance rather than any specific parameter.

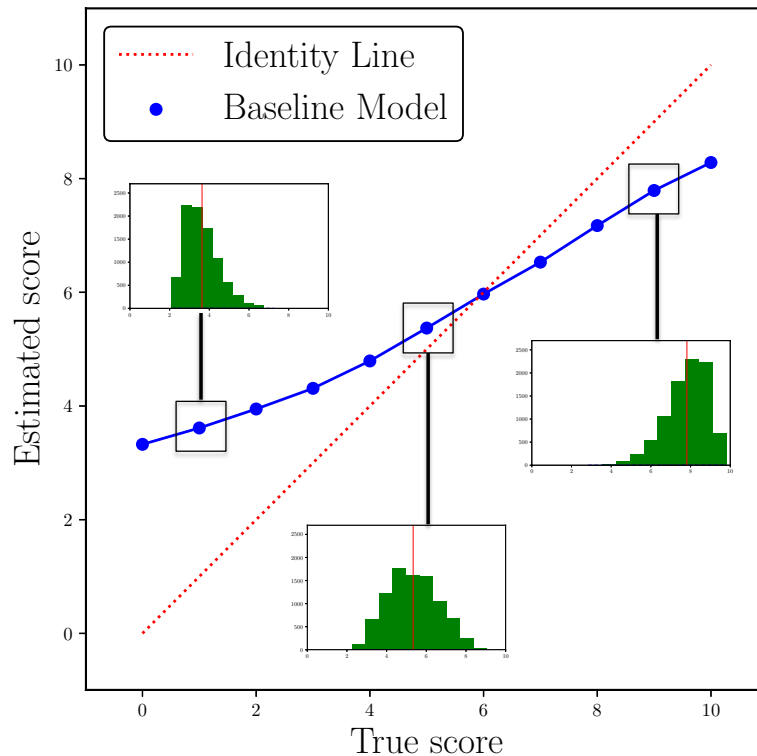


Figure 1.2: Model predictions in a toy example where participants solve 10 problems in the baseline model ( $\mu_\theta, \mu_\beta = 0, \sigma_\theta, \sigma_\beta = 1$ , and  $\epsilon = 0$ ). Each point shows the average of the posterior distribution on  $\theta$ , corresponding to estimated ability. Overlaid are histograms of MCMC estimates of the posterior distribution on  $\theta$  for three scores. The identity line represents completely accurate estimation, where a participant’s estimated score is equal to their true score. The guessing parameter  $g$  is equal to 0.2, which represents a task with five multiple-choice solutions per question.

### 1.3 Testing the model: The impact of immediate feedback

Our rational model predicts that people will inaccurately estimate their performance due to a combination of the effect of prior beliefs about their ability and imperfect skill at guessing their performance on each individual problem. To see how well models with differing parameters capture reality, we set up an experiment similar to those done by earlier researchers. One condition replicates the approach taken in previous work (Kruger & Dunning, 1999). In

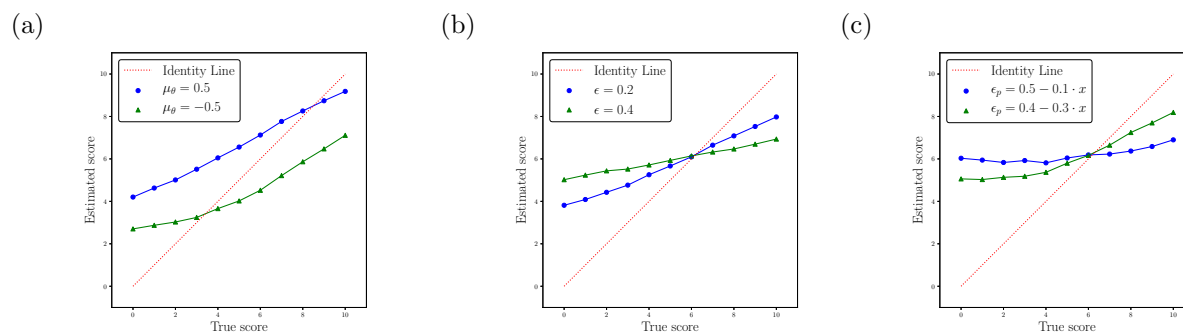


Figure 1.3: Model predictions in a toy example where participants solve 10 problems (a) when the mean on ability ( $\theta_p$ ) is adjusted ( $\mu_\theta = 0.5$  or  $-0.5$ ), (b) when the parameter  $\epsilon$  is adjusted ( $\epsilon = 0.2$  or  $0.4$ ), (c) demonstrating the Dunning-Kruger effect. Note that  $x$  in the equations in panel (c) refer to normalized score (score divided by maximum score). Again, we set  $g = 0.2$ .

the other condition, participants were given immediate feedback after each problem. In our model, this can be interpreted as reducing the  $\epsilon$  parameter – if people know whether they were right or wrong,  $\epsilon$  should be effectively 0.

## Methods

### Participants

A total of 100 participants (50 per condition) were recruited on Amazon’s Mechanical Turk and were each compensated \$3 for their time.

### Procedure

All participants completed 20 logical reasoning problems adapted from the 2007 LSAT (Law School Admissions Test).<sup>1</sup> Participants rated their absolute ability (“how many of the 20 logical reasoning problems will/did you answer correctly?”), their relative ability (“compared to other participants in this study, how well do you think you will do/did you do?”), the difficulty of the task for themselves, and the difficulty for others. To analyze self-assessment, we focus on people’s ratings of their absolute ability. At the conclusion of the study, participants were directed to a short demographics questionnaire. The “no feedback” condition was a direct reproduction of the design used in previous studies. In the “feedback” condition, participants additionally received immediate feedback after each problem they solved, which consisted simply of learning whether or not their answer was correct.

<sup>1</sup>Link to problems: [www.lsac.org/docs/default-source/jd-docs/sampleptjune.pdf](http://www.lsac.org/docs/default-source/jd-docs/sampleptjune.pdf)

	Actual Score	Estimated Score
No Feedback	9.90 (3.83)	8.70 (4.00)
With Feedback	9.52 (4.10)	8.26 (4.02)

Table 1.1: Mean scores and perceived scores by condition (standard deviations in parentheses).

## Results

Among the 100 participants (60 male, 37 female, 2 other, and 1 unspecified; mean age = 33.33 years), the average completion time was 36 minutes. On average, participants answered 9.71 problems correctly out of 20 (sd = 3.95) and the mean perceived score was 8.48 (sd = 4.00). The difference between actual score and perceived score was deemed significant by a paired-samples t-test ( $t(99) = 3.80$ ,  $p < .001$ ). Table 1.1 shows that this pattern of underestimation held for both conditions. The overconfidence of the worst performers was limited, presumably given that this test of logical reasoning – from the 2007 LSAT – was significantly more difficult than that of Kruger and Dunning (1999) – taken from the 1993 LSAT.

Pre- and post- self-assessments were correlated with one another in the no feedback condition, but not in the feedback condition. In both conditions self-assessments became better correlated with actual performance after completing the assessment (see Table 1.2), though there was still a pattern of underestimation in the data.

	Pre/Post	Pre/Score	Post/Score
No Feedback	.63***	.28*	.42**
With Feedback	.12	-.05	.90***

\* $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\* $p < 0.001$

Table 1.2: Pearson correlations between pre- and post- self-assessments and between self-assessments and actual score in both conditions.

The difference in self-assessment calibration between the conditions was deemed significant by a Fisher  $r$ -to- $z$  transformation between the Pearson  $r$  values ( $z = 4.97$ , two-tailed  $p < .001$ ), meaning those in the feedback condition, as anticipated, were much more accurate in estimating their score after the task than those in the no feedback condition.

In a linear model predicting estimated score from true score and condition, a significant regression equation was found ( $F(3, 96) = 31.45$ ,  $p < .001$  with an  $R^2$  of .50). Specifically, there was no effect of true score, but there were statistically significant effects of condition ( $\beta = -4.53$ ,  $p < .01$ ) and the interaction of true score with condition ( $\beta = .45$ ,  $p < .01$ ),

demonstrating that the effect of score on perceived score also depends on the condition, as predicted by our rational model.

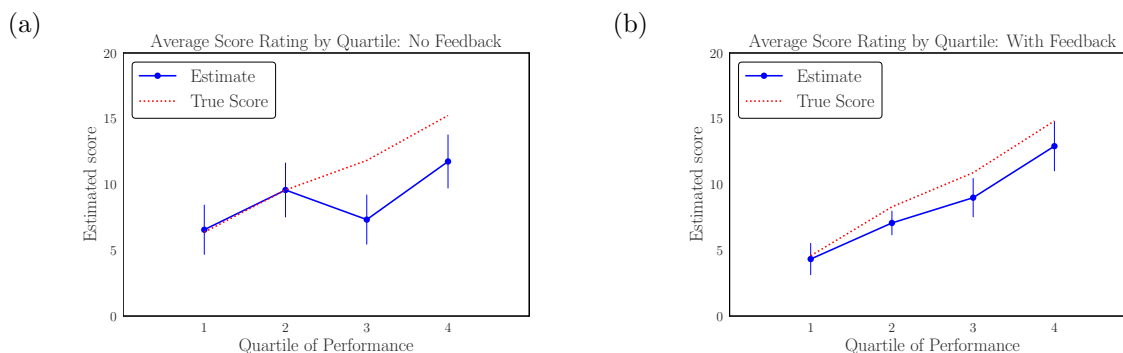


Figure 1.4: Mean estimates of score (out of 20) by quartile of actual performance in (a) no feedback and (b) with feedback conditions of Study 1. Error bars show 95% confidence intervals.

To fit the model to the data, we compare the model predictions to participants' estimates of their scores relative to their true score. Results from studies of self-assessment have typically organized their data by quartile of performance, as in Figure 1.4. However, this portrayal of the data eliminates much of its nuance. In the no feedback condition, grouping the self-assessments by true score instead of by quartiles shows significant variability, including among the worst performers (see Figure 1.5a). There is more accuracy in self-assessments than can be told from the quartile plots.

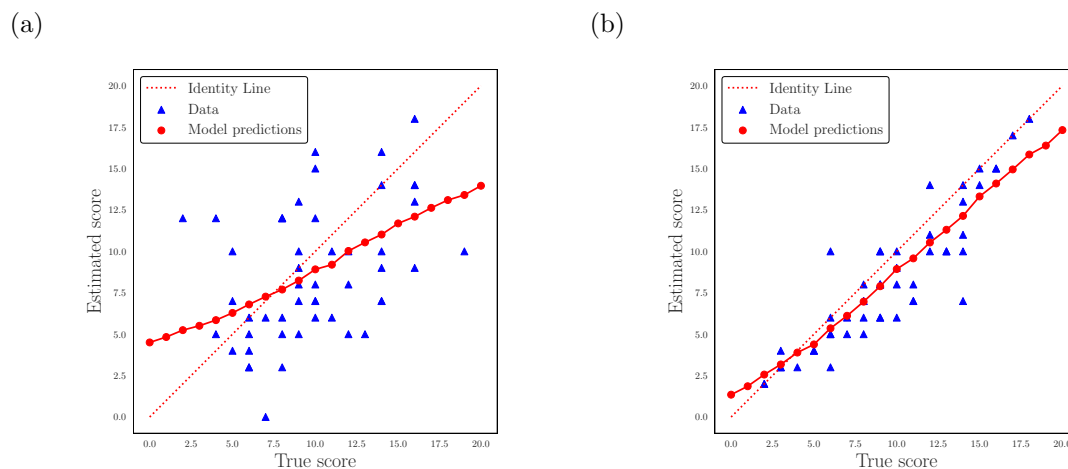


Figure 1.5: Participants' estimates of score (out of 20) by true score as compared to the best model predictions in (a) no feedback and (b) with feedback conditions of Experiment 1.

To find the best-fitting parameters for the model given the data, we perform a grid search over  $\mu_\theta$  and  $\epsilon$  where we consider values of  $\mu_\theta \in [-1, 1]$  and  $\epsilon \in [0, 0.5]$ , each varying by 0.05. Baseline values were used for the other parameters ( $\sigma_\theta = \sigma_\beta = 1$ ;  $\mu_\beta = 0$ ). The best-fitting model is that with the lowest sum of squared-errors ( $SSE$ ) between each individual's estimate and the model's prediction. For the no feedback condition, the best fitting model was parametrized by  $\epsilon = 0.4$  and  $\mu_\theta = -0.1$  ( $SSE = 656.68$ ), as shown in Figure 1.5a. Thus participants in the no feedback condition were characterized by low prior perceived ability as well as inaccuracy at estimating their performance on each problem. For the condition with feedback, the best fit model was parametrized by  $\epsilon = 0$  and  $\mu_\theta = -0.35$  ( $SSE = 146.04$ ), as seen in Figure 1.5b. The results from the feedback condition were thus best captured by a model with a low prior on ability and seemingly perfect accuracy guessing performance on each problem, as should be expected given that self-assessments are heavily impacted by people's 'self-concepts' (Ehrlinger & Dunning, 2003).

## Discussion

In this first study, we observe that the parameters are acting in ways we would expect: specifically, delivering immediate feedback effectively reduces the  $\epsilon$  parameter representing error detection following each problem to 0. Participants did not estimate their score perfectly following the task even with exact feedback owing to a low prior over perceived ability, or  $\mu_\theta$ . In Section 1.4, we adapt the model to demonstrate its effectiveness in hypothesis testing, specifically a very wide-spread and controversial theory about metacognition: the Dunning-Kruger effect.

## 1.4 The Dunning-Kruger effect

Studies seem to show poorer calibration for low than for high performers, which could indicate worse metacognitive ability among low performers relative to others (the "Dunning-Kruger" effect). By adapting our rational model of self-assessment, we show in this section that such an effect could be produced by two psychological mechanisms, either in isolation or conjunction: influence of prior beliefs about ability, or a relation between performance and skill at determining correctness on each problem. To disentangle these explanations, we conducted a large-scale replication of a seminal paper (Kruger & Dunning, 1999) with approximately 4,000 participants in each of two studies. Comparing the predictions of two variants of our rational model provides support for low performers being less able to estimate whether they are correct in the domains of grammar and logical reasoning.

Researchers have used analytical methods to contest the existence of the Dunning-Kruger effect. In one instance, researchers sought to demonstrate how measurement error could account for most of the Dunning-Kruger effect by formalizing this phenomenon in terms of true skill and overconfidence (Feld et al., 2017). In their sample, they found a relation between estimated and true ability, but it was much weaker than expected. Krajč and Ortman

(2008) set out to demonstrate a problem with biased subject pools in papers demonstrating the Dunning-Kruger effect (Kruger & Dunning, 1999), but did not compare their model to actual data. Schlösser et al. (2013) refuted this account by showing its statistical assumptions were inconsistent with data from the original paper by Kruger and Dunning (Kruger & Dunning, 1999). In contrast to these previous efforts, we construct different versions of a computational model of self-assessment that instantiate the theories we aim to distinguish between. Fitting these competing models to data and performing a model comparison allows us to evaluate which theory is a better explanation of the data. In doing so, we are able to explore subtle differences in the assumed mechanisms behind the effect – specifically whether we need to postulate a metacognitive deficit among underperformers (as originally suggested by Kruger and Dunning) or whether assumptions about priors (as instantiated in previous Bayesian models (Healy & Moore, 2007)) are sufficient.

These competing explanations engage with a different point than previous controversy over the Dunning-Kruger effect. Krueger and Mueller (2002) argued that the effect could be a statistical artifact of regression to the mean, or a general poor calibration among participants that leads most estimates to converge around the mean, paired with a “better-than-average” effect. Kruger and Dunning (2002) responded that regression to the mean did not adequately explain their results after a reanalysis of their original data. Through additional studies in more real-world settings and a meta-analysis, Ehrlinger et al. (2008) concluded that the Dunning-Kruger effect was the best interpretation of these new data. By correcting for potential measurement error, they compared this hypothesis to other prominent accounts including regression to the mean (Krueger & Mueller, 2002) and a task difficulty account in which perceived difficulty of the task produces a general trend of over- or under-estimation (Burson et al., 2006). While this debate has helped establish that the results often explained by Dunning-Kruger effect are not due simply to a statistical artifact of regression to the mean produced by the data, there remains the possibility that the internal regression to the (prior) mean produced by rational Bayesian inference – a psychological rather than statistical explanation for the data – is driving the effect.

To tease apart these two possible psychological explanations for the Dunning-Kruger effect, we use the mathematical framework for specifying rational models of self-assessment described above and modify the model to formalize the nuances of the competing psychological theories. Using this framework, we show that the two different theories predict different forms for the relation between performance and self-assessment. We run a pair of large-scale replications to more precisely identify the actual form of this relation, which has typically been measured relatively coarsely in previous research. This more fine-grained picture allows us to identify which psychological explanation best accounts for people’s errors in self-assessment.

## Performance-dependent estimation

The model described in Section 1.2 assumed that everyone is equally adept at knowing whether their responses were correct or incorrect, consistent with explanations that assume



similar metacognitive abilities on average regardless of true ability (Burson et al., 2006; Krueger & Mueller, 2002). The idea that poor performers are “metacognitively impaired” in comparison to high performers put forth by Kruger and Dunning (1999) can be captured by extending the model so that instead of an  $\epsilon$  parameter that is identical across all participants, there is instead a separate  $\epsilon_p$  associated with each person  $p$  that may differ across individuals in relation to their true ability ( $\mu_\theta$  remains constant across all individuals in this model). Namely, those who perform poorly will guess their performance on a single problem less accurately than those who perform well:  $\epsilon_p$  for lower performers will be larger than for higher performers.

One way to make  $\epsilon_p$  dependent on person  $p$ ’s ability is to use a simple linear function such that  $\epsilon_p$  varies linearly with score, which serves as our closest approximation of true ability:

$$\epsilon_p = \epsilon_0 - \alpha \cdot \frac{\sum_i x_i}{n}, \quad (1.6)$$

with slope  $-\alpha$ , intercept  $\epsilon_0$ , and maximum achievable score  $n$ . Thus  $\frac{\sum_i x_i}{n}$  represents someone’s scaled score. In the example in Figure 1.3c, we vary  $\epsilon_p$  gradually according to Equation 1.6 with  $\epsilon_0 = 0.5$ ,  $\alpha = 0.1$  and then with  $\epsilon_0 = 0.4$ ,  $\alpha = 0.3$ . In these examples, the worst performers have an  $\epsilon_p$  of 0.5 (at chance) or 0.4 and the highest performers have a slightly lower  $\epsilon_p$  (0.4 in the first example and 0.1 in the second), meaning they are more accurate in their beliefs about correctness. This produces greater overestimation at lower true scores.

In order to evaluate whether performance-dependent estimation – the explanation that Kruger and Dunning gave for their results (Kruger & Dunning, 1999) – actually occurs, we can compare how well two versions of the model that capture competing hypotheses about self-assessment fit the human data. The two variants of our Bayesian model are (1) a Bayesian inference model (where  $\epsilon$  is independent of true ability) which represents a simple explanation of the data and (2) a performance-dependent estimation model (where  $\epsilon_p$  is dependent on person  $p$ ’s score). It is worth noting that the two models differ most in their predictions about people’s beliefs in the tails of the plots, so experimentally differentiating between theories will require recruiting sufficient numbers of participants with extreme scores.

Previous arguments in favor of the Dunning-Kruger effect have performed an analysis by grouping participants based on quartile of performance (Ehrlinger et al., 2008; Krueger & Mueller, 2002). Having our distinct models will help tease apart possible interpretations of the data, but the way the data has been looked at previously is also not sufficiently high-resolution to demonstrate differences between the model fits because it only provides four data points (one for each quartile) to compare to a model. Therefore we compare each individual data point to the model rather than grouping the data by quartile of performance.

These considerations argue for conducting a large-scale replication of previous experiments on the Dunning-Kruger effect, which did not have sufficiently large samples to distinguish between the two explanations instantiated in our models. To address this, we conducted replications of two studies from Kruger and Dunning (1999) in which participants

solved a series of 20 multiple-choice questions about either grammar or logical reasoning and estimated their score following the assessment.

## Methods

This research complies with ethical obligations put forth by the UC Berkeley Internal Review Board. Informed consent was obtained from all human participants.

### Participants

One criticism of Kruger and Dunning is their use of a convenience sample composed of college undergraduates at elite universities (Krajč & Ortmann, 2008). We allay this concern by recruiting participants from Amazon’s Mechanical Turk (<https://www.mturk.com/>), where wider ranges of ages and educational backgrounds are represented (Mason & Suri, 2012).

We selected our sample size of 4000 participants per study by conducting a power analysis based on the effect sizes observed in preliminary studies of 250 participants each. To do so, we simulated increasing sample sizes and observed where the curves started to level out (see preregistration <https://osf.io/k28je> from March 6, 2019). With 4000 samples per bootstrap, 85.4% of the grammar data simulations and 81.6% of the logical reasoning simulations resulted in favor of the Dunning-Kruger effect. We declared a stopping rule (ending data collection once 4000 responses were collected or after 10 days) as well as all exclusion criteria in our preregistration. This power analysis was conducted with a previous version of the model that did not contain a guessing parameter. We re-ran our models to include this guessing parameter following a suggestion made by one of our reviewers.

Participants were paid \$2 to participate in the grammar study or \$3 for the logical reasoning study. We decided against awarding a bonus for more accurate self-assessment judgments because Kruger and Dunning also paid only a flat rate and because in other work, adding incentives did not reduce the amount of inaccuracy (Ehrlinger et al., 2008; Sanchez & Dunning, 2018). Since sample sizes were so large, participants were allowed to participate in both studies, which were presented as separate HITs on MTurk.

For the grammar study, there were 3860 responses. Of these, 164 failed the instructional manipulation, and so were excluded from analyses. 80 spent under 5 minutes on the task. There were 3 IP addresses used three times and 43 used twice, so these 95 responses were additionally excluded. Six responses had no associated IP address and were thus also excluded. This resulted in a total of 3515 responses viable for analyses (1698 self-identifying men, 1780 self-identifying women; mean age = 36.54; range: 18–88).

The logical reasoning study brought in 3901 complete responses. There were 154 failed attention checks, 77 participants who spent under 5 minutes and 55 repeated IPs (51 with 2, 2 with 3, 1 with 4 and 1 with 6 responses) and 9 without an IP address. This left 3543 responses for analyses, 1778 self-identifying women and 1731 self-identifying men; mean age = 36.59, range: 18–81).

## Materials

The closest approximation of the original 20 logical reasoning problems and 20 grammar questions from Kruger and Dunning (1999) that we could obtain from the original authors were made into surveys on [Qualtrics](#). We use their original materials to make a more compelling case for whether or not this effect exists.

These questions consist of multiple-choice items with five possible responses. In the logical reasoning test, participants were told “You will be presented with brief passages or statements and will be required to evaluate their reasoning or determine what inferences you can logically draw from the passage. In each case, select the best answer choice, even though more than one choice may present a possible answer.” In the grammar task, participants read the following instructions: “In each question, some part of each sentence is underlined; sometimes the whole sentence is underlined. Five choices for rephrasing the underlined part follow each sentence; one choice repeats the original, and the other four are different.”

## Procedure

Before beginning each study, all participants read a set of instructions and were asked two content-based questions about the instructions. They were given two opportunities to answer these questions correctly and were excluded from analyses if they failed the instructional manipulation on both attempts, as indicated in our preregistration. These exclusions were intended to prevent the low-performing individuals in our analyses from being composed of inattentive participants. Both before and after problem-solving, all participants rated their absolute performance (“how many of the 20 logical reasoning/grammar problems will/did you answer correctly?”), their relative performance as a percentile ranking out of 100 (“compared to other participants in this study, how well do you think you will do/did you do?”), the difficulty of the task for themselves, and the difficulty for others (both on a scale from 0 to 10). On each task, the 20 questions were presented in a randomized order as were the five multiple-choice solutions. All problems and self-assessment questions required a response for the participant to move ahead. The absolute performance ratings were displayed as a drop-down menu, the relative performance ratings as a sliding bar, and the difficulty ratings as horizontal multiple choice questions. At the conclusion of the study, participants were directed to a short demographics questionnaire where they optionally answered questions about their age, gender, race, and educational background. All analyses in this section are based solely on the absolute ratings of performance made after the test.

## Model fitting

To fit the Bayesian inference and performance-dependent estimation models to the data, we compare model predictions to participants’ estimates of their scores relative to their true score since “perceived performance” on the task is how we are operationalizing “perceived ability.” To generate a set of simulations with varying parameter values, we performed a grid search over  $\mu_\theta$  and  $\epsilon$  for the Bayesian inference model and these two parameters along with  $\alpha$

for the performance-dependent estimation model such that values of  $\mu_\theta \in [-1, 1]$ ,  $\epsilon \in [0, 0.5]$  and  $\alpha \in [0, 0.5]$  were considered. We do not consider values of  $\epsilon$  greater than 0.5 or worse than chance because we assume people are not systematically biased to believe they are incorrect when they are correct. We took steps of 0.05 for each parameter which resulted in a total of 41 considered values of  $\mu_\theta$  and 11 values each of  $\epsilon$  and  $\alpha$ , which produced 451 Bayesian inference model predictions and 2706 performance-dependent estimation model predictions. Note that  $\alpha$  cannot be lower than  $\epsilon_0$  in the performance-dependent estimation model as this would result in negative values of  $\epsilon_p$  which is impossible since this is a probability, so there will not be  $11 \times 11 \times 41 = 4961$  performance-dependent estimation model predictions, as one might expect. Baseline values were used for the other parameters ( $\sigma_\theta = \sigma_\beta = 1$ ;  $\mu_\beta = 0$ ). We ran five MCMC chains of 10,000 iterations where we marginalized over item difficulties ( $\beta_i$ ) in order to examine only perceived ability (removing the first 1,000 iterations each time for burn-in). We then took the mean of all remaining 9,000 sampled  $\theta_p$  values to generate predictions for each pairing or triplet of parameters. To better estimate the posterior, we calculated the mean across all samples from all five chains to compare to the human data. We calculated all these values for the performance-dependent estimation model and used the values from the Bayesian inference model for the cases when  $\alpha = 0$ . We then converted each simulated ability parameter value,  $\theta$ , generated by the models into a probability of a correct response using Equation 1.3. To transform this into estimated total score, which is what data we have to compare to the model, we then multiply this probability by the maximum score (20 in our data).

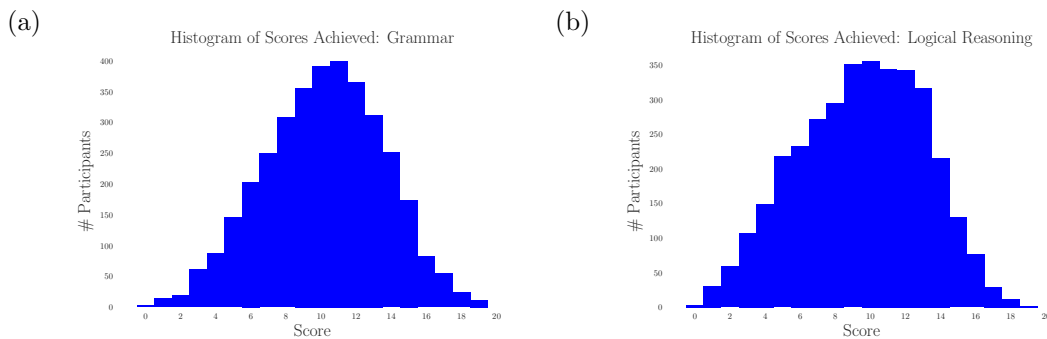


Figure 1.6: Histograms of scores achieved in (a) the grammar study and (b) the logical reasoning study.

## Results: Grammar study

For the 3515 participants who solved the grammar problems included in analyses (1698 self-identifying men, 1780 self-identifying women, and the remainder other or unspecified; 2560 White, 304 Black or African American, 246 Asian/Asian American, 215 Hispanic or Latino, and the rest selected multiple categories, other, or unspecified), the mean completion

time was 19.61 minutes. On average, participants scored 10.17 out of 20 ( $sd = 3.40$ ) (see Figure 1.6a for a distribution of the achieved scores) and the mean estimated score was 12.49 ( $sd = 3.91$ ). In the original study by Kruger and Dunning, participants scored an average of 13.3 and estimated the number correct at an average of 15.2. We attribute the lower performance in our study to the fact that their participants were undergraduate students whereas participants in this study had a wide range of backgrounds. The original authors did not report standard deviations. The overconfidence of the lowest scoring participants appeared substantial as can be seen in Figure 1.7. Participants made an average percentile estimate of 58.48 ( $sd = 20.57$ ), rated the task with a difficulty of 5.57 ( $sd = 2.26$ ) for themselves and 6.16 ( $sd = 1.84$ ) for other participants both out of 10, where 10 represents the highest difficulty.

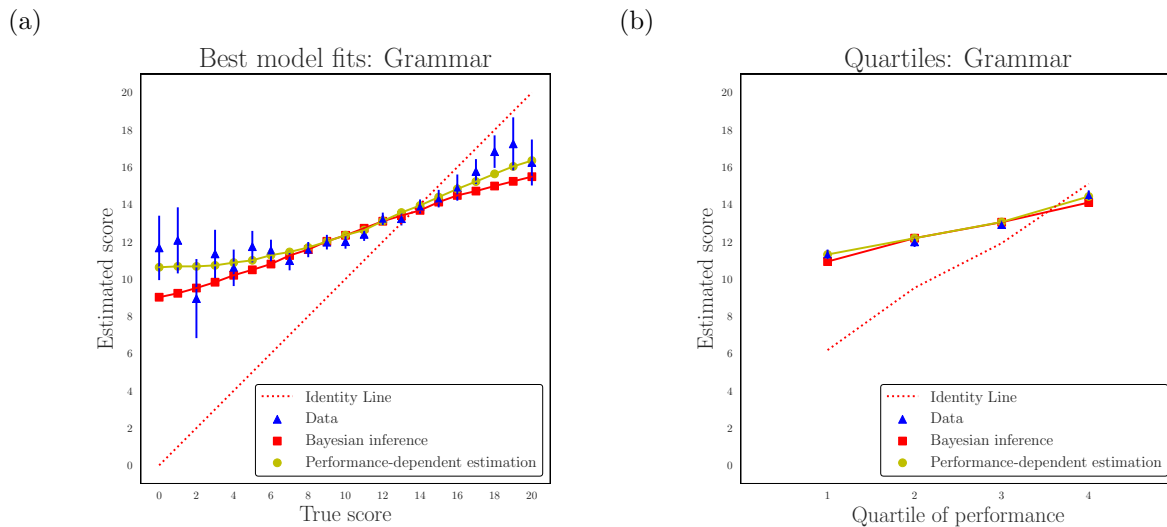


Figure 1.7: Results and best fitting models for the grammar study displayed by (a) score and (b) quartile of performance. The model where estimation accuracy is independent of score, or the ‘Bayesian inference’ model is parameterized by  $\epsilon = 0.4$  and  $\mu_\theta = 0.05$  while the performance-dependent model is parameterized by  $\epsilon_0 = 0.45$ ,  $\mu_\theta = 0.05$ , and  $\alpha = 0.1$ . Error bars represent 95% confidence intervals.

To fit the model to the data, we compare model predictions to participants’ estimates of their scores relative to their true score. Grouping the self-assessments by true score instead of by quartiles, as done in previous research, shows substantially more variability in performance. Comparing the data to the grid search of model estimates, the best fit Bayesian inference model (where estimation ability is independent of score) was parameterized by  $\epsilon = 0.4$  and  $\mu_\theta = 0.05$  ( $SSE = 49591.58$ ) while the performance-dependent estimation model (the model that instantiates the Dunning-Kruger effect) was parameterized by  $\epsilon_0 = 0.45$ ,  $\mu_\theta = 0.05$ , and  $\alpha = 0.1$  ( $SSE = 49073.04$ ). The best fit models are presented in Figure 1.7a alongside the associated data, grouped by actual score.

To compare these competing models, as described in our analysis pipeline, we calculated the Bayesian information criterion (*BIC*) for each model. The *BIC* for the Bayesian inference model with constant  $\epsilon$  ( $BIC = 19303.07$ ) was higher than that of the model with  $\epsilon$  dependent on score ( $BIC = 19274.29$ ). Because these are nested models such that the performance-dependent model contains one more parameter than the Bayesian inference model, we performed a likelihood ratio test which is equivalent to a  $\chi^2$  test with one degree of freedom. This yielded  $\chi^2(1) = 36.95, p < 0.001, \phi = 0.62$ , which far exceeds the threshold of 3.84 required to be significant. Thus, we have strong evidence to prefer the more complex model in this case, which is the performance-dependent model. We additionally computed the log Bayes Factor, the logarithm of the ratio of the marginal likelihood of the performance-dependent hypothesis to the likelihood of the Bayesian inference hypothesis, and obtained a value of 16.14, another indication of strong evidence for the performance-dependent estimation hypothesis. Given that the parameters of our model correspond to the intercept ( $\mu_\theta$ ), slope ( $\epsilon$ ), and curvature ( $\alpha$ ) of the data, we additionally fit linear and quadratic models to the data, finding that the quadratic model provided a better fit compared to the linear model ( $F(1) = 34.25, p < 0.001$ ). When no exclusions are applied (as described in the methods section) and data from all participants is included, results are substantially the same as those presented here.

We show in Figure 1.7b a fit of the data to the model by quartile of performance, as done in previous work. Specifically, we group participants in quartiles based on their scores and plot their average self-assessment judgments and overlay average model values for the best fit models. We see in this depiction that quartiles do not show the clear distinction between the two models that can be seen in the alternative plot.

## Results: Logical reasoning study

For the 3543 participants included in analyses who solved the logical reasoning problems (1778 self-identifying women and 1731 self-identifying men; 2553 White, 350 Black or African American, 246 Asian/Asian American, 196 Hispanic or Latino), the average completion time was 23.48 minutes. The mean score was 9.45 out of 20 ( $sd = 3.59$ ) and the mean estimated score was 10.86 ( $sd = 4.05$ ). In the original study by Kruger and Dunning, participants scored an average of 12.9 and estimated an average of 13.3. As in the grammar study, we observe considerable overconfidence by the worst performers (see Figure 1.8). Participants made an average percentile estimate of 52.44 ( $sd = 20.84$ ), rated the task with a difficulty of 6.78 out of 10 ( $sd = 2.04$ ) for themselves and 6.92 ( $sd = 1.79$ ) for other participants.

Comparing the data to the grid search of model estimates, the best fit Bayesian inference model was parameterized by  $\epsilon = 0.45$  and  $\mu_\theta = -0.1$  ( $SSE = 55801.41$ ) while the performance-dependent estimation model was parameterized by  $\epsilon_0 = 0.5$ ,  $\mu_\theta = -0.15$ , and  $\alpha = 0.15$  ( $SSE = 54912.32$ ), seen in Figure 1.8a. A quadratic model again was a better fit to the logical reasoning data as opposed to a linear model ( $F(1) = 56.87, p < 0.001$ ).

The *BIC* for the model with constant  $\epsilon$  ( $BIC = 19846.55$ ) was again higher than that of the model with  $\epsilon$  dependent on score ( $BIC = 19797.82$ ). A likelihood ratio test comparing

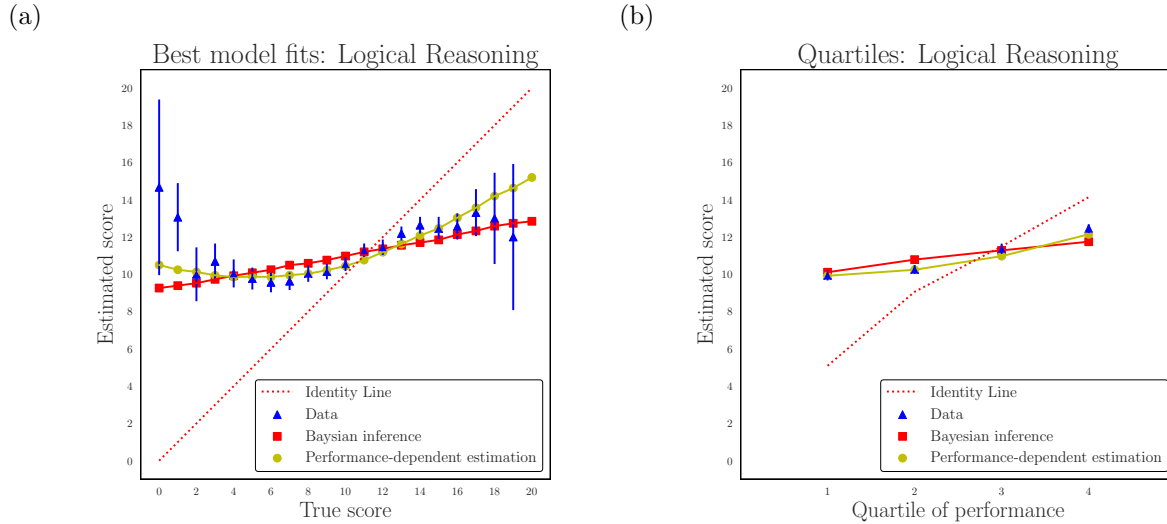


Figure 1.8: Results and best fitting models for the logical reasoning study displayed (a) by true score and (b) by quartile of performance. The simpler Bayesian inference model is parameterized by  $\epsilon = 0.45$  and  $\mu_\theta = -0.1$  while the performance-dependent model has  $\epsilon_0 = 0.5$ ,  $\mu_\theta = -0.15$ , and  $\alpha = 0.15$ . Error bars represent 95% confidence intervals.

the models was significant ( $\chi^2(1) = 56.91, p < 0.001$ ). Just as for the grammar study, when no exclusions are applied, results are substantially the same. A log Bayes Factor calculation yielded a value of 26.15. Thus, we again have sufficient evidence to prefer the more complex performance-dependent estimation model over the Bayesian inference model.

## Discussion

The observation that poor performers overestimate their ability could be given two possible psychological explanations: that it is a mere result of rational Bayesian estimation, or that it reflects a genuine decreased sensitivity to errors among low performers. In this section, we formalized these competing accounts as mathematical models and ran large-scale replications to identify the form of the function relating self-assessment to performance. Though even with such large samples there is limited resolution in the tails of the score distributions (as fewer participants obtained the most extreme scores), the model assuming reduced sensitivity among low performers is a statistically better fit to the data than the model assuming simple Bayesian inference.

The rational model we developed gives us a framework for testing even more hypotheses beyond whether or not there is a relation between true ability and accuracy of perceived ability. Specifically, it offers a precise specification of the form of the function relating true ability to perceived ability, which allows for testing differences at task-specific and individual levels. This model enables future work that could include testing out a combination of priors

over each parameter or identifying potential differences by domain, age, or other demographic variables. We see this as a promising approach for gaining more insight into what underlies learners' metacognitive abilities. The flexibility of this model can also allow for fitting to different types of metacognitive data, including relative perceived performance (here we focused on absolute estimates of performance) or incorporating new parameters to formalize other areas of research into metacognition, such as the bias blind spot (Pronin et al., 2002). In the next section, we further enhance the model to analyze self-assessment on an individual level and over time.

## 1.5 Sequential self-assessment

There has been some debate as to whether self-assessment ability should be measured via one-off judgments (as in these studies) or by aggregating an individual's confidence judgments made before solving each problem (Krueger & Mueller, 2002). In this section, we implement a version of the model that predicts a participant's confidence judgments over time to provide a more accurate representation of each individual's beliefs about their ability.

People's assessment of their ability varies in whether it is measured once following a task or sequentially via confidence judgments recorded throughout. Multiple models have been developed to predict one-off judgments of performance, which have often distinguished between peoples' biases about their general ability in a domain and their sensitivity to correctness. We modify our rational model to predict *sequential* self-assessment which allows for making predictions about each individual separately—unlike in the one-off case which looks exclusively at the population level—and for identifying, in addition to bias and sensitivity, the extent to which individuals' beliefs are responsive to their most recent evidence over the course of a task. We adapt the model to fit multiple datasets: four multiple-choice tasks (two trivia and two math), two non-multiple-choice math tasks, and math and reading tasks with students in India. With a diverse set of problem types and age groups, we show that bias, sensitivity, and responsiveness, all represented as separate parameters in the probabilistic model, vary meaningfully across participants and demonstrate an additional dimension of versatility of our rational framework.

Results from studies of self-assessment show that multiple individual-level characteristics may cause differences in calibration to performance on a task. Ehrlinger and Dunning (2003), for example, proposed that a person's "self-concept," their beliefs about their overall skill in a domain, is foundational to their beliefs about their performance on a specific task. These views about the self are likely to be very different, especially in domains like math where variable self-concepts have been widely documented (Seaton et al., 2014). Thus, analyzing individuals' confidence judgments can assist in capturing even smaller individual-level differences across domains.

In Ehrlinger and Dunning (2003), men and women performed comparably on a science test, but women underestimated their ability compared to men. In a similar vein, Correll (2001) argued that cultural beliefs about gender and math ability harmed girls' perceptions



of their competence. A model of consecutive self-assessment can still account for these sorts of group-level differences (in addition to individual differences) because this type of analysis will specify the full distribution of individual parameters within groups.

Some have argued that aggregating confidence judgments is a superior method to requesting a single judgment per participant (Krueger & Mueller, 2002), but really these are different types of judgments that may both be important in distinct ways: single judgments following a task are useful for a person to determine what they will be capable of in the future, while confidence judgments, which convey someone’s tracking of their ability, are necessary for determining which more specific skills require targeted study. On a linear equation-solving task, for example, self-assessments made following the task will be used by someone to decide whether to keep practicing at their level or to move on to quadratic equations or another more advanced topic. Tracking performance throughout this task, on the other hand, will provide insight into whether there is a specific algebra skill they are having trouble with (e.g., distribution or combining terms). We will be able to analyze how different individuals update their perceptions of their ability throughout a task with this formal model of sequential self-assessment.

## Modeling Sequential Confidence Judgments

In this section, we tailor our computational model to generate a more accurate representation of the form of the function that links each individual participant’s confidence throughout a task to their actual performance. We start by describing the revisions to our rational model and the predictions it will make under a variety of circumstances. Following this, we fit alternative versions of the model to multiple sets of data: adults online solving multiple-choice trivia and math questions (studies 1–4), adults online solving math problems in an interactive tutor (studies 5–6), and students in India answering math and English grammar questions (study 7).

### Model assumptions

This model makes the same assumptions as the model described above, but at the problem-by-problem level rather than at task completion and treats individuals as making confidence judgments that are consistent with Bayesian inference about their ability. We assume a rational agent makes each judgment based on their beliefs about their ability so far (which includes both their prior beliefs before beginning the task and their performance on already solved problems), the task’s difficulty, and individuals’ sensitivity to their correctness on each problem. We assume that the priors over a person  $p$ ’s perceived ability before beginning the task ( $\theta_{p,1}$ ) and the difficulty of each problem ( $\beta_t$ ) are normally distributed and we compute the probability of a person’s correctness at a particular time point  $t$  ( $X_{p,t}$ ) which is dependent on perceived ability and difficulty parameters up to and including the current time point.<sup>2</sup>

---

<sup>2</sup>For ease of reading, we drop the  $p$  in the subscripts, as we assume that the model is run separately for each individual.

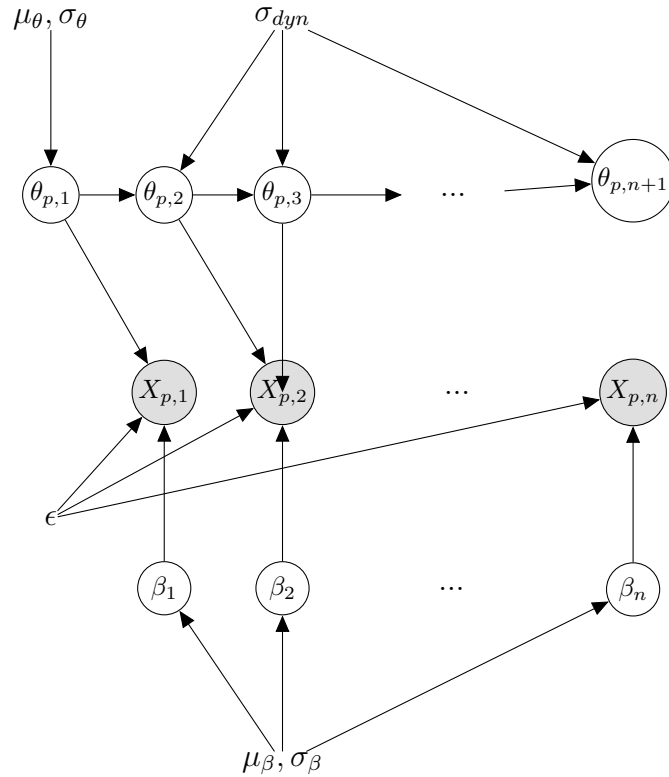


Figure 1.9: Graphical representation of the model: each observed item  $X_{p,t}$  is influenced by latent variables  $\beta_t$  (difficulty of problem at time  $t$ ) and  $\theta_{p,t}$  (perceived ability of person  $p$  at time  $t$ ) as well as a constant  $\epsilon$  (ability to determine correctness). The difficulties of all problems  $\beta_t$  and the prior over perceived ability  $\theta_{p,1}$  are drawn from normal distributions with means  $\mu_{\beta}$ ,  $\mu_{\theta}$  and standard deviations  $\sigma_{\beta}$ ,  $\sigma_{\theta}$ . Each subsequent  $\theta_{p,t>1}$  is drawn from a  $N(\theta_{p,t-1}, \sigma_{dyn})$ .

At each time point, we compute the probability of responding to a problem correctly or not given the person's prior perceived ability and the difficulty of this problem which acts as the *likelihood* in this Bayesian computation. To do so, we, as above, use the 1-parameter IRT model (Equation 1.1) and incorporate the error parameter  $\epsilon$ , which represents the probability of incorrectly guessing performance on an individual problem (Equation 1.2).

Because we are modeling sequential confidence judgments, perceived ability at time  $t$  depends on perceived ability at all problems up to  $t - 1$ . At time  $t = 1$ ,  $\theta_1$  is drawn from a normal distribution with mean  $\mu_{\theta}$  and standard deviation  $\sigma_{\theta}$ . At all subsequent time points, the dynamics governing how  $\theta_t$  is related to  $\theta_{t-1}$  is specified by  $p(\theta_t|\theta_{t-1})$ . We assume a normal distribution for  $\theta_t$  centered at  $\theta_{t-1}$  and that the variance of this distribution,  $\sigma_{dyn}$ , is a parameter of the model which controls how reactive people are to their most recent data.

We additionally need to adjust the likelihood function for all  $t > 1$  to incorporate all previous problems, so we define the probability of responding to a question correctly given

someone's perceived ability and the difficulty of the problems so far as the product of all likelihoods up through the current problem. We combine this likelihood with a person's previous ability belief  $p(\theta_{t-1}, X_{1:t-1})$  and the dynamics of perceived ability  $p(\theta_t|\theta_{t-1})$  via Bayes' rule to compute each person's posterior beliefs about their own ability on each problem at time  $t$ :<sup>3</sup>

$$\begin{aligned}
 p(\theta_t|\theta_{1:t-1}, X_{1:t}) \propto & \int_{\beta_k} p(X_k|\theta_k, \beta_k, \epsilon)p(\beta_k)d\beta_k \\
 & \cdot p(\theta_{t-1}|X_{1:t-1}) \\
 & \cdot p(\theta_t|\theta_{t-1}).
 \end{aligned} \tag{1.7}$$

A graphical representation of the model dependencies is shown in Figure 1.9. In model simulations presented next, we vary prior beliefs about ability via  $\mu_\theta$ , the likelihood by increasing  $\epsilon$ , and the dynamics of perceived ability through changes to  $\sigma_{dyn}$ .

## Generating model predictions

Because the integrals in Equation 1.7 are intractable to calculate exactly, we require an algorithm that can dynamically update the posterior on  $\theta_t$  in light of new data. We use a standard sequential Monte Carlo method known as a *particle filter* (for an overview, see Doucet and Johansen (2009)). To produce a model simulation for a given set of parameters with a particle filter, we follow Algorithm 1 to generate posterior distributions of each  $\theta_t$  given  $X_{1:t}$ . At each time point  $t > 1$ , we represent the posterior with a set of  $n$   $\theta_t$  values, or *particles*, from a probability distribution based on the particles at the previous time  $t - 1$ . Each vector of particles has a normalized set of  $n$  weights equal to the cumulative likelihoods as in Equation 1.7. If the variance of these weights is large we want to remove particles with low weights and multiply ones with higher weights, so we resample  $n$  new particles using the normalized weights as a distribution, and then adjust the weights to be uniform. The likelihoods then accumulate again as new data come in. For each time point, we convert the

---

<sup>3</sup>To obtain estimates of people's inferences over their ability, we marginalize over the difficulty parameters ( $\beta_t$ ).

---

### Algorithm 1: PARTICLE FILTER ALGORITHM

---

1. **Sample** a set of  $n$  particles  $\theta_t^i$ , ( $i = 1 \dots n$ )
    - (a) If  $t = 0$ : from the prior  $N(\mu_\theta, \sigma_\theta)$ ;
    - (b) If  $t \geq 1$ : from the particles at time  $t - 1$   $p(\theta_t^i|\theta_{t-1}^i)$ ;
  2. **Compute weights**  $w_t^i(\theta_{1:t}^i)$  which are equal to the product of the previous likelihoods  $p(X_{1:t}|\theta_{1:t}^i)$  since resampling:  $w_t^i(\theta_{1:t}^i) \propto w_{t-1}^i(\theta_{1:t-1}^i)p(X_t|\theta_t^i)$ ;
  3. **Resample**: if resampling criterion satisfied, resample  $\{W_t^i, \theta_{1:t}^i\}$  to obtain  $n$  new equally weighted particles  $\{\frac{1}{n}, \widehat{\theta_{1:t}^i}\}$
-

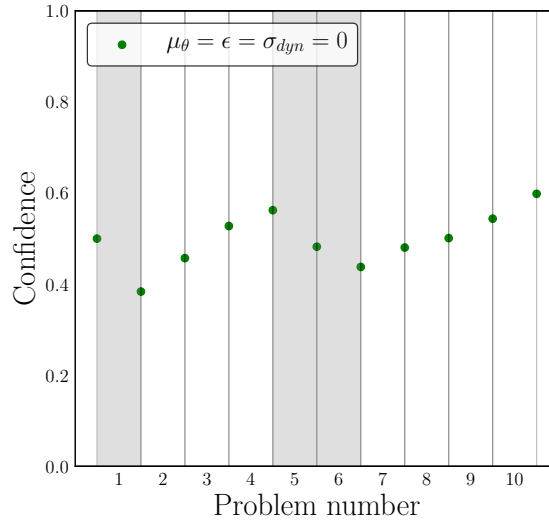


Figure 1.10: Model predictions in a toy example where participants solve 10 problems in a baseline model ( $\mu_\theta, \mu_\beta = 0, \sigma_\theta, \sigma_\beta = 1$ , and  $\epsilon = 0$ ). Each point shows the weighted average of the posterior distribution on each  $\theta_t$ , corresponding to confidence judgments at each time point. The first judgment made before the first problem is simply based on the prior over ability. The shaded areas demarcate problems solved incorrectly, while the unshaded regions correspond to correctly solved problems.

vector of associated particles into probabilities of a correct response via Equation 1.1 (which is a sigmoid function) and take the weighted average of all particles to obtain a model prediction of each confidence judgment between 0 and 1 over time (see Figure 1.10 for a baseline example of model predictions where we set all adjustable parameters  $\epsilon$ ,  $\mu_\theta$ , and  $\sigma_{dyn}$  equal to zero).<sup>4</sup>

### Changing the prior

When we adjust the prior over a person’s beliefs about their ability ( $\mu_\theta$ ), we observe changes to their overall beliefs. In the toy example in Figure 1.11a, when we assign a higher mean over ability ( $\mu_\theta = 1$ ), confidence judgments tend to be higher overall. Shifting the prior mean downward ( $\mu_\theta = -1$ ) most depresses confidence judgments early on, when the person has limited data from the task, but as they have more experience, their estimates of their ability become more similar to the case with the higher prior mean.

<sup>4</sup>In all models implemented here, we opted to generate  $n = 10,000$  particles and used the Effective Sample Size as threshold for determining when to resample:  $ESS = (\sum_{i=1}^N (W_n^i)^2)^{-1}$ .

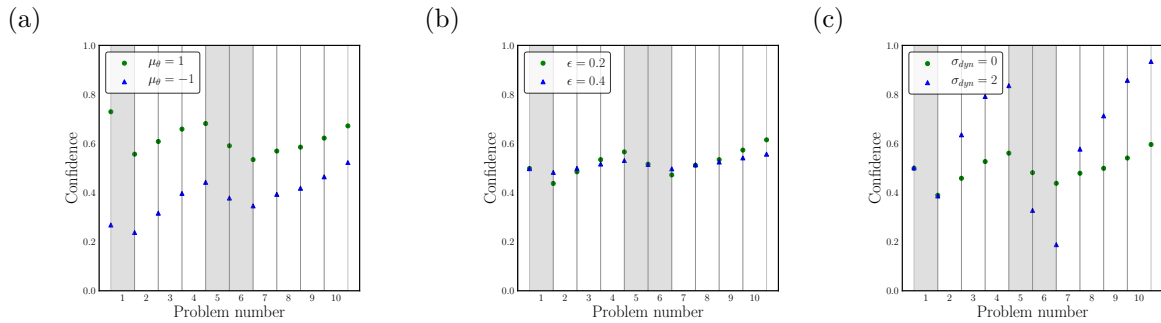


Figure 1.11: Model predictions on the same toy example as in Figure 1.10 for (a) when the mean on ability ( $\theta_{pt}$ ) is adjusted ( $\mu_\theta = 1$  or  $-1$ ), (b) when the error parameter  $\epsilon$  is adjusted ( $\epsilon = 0.2$  or  $0.4$ ), and (c) when ( $\sigma_{dyn}$ ) is zero or nonzero ( $\sigma_{dyn} = 2$ ). In all three plots, the parameters we are not adjusting are equal to zero.

### Changing the likelihood

Increasing  $\epsilon$  to include more error in individual judgments of correctness lowers confidence following correct responses and raises them after incorrect responses (see Figure 1.11b).

### Changing dynamics

By varying the dynamics of our model, we can control the extent to which participants learn from their entire set of previous responses. When  $\sigma_{dyn} = 0$ , a new particle at time  $t$  will be exactly the same as the old particle at time  $t - 1$  because the probability distribution places all the mass at the one location. As  $\sigma_{dyn}$  increases, there is a higher chance of the particle moving farther away from its previous location. In Figure 1.11c, we observe that a larger value of  $\sigma_{dyn}$  results in recent observations having a greater influence on beliefs. In particular, as seen in this example, when the simulated participant answers multiple problems correctly in a row, their confidence increases more steeply when  $\sigma_{dyn} = 2$  than when it is zero (and decreases similarly after multiple sequential incorrect answers).

### Fitting the models to data

To see how well models with different sets of parameter values compare to actual judgments, we designed studies to elicit sequential self-assessments from individuals. To see whether the parameter adjusting the dynamics of the model ( $\sigma_{dyn}$ ) is necessary to generate better model predictions, we fit each individual's data to a version of the model with no dynamic updating ( $\sigma_{dyn} = 0$ )—which we refer to as the *static* model—and a second version, the *dynamic* model, where  $\sigma_{dyn} > 0$ . This distinction is consistent with the idea of ‘mindset’ (Ehrlinger et al., 2016) such that the static model captures a fixed mindset (because  $\theta$  is fixed) while the dynamic model represents a growth mindset (since  $\theta$  varies and we can track how it changes).

## Model simulations

We compared each individual’s data to both the static model (where  $\sigma_{dyn} = 0$ ) and the dynamic model ( $\sigma_{dyn} > 0$ ). In order to make this comparison, we generated collections of model simulations for each participant given their sequence of correct and incorrect responses by performing a grid search over  $\mu_\theta$  and  $\epsilon$  for the static model and these two parameters along with  $\sigma_{dyn}$  for the dynamic model such that values of  $\mu_\theta \in [-3, 3]$ ,  $\epsilon \in [0, 0.5]$ ,<sup>5</sup> and  $\sigma_{dyn} \in [0.01, 6]$  were considered. We took steps of 0.05 for  $\epsilon$ , giving 11 possible values, steps of 0.2 for  $\mu_\theta$ , resulting in a total of 31 values,<sup>6</sup> and increasing steps of  $\sigma_{dyn}$ <sup>7</sup> for a total of 31 values, which produced 341 static model predictions and 10,571 dynamic model predictions. These parameter values were chosen based on initial attempts to model a subset of participants such that a representative spectrum of possible parameters were considered. For each participant, we compared their sequence of confidence judgments to each set of model predictions by calculating the sum of squared errors (*SSE*). We took the models with the smallest *SSE* amongst the static models and then the dynamic models to identify the parameters associated with the best fit model in each case.

## Post-processing for data visualization

To visually represent each individuals’ best-fit models, we generated plots that showed an individuals’ confidence judgments as a function of the problem number, indicating actual confidence ratings (cerulean triangles) between problems, and simulated values for the best-fit model (cyan circles for the static model and magenta squares for the dynamic model) overlaid (see Figure 1.12 for examples from the algebra multiple-choice study). In each plot, correct and incorrect responses are illustrated with blank or shaded regions, respectively.

## Studies 1–4: Multiple-choice trivia and math

We first conducted four studies to elicit confidence judgments from individuals on Amazon’s Mechanical Turk solving multiple-choice problems. These consisted of two trivia (college acceptance rates and years of Nobel prize awards adapted from Burson et al. (2006)) and two math tasks (algebraic equation solving and fraction arithmetic).

## Model specifications

Because all of these tasks included multiple-choice solutions, we need to account for the possibility of guessing, as described in the previous sections.

<sup>5</sup>Because  $\epsilon$  is a probability, we only consider values from 0 to 0.5 in our simulations because this value signifies guessing at chance.

<sup>6</sup>We explored more values of  $\mu_\theta$  when it seemed like there was a plateau, but did not find that it made much of a difference in precision of model estimates, especially in the cases where guessing was relevant.

<sup>7</sup>The values considered were (0.01, 0.02, 0.03, ..., 0.19, 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 2.5, 3, 4, 5, 6).

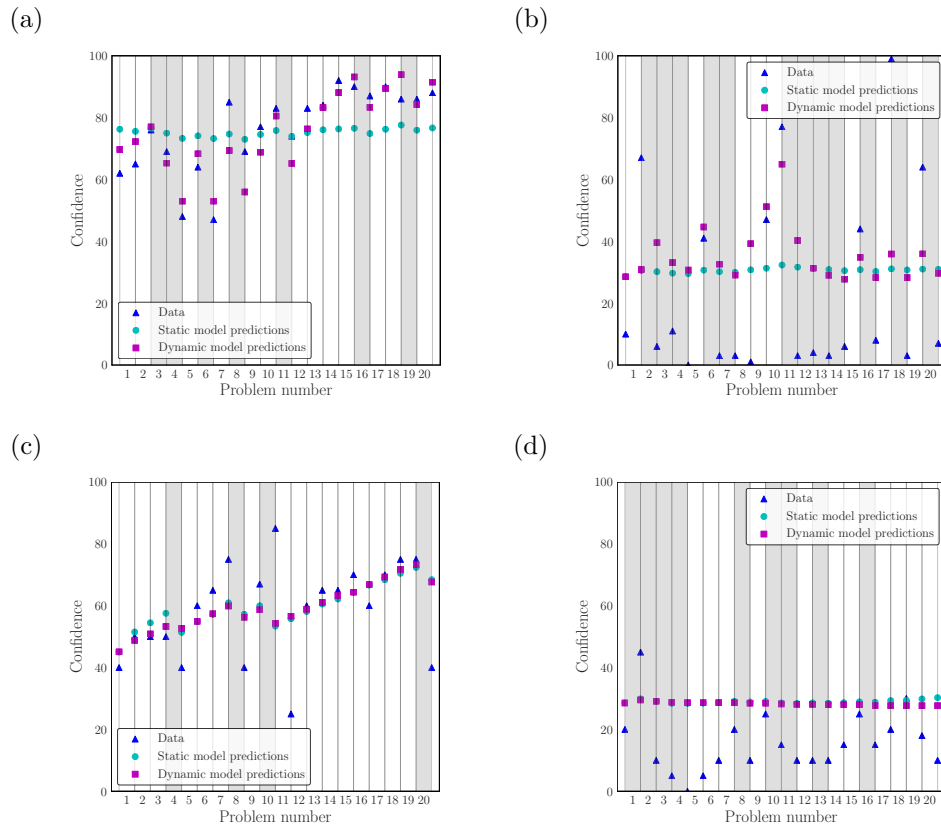


Figure 1.12: Model predictions for four example participants from the algebra multiple-choice task (Study 2) where shaded regions signify incorrect responses: (a) the best fit static model was parametrized by  $\epsilon = 0.25$  and  $\mu_\theta = 1$  and the best fit dynamic model had parameters  $\epsilon = 0.2$ ,  $\mu_\theta = 1$ , and  $\sigma_{dyn} = 0.6$ ; (b) was fit by  $\epsilon = 0.05$  and  $\mu_\theta = -1.7$  while the dynamic model is fit by  $\epsilon = 0.1$ ,  $\mu_\theta = -1.3$ , and  $\sigma_{dyn} = 2$ ; (c) the best fit static model for this participant was parametrized by  $\epsilon = 0.3$  and  $\mu_\theta = -0.3$  while the dynamic model is fit by  $\epsilon = 0.45$ ,  $\mu_\theta = -0.1$ , and  $\sigma_{dyn} = 1.6$ ; (d) was best fit by  $\epsilon = 0.25$  and  $\mu_\theta = -1.9$  for the static model and  $\epsilon = 0.35$ ,  $\mu_\theta = -2$ , and  $\sigma_{dyn} = 0.2$  for the dynamic model. The dynamic model fit the data significantly better than the static model in (a) and (b). In (c) and (d), the dynamic and static models fit the data equally well.

Studies 1–4 each consist of 20 problems and provide four possible solutions per question, making  $g$  equal to 0.25 for all simulations. The inclusion of a guessing parameter produces slightly different predictions than the basic model, as seen in Figure 1.13. The model predicts generally higher confidence judgments on multiple-choice tasks because there is a one in four chance of guessing the correct answer to any question.

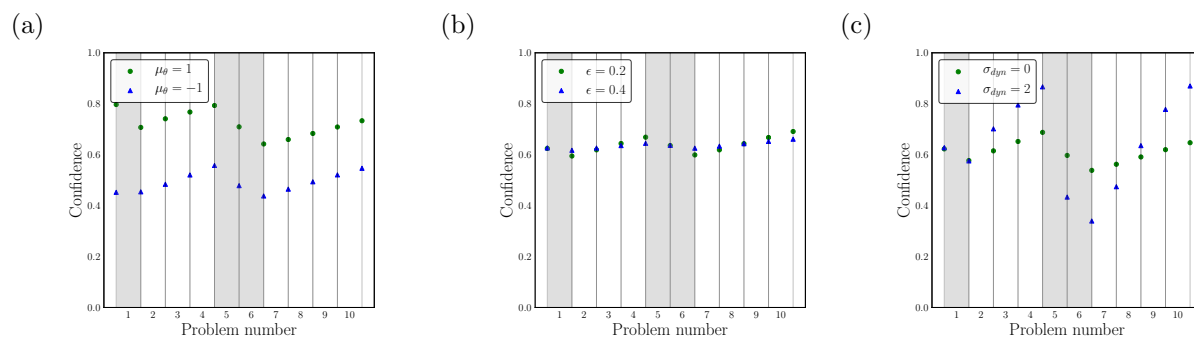


Figure 1.13: Model predictions on the same toy example as in Figure 1.11 with the same parameter values, but including a guessing parameter  $g = 0.25$ .

## Procedure

Participants were directed to one of four tasks where they solved 20 multiple-choice problems about college acceptance rates (Study 1), years when Nobel prizes were awarded (Study 2), algebraic equations (Study 3), or fraction arithmetic (Study 4). Prior to each question, participants were asked “You are about to solve a problem. How confident are you that you will solve it correctly?” as well as following the final problem (on a scale from 0 to 100), which resulted in a total of 21 judgments per person. All participants received the same problems in the same order, so that problem difficulty was preserved and individuals could be directly compared. To vary trivia difficulty, versions of the questions were pilot tested until the average score was about 10 out of 20 for both domains. The final version of the Nobel prize task consisted of asking when a Nobel prize in literature or Nobel peace prize was awarded and possible answers were 20-year increments (e.g., “when did Bob Dylan win the Nobel Prize in literature?” between 1940 and 1959, 1960 and 1979, 1980 and 1999, or 2000 and 2019). The other trivia task asked participants to estimate college acceptance rates within ranges of 15% (e.g., “What was the acceptance rate at the Massachusetts Institute of Technology (MIT) in 2016?” 0-15%, 15-30%, 30-45%, or 45-60%). To vary math problem difficulty, algebraic equations required different amounts of steps and types of skills (e.g., combining like terms, fractions) to solve, such as  $15 - x = 19$  and  $6(-10 + 3x) + 2(5x + 6/5) = -10x$  and fraction arithmetic problems involved different operations (addition, multiplication, subtraction, division) and some had common denominators while others did not, such as  $\frac{3}{2} - \frac{1}{2}$  and  $\frac{1}{2} \times \frac{7}{4}$ . On both math tasks, as in the trivia tasks, there were four multiple-choice options per problem and the three distractor solutions were designed to be the results of different errors a participant might make (complete surveys can be found on the Open Science Framework: <https://osf.io/ak2dr/>).

We collected 200 responses per task and excluded participants who failed to correctly respond to instructional manipulations after two tries at the start of the task or who claimed to have searched the internet or used other assistive technology to answer the questions. Table 1.3 indicates how many participants were included in analyses for each task.



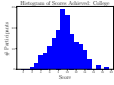
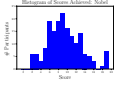
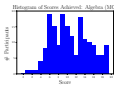
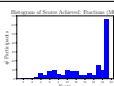
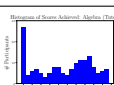
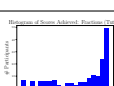
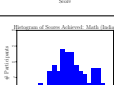
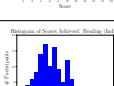
	DOMAIN	$n$	MEAN SCORE	SCORE DIST.	SA	TIME (MIN)	$r$ (SCORE, SA)
Studies 1–4	COLLEGE	178	9.42		9.53	10	.08
	NOBEL	173	9.34		7.51	11	.47
	ALG (MC)	179	10.85		8.25	26	.72
	FRAC (MC)	172	14.55		12.93	19	.82
Studies 5–6	ALG (TUTOR)	160	9.08		8.70	41	.70
	FRAC (TUTOR)	170	13.56		11.96	22	.78
Study 7	MATH (INDIA)	114	10.96		15.11	–	.35
	READING (INDIA)	113	6.51		13.92	–	.20

Table 1.3: Summary table (number of participants, mean score, mean post score estimate, mean time, and Pearson correlation between scores and post score estimates).

## Results

As expected, individuals' confidence judgments were represented by different combinations of parameter values for all tasks (see Figure 1.12 for some interesting examples) and though there were some parameters that were more common amongst participants, values varied across individuals (see Figure 1.14 for histograms of best-fit parameters). We determined which model best fit each individual's data by calculating the SSE for each model. For the majority of participants, the dynamic model fit the data better (i.e., the best-fit model's  $\sigma_{dyn}$  was nonzero).

We can clearly see that these parameters return interpretable and meaningful results on an individual rather than group level which contributes information beyond what previous metacognitive modeling efforts have provided. For example, the participant represented in Figure 1.12c had a higher than average  $\sigma_{dyn}$  parameter which is clear based on their increasing confidence judgements as they achieve multiple correct responses consecutively.

Participants with a negative  $\mu_\theta$  parameter gave low confidence ratings in general (such as the participant represented in Figure 1.12d) and a small  $\epsilon$  alluded to more “correct” judgments – that is lower confidence following an incorrect response and higher confidence after a correct response (for example Figure 1.12b).

In many of the datasets, there was a spike of best-fit models with  $\epsilon = 0.5$ , meaning these participants were at chance for guessing whether they were correct or not on each problem. This makes it seem that many participants are not tracking their performance on individual multiple-choice items, perhaps because this sensitivity is being overridden by other influences.

There are many cultural influences on self-views in math especially that may impact participants’ prior beliefs about their ability (captured by  $\mu_\theta$ ). The individuals who completed the algebra and fractions tasks have very low  $\mu_\theta$  parameters compared to those who were asked trivia questions, perhaps owing to negative attitudes about math in particular (which I discuss in more detail in the next Chapter). Adult participants will have had different amounts of math education, which shows in the variability of their sensitivity to their correctness (or  $\epsilon$  parameter). Compared to both trivia tasks, participants were more reactive to their recent evidence as shown by the variety of  $\sigma_{dyn}$  parameters as well as their higher average. We also observe an unanticipated contrast between the trivia tasks: on the college acceptance rates task, much more than the Nobel prizes assessment, there was a lot of error in guessing accuracy (shown in the large number of individuals with  $\epsilon = 0.5$ ) and very little reactivity to recent evidence (because the static model with  $\sigma_{dyn} = 0$  prevailed for most participants).

## Studies 5–6: Open-ended math

Here, we look at data from studies involving the same problems as on the math tasks (Studies 3 & 4), but with an open-ended problem solving interface rather than multiple-choice.

### Model specifications

In this pair of studies, there were no multiple-choice options, so we forgo the  $g$  parameter used for the model in Study 1. All problems had to be solved to continue to the next question.

### Procedure

Participants were directed to one of two tasks where they solved 20 math problems in an interactive math tutor, Emmy’s Workshop<sup>8</sup> (Rafferty et al., 2020), either solving algebraic equations (Study 5) or fraction arithmetic (Study 6). As in the previous studies, participants were asked prior to each question “You are about to solve a problem. How confident are you that you will solve it correctly?” as well as following the final problem (on a scale from 0

---

<sup>8</sup><https://emmysworkshop.com/>

to 100), which resulted in a total of 21 judgments per person. All participants received the same problems in the same order and the problems were identical to those from Studies 3 & 4.

## Results

As with the multiple-choice math tasks, the dynamic model fit participants better overall and as is visible in Figure 1.14, there was even more variability in best-fit dynamic parameters ( $\sigma_{dyn}$ ), with algebra in particular (both multiple-choice and in the tutor) accounting for a larger proportion of participants with very large  $\sigma_{dyn}$  values. There was also more variability in priors over ability compared to the multiple-choice counterparts ( $\mu_{\theta}$ ) and for fractions in the tutor, slightly higher error parameters ( $\epsilon$ ) than for multiple-choice.

## Study 7: Math and reading in India

After observing interpretable parameter values across different studies online with adults, we hoped to test out a similar paradigm, but within-subjects (where the same participants do multiple tasks in different domains) and rather than compare math to trivia, use a more educational comparison domain: reading. This study was conducted in India with 9th grade students in a school.

### Model specifications

This pair of studies again did not include multiple-choice solutions, however, there was one significant difference from the previous studies: these tasks were conducted on paper due to constraints of the testing site, so all participants could skip problems if they pleased. We therefore had to adjust the model to account for questions left blank. In the model itself, this meant that we set the likelihood for the current step equal to 1 for all particles on an iteration when there was missing data, meaning the weights were unchanged from the previous time step when participants gave a blank response.

### Procedure

Participants were recruited from a school in Gujarat, a province of India where English is treated as the primary language. Of the 128 consented ninth grade students at this school, a total of 111 completed all parts of the study.<sup>9</sup> One student was eliminated for refusing to respond to a part of one of the surveys. There was an attention check procedure, described below, which excluded an additional five students from analyses. The analyses reported below include the remaining 105 participants (54 boys and 51 girls).

---

<sup>9</sup>Five students completed only the first part, but were absent for the second part of the study so were excluded from all analyses.

The study took place in two parts: participants first completed either the math or reading task, and returned one week later to complete the remaining task. Testing occurred in groups of up to 20 at once, and all participants in one group were given the same task. Participants completed a 20-question assessment in math or reading. Assessments of math and English reading/writing ability were taken from the Woodcock-Johnson III Tests of Achievement (Woodcock et al., 2001). For both assessments, all questions were placed in a randomized order so that place in test did not predict difficulty, and all students received the same order. Students were asked questions ranging from fairly easy (e.g.,  $8 \times 5$ ) to more difficult (e.g., 12% of 6.0). In the reading task, students were asked to locate the grammar, spelling, punctuation, or capitalization mistake in a sentence (e.g., ‘She likes to drink melk.’) and identify the correction. There was exactly one mistake per item and scoring was done as per the Woodcock-Johnson III test manual. Questions were slightly adapted: all names were changed to high frequency Indian names (keeping the gender the same as in the original question and varying whether the name was stereotypically Muslim or Hindu) and one word (‘color’) was changed to the group’s known spelling (‘colour’).<sup>10</sup> To track students’ assessments of their own abilities throughout the task, as in Studies 1–6, they reported confidence judgments about their performance prior to each problem before seeing it (“how sure are you that you will solve the next problem correctly?”). Finally, a post-survey was distributed in which they were asked to estimate their performance on the finished assessment, and answer questions about their mindset, perceived competence, values, enjoyment, and stereotypes in the domain being tested.

## Results

There was a great deal more overconfidence from these participants compared to those in the previous studies. This is particularly visible in the high  $\mu_\theta$  parameters compared to all other domains (see Figures 1.14s and 1.14v). There was additionally substantial variability in the dynamic parameters (shown in Figures 1.14u and 1.14x), but also much more error (Figures 1.14t and 1.14w). This final point is made more clear by the  $n$ ’s listed in Table 1.4 — approximately half of participants in Study 7 were best fit by a model with an  $\epsilon$  value of 0.

## Discussion

By deploying a version of our rational model of self-assessment that tracks judgments made over time and eliciting multiple judgments from participants rather than a single post-task score guess, we have made clear that each individual has a different recipe for judging their own ability. On all tasks, there is indeed variability across participants in their best-fit parameter values and the majority make dynamic judgments, with greater sensitivity towards more recent evidence (instantiated by  $\sigma_{dyn} > 0$ ).

---

<sup>10</sup>All participants filled out a pre-survey before solving the problems where they predicted their performance and the difficulty of the assessment they were to complete.





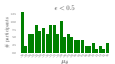
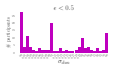
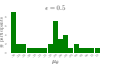
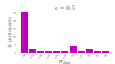
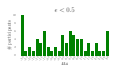

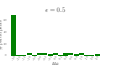







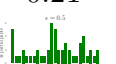













STUDY	$\epsilon < 0.5$				$\epsilon = 0.5$		
	n	$\mu_\theta$	$\epsilon$	$\sigma_{dyn}$	n	$\mu_\theta$	$\sigma_{dyn}$
1. COLLEGE ACCEPTANCE	84	-0.67	0.25	0.08	94	-0.29	0.13
							
2. NOBEL PRIZES	133	-1.14	0.17	0.83	40	-1.26	0.23
							
3. ALGEBRA (MC)	76	-0.10	0.17	1.51	103	-1.94	1.06
							
4. FRACTIONS (MC)	117	0.34	0.16	0.90	55	-1.85	0.26
							
5. ALGEBRA (TUTOR)	107	-0.37	0.20	1.50	53	0.21	1.66
							
6. FRACTIONS (TUTOR)	124	-0.12	0.24	0.85	46	-0.18	1.03
							
7A. MATH (INDIA)	60	1.11	0.28	1.07	54	0.70	1.24
							
7B. READING (INDIA)	56	1.22	0.30	0.73	57	0.72	1.09
							

Table 1.4: Average model parameters separating those with  $\epsilon = 0.5$  and those with  $\epsilon < 0.5$ .

A few patterns across all studies emerge that seem to align with the nature of the tasks and their respective participants. For one, prior beliefs about ability ( $\mu_\theta$ ) are highest for both tasks run in India: there are many differences that set this group apart from the others, but notably they are children. Previous work has indicated that children tend to be more overconfident than adults (Bjorklund & Green, 1992), so I surmise that this result reflects that idea. Additionally, the algebra task from studies 3 and 5 was more difficult than the fractions task from studies 4 and 6 (the algebra problems required knowledge of fraction arithmetic as well as other mathematical skills). This is reflected in the lower  $\mu_\theta$  values on the algebra task compared to its identically formatted fractions counterpart.

Overall, results from this series of studies of human confidence judgments suggest that we can glean more information from individual-level analyses than from aggregated group-level

analyses and that there is incredible variability across individuals.

## 1.6 General discussion

In this chapter, I proposed a highly customizable rational model that detects a number of psychological traits that are necessary for self-assessment. The parameters that represent these traits are interpretable (Section 1.3), can distinguish between theories (Section 1.4), and are able to be discovered at the individual level (Section 1.5).

To unpack differences between math and other domains in this chapter, there are a few observations that jump out. One is that the mathematical tasks presented in the open-ended tutoring environment (Studies 5 and 6) led to more variability in all parameters compared to the same tasks in a multiple-choice format (Studies 3 and 4) as is clear in Figure 1.14. There are notably fewer at-chance estimates of performance (where  $\epsilon = 0.5$ ) for participants who completed the tutor versions (see Table 1.4). There are also higher  $\sigma_{dyn}$  parameters for the algebra than the fractions tasks. Another difference is apparent in the dynamic parameters in Study 7: for the students in India,  $\sigma_{dyn}$  is higher on average on the math than on the reading task, suggesting that they are able to incorporate their own recent performance when making confidence judgments about math problems, but less so for reading.

The large data sets that we collected through our experiments, particularly in Section 1.4, provide an impressively clear picture of the form of the relation between self-assessment and performance. Over the last few years, psychology has struggled with concerns about methodology and the replicability of prominent findings. We believe that the way forward is not to merely try to replicate our past results, but to combine the large sample sizes made possible by modern technology with advances in computational modeling to pursue new psychological research that gives us deeper insight into the phenomena behind those past results. As in the present case, the primary outcome of this research confirms the existence of a key phenomenon, and providing definitive evidence for such phenomena – particularly those that enjoy the public profile of ideas such as the “Dunning-Kruger” effect – is an important step towards re-establishing the validity of psychological research. We view our results as an example of what this approach can achieve, providing a high-resolution picture of the nature of human miscalibration. The next chapter also makes use of a large dataset.

This chapter provides a framework for better measuring people’s perceptions of their ability and it allows for comparing across domains, task types, ages, and individuals. Self-assessment is a simple and low-cost way of measuring a facet of metacognition, and I have made some strides beyond the scope of this dissertation toward incorporating these tools in applied settings, namely educational assessments and with clinical populations. I hope that the openness of my data and modeling apparatus can lead to more information gathering about the nature of perceived ability. In the next two chapters, perceived ability remains relevant as it is brought up in people’s narrations of their experiences with math (Chapter 2) and, combined with attitudes about math, relates to judgments about what types of activities involve math (Chapter 3).

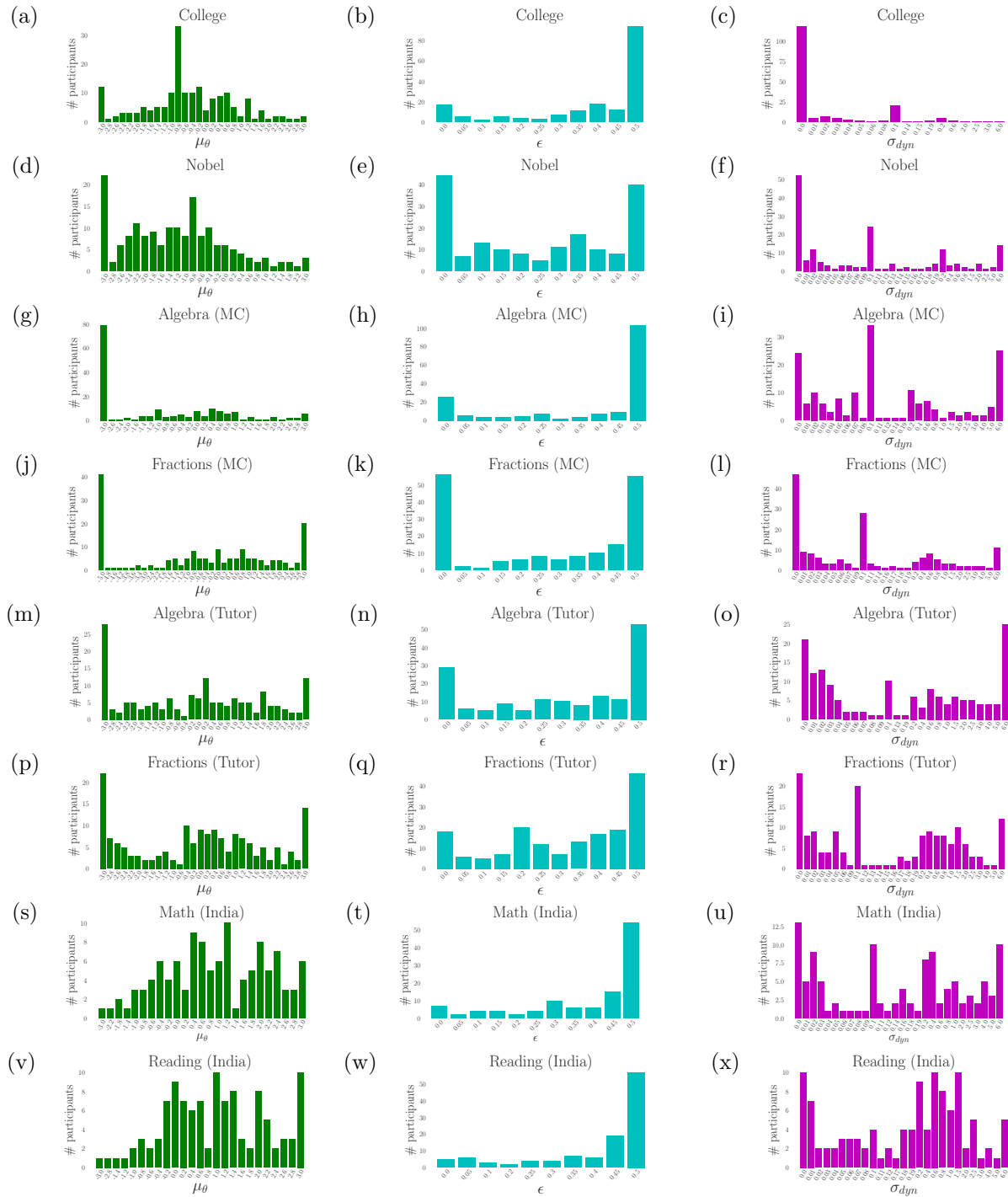


Figure 1.14: Histograms depicting number of participants in each study who were best fit by each possible  $\mu_\theta$ ,  $\epsilon$  and  $\sigma_{dyn}$  value for each study (College, Nobel, Algebra Multiple-choice, Fractions multiple-choice, Tutor Algebra, Tutor Fractions, India Math, and India Reading). Note that when  $\sigma_{dyn} = 0$  this indicates that the static model fit the data better for that individual.

*The truth is a matter of the imagination.*

Ursula Le Guin, *The Left Hand of Darkness*

## 2

# Sense of Belonging

*How do you feel about math?*

## 2.1 Introduction

MATH, PERHAPS MORE THAN ANY OTHER SUBJECT taught in school, is associated with more intense – and more negative – emotions, for many students across ages (OECD, 2016). The ways that math are typically discussed perpetuate pervasive gender stereotypes (Chestnut & Markman, 2018), and many students are gripped by math anxiety that causes them to perform more poorly on math assessments than they should given their actual ability (E. Carey et al., 2016). Teachers in the United States receive inadequate subject matter training and are expected to teach more topics in less depth compared to those in other countries (Schmidt et al., 2002). Researchers have long sought to relieve students of their negative feelings about math in the hopes of improving performance on mathematical assessments. At the societal level, when students perform better and have more positive feelings about math, they might then be more keen to pursue higher degrees and careers in science, technology, engineering, and math (STEM) domains (Lauermann et al., 2017). In this paper, we present data from a survey consisting of open-ended responses that provides a lens through which we can observe the complexity of people’s emotions related to math. We demonstrate a methodology for studying emotions surrounding individuals’ experiences with mathematics, and how those emotions develop across the lifespan, which uses computational text analyses for thoroughly exploring written narratives about math attitudes.

People’s current feelings about math, or their *math attitude*, have been primarily studied using closed-form surveys where respondents must choose from a set of pre-selected options, including ones to measure math anxiety (Dowker et al., 2016; Hembree, 1990b), positive attitude about math (Chen et al., 2018), and math self-concept (Seaton et al., 2014). However, most research about math attitudes begins from the presumption of a negative (e.g., anxiety (Foley et al., 2017)) or positive (e.g., self-efficacy (Dowker et al., 2016)) stance. Beginning with a particular presumption about attitudes inhibits respondents from expressing a full range of possible feelings about math, and prevents researchers from recognizing



that individuals may, for instance, be feeling excitement or uncertainty rather than anxiety. In this paper, we argue for a more holistic exploration of math attitudes and the use of computational methods as a complement to statistical analyses of surveys to examine 1) overall attitudes towards mathematics 2) when these attitudes were formed, and, crucially, 3) people's own narratives about what contributed to their attitudes. Understanding how personal narratives vary based on individuals' feelings about math, personal characteristics, and when and how their math attitudes were formed will provide a more complete picture of how autobiographical experiences relate to math attitudes, which could in turn inform future educational interventions.

Exploring people's narratives about their math attitudes builds on existing studies collecting open-ended responses to questions related to math attitudes in which researchers primarily perform qualitative analyses and the data exhibit significant emotional richness. Towers et al. (2017) reviewed the literature about autobiographical accounts of experiences with math and noted that most analyses were qualitative in nature and a large portion of the participants were practicing or pre-service teachers, a group consisting of more women than men and individuals with more negative than positive attitudes. In studies of students, Nasir and Hand (2008), for example, observed that even for just two students in a very similar educational environment (at the same school and even on the basketball team together), experiences with and feelings about math were both rich and divergent. Studies using qualitative methodologies find results that are often consonant with findings from closed-form surveys. Larkin and Jorgensen (2016) had students record themselves talking about their math experiences in a private room and observed gender differences and negative attitudes forming from a young age, consistent with previous work based on closed-form surveys (Cvencek et al., 2011; Else-Quest et al., 2010). Students participating in this study also alluded to boredom, frustration, and other feelings that cannot be captured by the use of a single-scale measure. The emotional richness of these data suggests that a more unrestricted method for assessing math attitudes is necessary for capturing the wide range of feelings people might harbor. This method may also provide additional insight for future study design – studies like Geary et al. (2020) that seek to understand what traits are most predictive of later math ability could be informed by detailed explorations of rich open-ended narratives which are likely to generate traits of students worth exploring.

More systematic analysis approaches help make more definitive conclusions about the consistency of results across open-ended and closed-form responses and enable researchers to work with much bigger data. One example of this was a recent study of math narratives by John et al. (2020) who collected life story narratives from emerging adults and qualitatively identified themes in their narratives; they related the manually coded themes to closed response questions to uncover any systematicity to the observed themes. Some previous work has employed computational text analyses to examine open-ended responses related to feelings about math. Park et al. (2014) used word frequency-based categorization to analyze expressive writing of high versus low math anxiety participants before taking a math assessment, finding associations between word usage and math performance. Kle-

banov et al. (2017) asked participants to write reflections about the usefulness of science for themselves or for society. They specifically highlighted how believing course material to be important in the future (it's *utility*) was related to participants' retention of course material. The authors used computational methods to see whether these reflections could be automatically graded by a trained model, based on how they spoke of science's utility value. Various computational text analysis methods have additionally been used to predict math performance (Crossley & Kostyuk, 2017) and ratings of personal identification with math (Crossley et al., 2018) from transcripts of students' written interactions with an online intelligent tutoring system.

Here, we address the problem of systematically evaluating the scope of emotions about math without manually coding responses by combining the two approaches described above: collecting open-ended narratives about experiences with math and employing computational text analysis methods. With a survey administered to a sample of 1,000 participants online, we show that our methodology can (1) obtain results that are consistent with previously observed patterns of attitudes and (2) capture more nuance in math attitudes. Given that research using autobiographical narratives about math experiences has primarily been studied in teachers (Towers et al., 2017), we aim to collect from a broad sample of US participants to understand the role of teachers from students' perspectives. To further probe the trajectory of a person's *remembered* math attitudes, we ask participants to write separately about multiple time periods in their education. To systematically analyze results from this rich qualitative measure, we demonstrate a case study use of computational text methods, specifically classifiers to identify how word usage predicts differences between groups, and topic modeling as an exploratory means of uncovering frequently arising themes in participants' narratives. These results show that the language used to write about math experiences is related to participants' Likert-style responses about their math attitudes and that the valence of the language used and the most relevant themes do change over time.

## 2.2 Related literature

### Math attitudes

Previous work has established relationships between demographic factors and math attitudes that we expect to replicated in a systematic analysis of open-ended narratives. Gender differences in self-efficacy, motivation, and levels of math anxiety have been reported (Else-Quest et al., 2010), with research consistently finding that girls feel more negatively toward math than boys (Gunderson et al., 2012). Prior research has found that women report forming math attitudes earlier than men (Cvencek et al., 2011), and adopt stereotyped beliefs from a young age (Bian et al., 2017).

We predict that open-ended narratives will highlight the influence of parents and teachers, who may transmit their math anxiety to students (Eccles & Jacobs, 1986; Gunderson et al., 2012; Maloney et al., 2015). Parents (Elliott et al., 2017; Levine et al., 2010) and

teachers (Boonen et al., 2011) are known to influence children’s math *ability* from a very young age via their “number talk,” which could in turn influence the formation of children’s variable math attitudes.

The focus of research about math attitudes is often on gender differences in ability and perceived ability, but socioeconomic status (SES) is reliably related to mathematics achievement (Kalaycioglu, 2015) and lower quality education and thus may also influence attitudes about and experiences with math. Levine et al. (2010) found that SES correlated with the amount of number talk parents used with their young children, so though we may not find explicit SES-based differences in participants’ attitude ratings, we might expect different themes to appear in their narrative responses. For example, if parents spoke less to their children about math, participants may cite parents less frequently as influences on their attitudes toward math. If STEM jobs are paid higher, then those with lower childhood SES might also be less likely to have a parent with a STEM job. Given that the personal relevance of STEM subjects contributes to students’ persistence in these domains (Durik & Harackiewicz, 2007; Harackiewicz et al., 2016), it may be that those with positive attitudes and/or who ended up with a career in a STEM field will be more likely to cite instances of using math outside of school settings.

## Computational text analyses

As noted above, previous work regarding open-ended surveys about math attitudes has primarily relied on qualitative methods to make sense of their data. Here, we collect large samples that would make relying on manual coding or annotation cumbersome. Computational text analysis, on the other hand, is both much faster to implement with large sets of data and is invulnerable to coding errors or inconsistencies arising from human misinterpretation, although it brings its own challenges (e.g., important themes might be overlooked and interpretability could pose a problem). Additionally, Roberts et al. (2014) noted that manual and computational methods achieve comparable results, but only hand-coding becomes less reliable as datasets grow in size. To examine how people’s narratives vary based on their personal characteristics and overall attitudes about math, we employ two types of computational text analysis: (1) *classification* of groups based on narratives to find predictive features, and (2) *topic modeling* to identify themes in word usage and differences across groups.

Classification analyses are used to assign responses to pre-specified categories based on differences in content across members of these categories. These methods have been used to classify emotions in text (Mohammad, 2011; Neuman et al., 2012), and have performed well in past sentiment analyses (Liu et al., 2013; Narayanan et al., 2013; Ting et al., 2011). Mohammad and Yang (2011) found gender differences in email correspondence, where women used more terms on the joy–sadness axis, while men preferentially used terms relating to fear and trust. Our initial text analyses employ Naïve Bayes classifiers, which are known for being simple and performing well (Ting et al., 2011). These classifiers use a subset of the data to learn a model relating the input to the categories and then are tested on the

remaining data to see if they can make accurate predictions about the category to which each response belongs (e.g., if the text was written by a man or a woman). We analyze the learned Naïve Bayes classifiers to determine which words were most valuable for determining whether the response belonged to one group or another, as well as how much more frequently the informative words came up for the relevant group in the training phase.

While classification is known as a supervised method because we provide labels to the classifiers (e.g., ‘man’ vs. ‘woman’) and then make predictions about those labels, topic modeling (Blei et al., 2003), an unsupervised method, is more exploratory because it allows the model to discover patterns in unlabeled data. Topic modeling is a method that draws from the idea that sets of related documents (in this case, the responses from each individual) are connected by *topics*. By applying this method, we seek to uncover the abstract topics that occur across a set of documents — in this context, the narrative themes that occur across individuals. We specifically employ a Latent Dirichlet Allocation (Blei et al., 2003) topic model that identifies topics based on word frequency across a set of documents. Supervised methods have been used in related work about math attitudes, such as Park et al. (2014), discussed above, who identified differences between high and low math anxiety students and Crossley et al. (2018) who made judgments based on their participants’ ratings of their identification with math. With topic modeling, we are able to uncover patterns in the narrative responses without giving the model any information beyond what is written (e.g., the gender or age of the respondent). Each narrative is probabilistically modeled as a collection of topics, with each topic generating different words at different rates. Topic modeling has been used to identify semantic patterns for a range of applications, including recommending academic papers (C. Wang & Blei, 2011) and supporting scholars of literature (E. Alexander et al., 2014; Buurma, 2015). The combination of unsupervised and supervised computational text analysis methods will allow us to characterize what themes occur in narratives and how they differ across time, attitudes, and demographic characteristics.

## 2.3 Study 1: Math narratives in a large sample

We obtained data from a survey posted by an online science publication and used these data as an initial test of the text analysis methods described above, as well as to generate hypotheses for a second study. To validate the proposed method of posing open-ended questions about math attitudes and using computational text analyses to systematically explore responses, we hoped to show that, in this convenience sample, people’s narratives about their math attitudes are aligned with previous research.

### Materials

Data was provided by Quanta Magazine,<sup>1</sup> an online science publication that aims to “illuminate basic science and math research through public service journalism.” On October 20,

---

<sup>1</sup><https://www.quantamagazine.org>

2016, as part of a series of articles exploring STEM education, Quanta released a survey<sup>2</sup> asking how respondents felt about math, when they formed their opinions, and to explain in an open-ended text box what shaped their opinions (see Figure 2.1a). Anyone who found the site could respond to the survey, though all questions posed were optional (e.g., age, name). The full data set considered in this paper consists of a total of 3030 responses (73% men, largely a result of the magazine’s male-skewed readership<sup>3</sup>) and was received for analysis on March 24, 2017.

## Data analysis

In both this study and the following, we perform two different types of computational text analysis that we detail here. Prior to any text analyses, we use standard text preprocessing techniques on the open-ended responses by removing stop-words and punctuation, then stemming and lemmatizing the remaining words to reduce all words to their roots (Vijayarani et al., 2015). To examine the narratives, we utilize two distinct computational text analysis methods: *classification* of groups to find predictive features, and *topic modeling* to identify themes in word usage.

### Classification

We used Naïve Bayes classifiers to examine word usage in men’s versus women’s narratives and in those who have positive versus negative attitudes about math. Specifically, we employ the `NaiveBayesClassifier` function from Python’s Natural Language Toolkit (nltk version 3.2.2) package.<sup>4</sup> The nltk package contains many easy-to-use functions for text preprocessing and analyses.

To test the accuracy of each classifier, we shuffled the data and separated it into a training set consisting of 80% of the data and a test set comprising the remaining 20%, a typical practice in machine learning. We train the classifier on the training set, then report the classifier’s accuracy predicting responses on the test set as well as details of ‘precision’ and ‘recall.’ Along with these statistics, we report a subset of informative features (meaning words that are more common for one specific subgroup) as well as example uses in context, but the Appendix contains a longer list of features along with their usage ratios for classifiers from both studies.

### Topic modeling

We next employed topic modeling as a more exploratory analysis of the open-ended responses using Python’s gensim package (version 1.0.1).<sup>5</sup> Since the number of topics must not exceed

---

<sup>2</sup><https://www.quantamagazine.org/20161020-science-math-education-survey/>

<sup>3</sup>In October 2016, 82% of Quanta’s viewers were men.

<sup>4</sup><https://www.nltk.org/>

<sup>5</sup><https://pypi.org/project/gensim/>

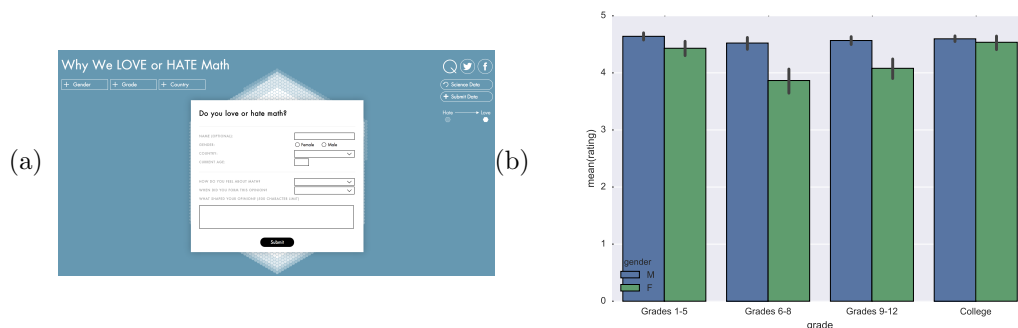


Figure 2.1: (a) Screenshot of the Quanta magazine survey and (b) mean attitude ratings (from 1, meaning “hate,” to 5, meaning “love”) by gender and time when attitudes were formed. Female respondents (green) gave significantly lower ratings for grades 6-8 and grades 9-12 than other groups. Error bars reflect one standard error.

the number of words in a given document, we removed any response containing less than 10 words, leaving 2105 responses in this study. All topic models were designed to output 5 topics in this paper, though this number may be tuned to deliver the largest possible set of meaningful topics.

## Results

Feelings toward math were mostly positive, unsurprising given the self-selectedness of this group. Men had more positive attitudes about math than women ( $t(3028) = 8.77, p < .001, d = .36$ ), consistent with previous research. There was a significant effect of time when opinions were formed on rating ( $F(3, 3026) = 11.74, p < .001$ , see Table 2.1), though these differences over time varied for women and not men (see Figure 2.1b). The largest portion of women (39%) reported forming their attitudes about math early on (in grades 1-5) while the largest portion of men (33%) reported doing so in college (see Table 2.2).

	NUMBER OF PEOPLE	MEAN RATING	STD. DEV
GRADES 1-5	904	4.56	.93
GRADES 6-8	465	4.31	1.13
GRADES 9-12	773	4.44	0.98
COLLEGE	888	4.58	0.71

Table 2.1: Number of people in Study 1 who said they formed their opinions in each time-period as well as mean ratings of those individuals and standard deviations.

	FEMALE				MALE			
	#	%	mean	std. dev	#	%	mean	std. dev
GRADES 1-5	318	39%	4.43	1.09	586	27%	4.64	0.82
GRADES 6-8	149	18%	3.87	1.34	316	14%	4.52	0.94
GRADES 9-12	201	24%	4.08	1.25	572	26%	4.57	0.84
COLLEGE	155	19%	4.54	0.77	733	33%	4.60	0.69

Table 2.2: Number and percentage of female and male respondents who decided their feelings in each time-period as well as mean ratings in each bracket about math and standard deviations.

### Text analyses

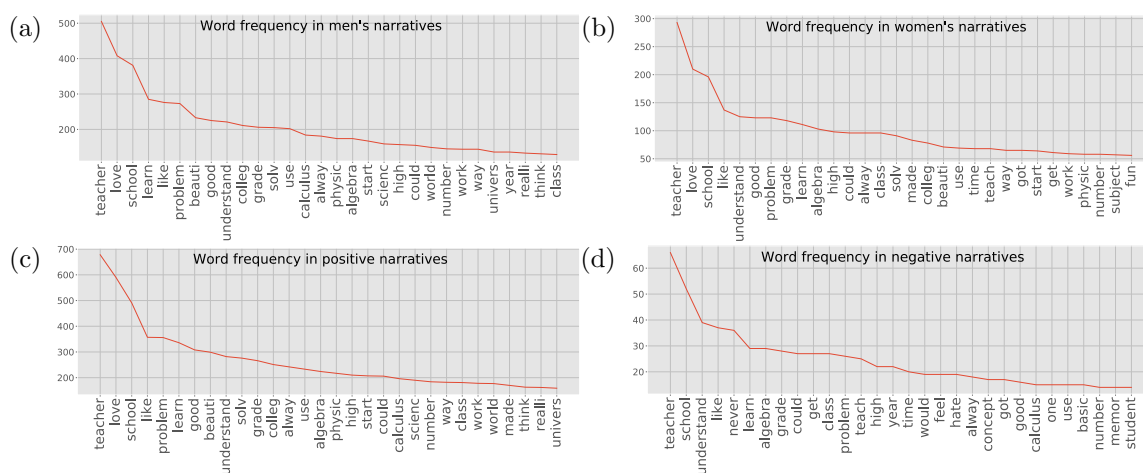


Figure 2.2: Distribution of 30 most frequent word roots in Study 1 for (a) men, (b) women, (c) those with positive feelings about math and (d) those with negative feelings about math. All plots exclude words with root “math.” One observation is that the top words for all groups include “teacher” and “school” but “beauti” and “love,” as expected, appear in all groups except the negative attitude group which instead includes “hate,” “feel,” and “never.”

While some differences in word frequencies across groups can be seen in Figure 2.2, it is difficult to determine from these plots alone whether any particular words are reliably more common in one group versus another, so we perform text classification to distinguish word usage across groups. Word usage in the open-ended responses was consistent with the closed-ended responses about attitudes: women and people who rated math more negatively used more pejorative language. Words like “inability” and “difficulty” were predictive of a negative attitude, indicating that people might come to dislike math because of poor performance. Women were more likely to use words such as “girl,” “stupid,” “trick,” and

“confuse.” The valence of words was expected, but the types of words revealed greater depth to respondents’ attitudes as well as potential causes of these attitudes. For example, the use of terms like “girl” and words related to internalization of feelings (“stupid”) by women shows that gender may influence how they perceive their own ability.

Of the five topics constructed via our topic model, one focused on schooling (math, school, teacher, high, college, algebra, calculus, grade, did, good like, class, really, got, learned), another on the elegance of math (e.g., mathematics, world, physics, language, beauty, love, real, understand, universe), and another on puzzle solving and math subfields (e.g., problems, solve, algebra, teacher, grade, geometry, calculus, love, class, puzzles, problem). This set of topics demonstrates that there are distinctions in how individuals think about their most salient experiences with math and that topic modeling can help illuminate these differences. In the following study, we collect a great deal more narrative data from individuals which allows us to use topic modeling to distinguish between groups more than is allowed for in the present study.

## 2.4 Study 2: Math attitudes across time

Though the results from the online survey analyzed in Study 1 were biased by the male- and positive attitude-skewed readership of Quanta magazine (it could also be that those with positive attitudes were more inclined to complete the survey), they provide insight into how math attitudes differ by gender and show that attitudes can be formed at any point in schooling. These results also demonstrate the robustness of this methodology: even asking only one short open-ended question without length restrictions to a skewed group, interpretable and consistent patterns are observed via computational text analyses. For example, women gave math more negative ratings on the scale and also used more negative language in their narratives. Study 2 expands upon the survey posted by Quanta and gives a fuller picture of math attitudes through a large and more balanced sample of participants.

To gain a full picture of people’s math attitudes and further explore the use of computational analyses beyond previous work, we conduct a study designed to elicit more detailed narratives through the posing of multiple probing open-ended questions. We explore gender in our analyses, which is frequently the focus of investigations of math attitudes, but also address how other personal characteristics might interact with experiences learning math. We based our survey design, sample size, and hypotheses on analyses of a short survey posted by the science publication Quanta magazine<sup>6</sup> and preregistered hypotheses and associated analyses on AsPredicted.org<sup>7</sup> before collecting any data.

---

<sup>6</sup><https://www.quantamagazine.org/20161020-science-math-education-survey/>

<sup>7</sup><https://aspredicted.org/blind.php?x=28z639>



## Participants

We recruited 1,000 participants from Amazon’s Mechanical Turk (MTurk),<sup>8</sup> a platform used frequently in psychological research that produces high-quality responses and where wide ranges of ages and educational backgrounds are represented (Mason & Suri, 2012). All participants were required to be based in the US and were compensated \$4 upon completing the survey. We found in piloting that participants wrote lengthy and on-topic responses to our open-ended questions, suggesting that they were focused on the task at hand (see Appendix for details about the power analyses used to select the sample size).

## Procedure

Participants were directed to a survey on Qualtrics where they initially provided their consent to participate and were asked a pair of attention check questions before they could proceed to the main survey (see Appendix for details). Once they consented and passed the attention checks, they were asked to answer a series of multiple-choice questions probing their feelings about math, when they believe their attitudes were formed, and what factors they think contributed to their attitudes (e.g., teachers, parents, job). They were asked one open-ended question on this first page: “Please describe what shaped your general feelings about math, including specifically how the factors you mentioned above shaped those feelings.”

They were then directed to a series of separate pages asking them about their feelings in each phase of schooling (grades 1–5, grades 6–8, grades 9–12, and college if they attended). Along with these rating questions, for each phase of school, they were asked two open-ended questions: “Please describe a memorable event related to math that happened in [grades 1–5/grades 6–8/grades 9–12/college]” and “How did this event contribute to your personal feelings about math?” (see Figure 2.3). To encourage participants to write more detailed responses to the open-ended questions, they were not able to advance through the survey unless they wrote at least 10 characters. If they did not meet this threshold, they were prompted to elaborate their response.

After all sections of the survey eliciting open-ended narratives and attitude ratings, we asked participants to rate their math anxiety using the single-item math anxiety rating scale: “On a scale from 1 to 10, how math anxious are you?” (Núñez-Peña et al., 2014). We included this item given that math anxiety is the most widely used measure of feelings about math. Including a measure of math anxiety allows us to compare general feelings about math and how narratives vary based on how we are assessing math attitudes. Finally, participants were directed to an optional questionnaire about their demographics and exposure to math at school and at work.

---

<sup>8</sup><https://www.mturk.com/>

## Data analysis

### Exclusions

Participants were given two chances to respond to two multiple-choice questions about a short set of instructions upon beginning the survey. If they failed on the second try, they were asked to return the study on MTurk and were not compensated. A total of 147 failed on both attempts so did not complete the survey and we collected 1,001 responses to the full survey. As described in the preregistration, if more than half of a given participant's open-ended responses were off-topic,<sup>9</sup> they were excluded from analyses. Two independent coders agreed on 59 exclusions based on written responses that were unrelated to the questions being answered. The raters had an initial Cohen's  $\kappa$  of 0.83 and the responses about which there was disagreement were dealt with through discussion. Responses from a total of 942 participants were thus included in analyses. Since all demographics questions were optional, for some analyses we do not have relevant data for every participant.

### Calculating socioeconomic status

We are interested in participants' childhood SES since open-ended narratives refer to their educational experiences. SES was measured in two ways: (1) as mother's level of education and (2) calculated via Hollingshead's Four Factor Index of Social Status (Hollingshead, 1976). The details of calculating this index are located in the Appendix.

### Hypotheses

This dataset of attitude ratings, narratives, and additional demographic data were used to evaluate the following hypotheses (both pre-registered), focused on the relation between gender, math attitudes, and time period when attitudes were solidified:

1. Hypothesis 1: Men will have more positive attitudes about math than women and this will come across in the types of words and topics present in their narratives (women will use more negatively valenced words).
2. Hypothesis 2: Women who report forming their attitudes in middle and high school will have more negative attitudes than the other groups.

Traditional statistical analyses were used to address Hypotheses 1 and 2; computational text analysis methods were used to systematically explore the collected narratives of men and women, as indicated in Hypothesis 1. We preregistered specific analyses to test the proposed hypotheses.

---

<sup>9</sup>There were a few optional open-ended responses in the demographics questionnaire, so we only performed exclusions based on the non-optional questions about math attitudes. There were thus a total of 9 of these responses used for exclusions (or 7 for those who did not attend college).

### Statistical analyses

We performed a series of ANOVAs with math attitude as the dependent variable and then gender (hypothesis 1), time of attitude formation, race, and childhood SES as independent variables, as well as a gender by time interaction (all indicated in the preregistration). We included race and SES in this model because the majority of research about math attitudes focuses on gender differences. Because SES was measured in two ways, this ANOVA was performed twice for each math attitude rating, once with just mother’s level of education as a stand-in, and the other based on Hollingshead’s Four Factor Index of Social Status (Hollingshead, 1976). This resulted in fourteen ANOVAs because people rated their feelings seven separate times (about math in general, doing math in school, doing math at work, in grades 1–5, grades 6–8, grades 9–12, and college). Because of this large number of tests, we adjusted the  $\alpha$  value of 0.05 via a Bonferroni correction by dividing  $\alpha$  by fourteen (giving  $\alpha = .0036$ ).

For analyses using mother’s education as our measure of SES, we excluded those who did not indicate their gender, race, or mother’s education as these were all optional questions. To reduce the degrees of freedom, we additionally excluded those who identified with a gender other than man or woman, who listed more than one race, and who said ‘other’ for their mother’s education. This left 889 participants for this analysis (and 693 for the model predicting attitudes in college since those who did not attend college were excluded).

We then did the same analyses but with Hollingshead score as our measure of childhood SES. Excluding those in the set of 846 with Hollingshead scores who did not specify either ‘Woman’ or ‘Man’ for their gender ( $n=3$ ), who did not indicate their race ( $n=1$ ), or selected more than one racial category ( $n=43$ ), we ran this set of ANOVAs with the remaining 799 participants. For the model predicting college rating, we only had responses of those who attended college, so this particular model made use of 629 responses.

### Text analyses

We perform two different types of computational text analysis that we detail here. Prior to any text analyses, we use standard text preprocessing techniques on the open-ended responses by removing stop-words (e.g., “the,” “and,” “a”) and punctuation, then stemming and lemmatizing the remaining words to reduce all words to their roots (Vijayarani et al., 2015). To examine the narratives, we utilize two distinct computational text analysis methods: classification of groups to find predictive features, and topic modeling to identify themes in word usage (see Figure 2.3 for details).

Classification: We used Naïve Bayes classifiers to contrast word usage in men’s versus women’s narratives, those who have positive versus negative attitudes about math, those who have high versus low math anxiety, and those with high or low childhood SES. Specifically, we employ the NaiveBayesClassifier function from Python’s Natural Language Toolkit (nltk version 3.2.2) package. Naïve Bayes classifiers learn a model for the probability that each word will be generated given a category (e.g., ‘man’ or ‘woman’). For instance, the model might learn that when the respondent is a man, there is a 2% chance of generating

‘teacher’ for a particular word. The model makes the simplifying assumption that words are generated independently given the category, using a bag-of-words representation that ignores the ordering of words.

To test the accuracy of each classifier, we shuffle the data and separate it into a training set consisting of 80% of the data and a test set comprising the remaining 20%, a typical practice in machine learning. We train the classifier on the training set, then report the classifier’s accuracy predicting responses on the test set as well the precision (the proportion of selected items that are correct) and recall (the proportion of correct items that are selected). Along with these statistics, we report a subset of informative features (words that are more predictive of one specific subgroup) as well as example uses in context. In the Appendix, we include the top 25 most informative features for each classifier along with their “informativeness” ratio which is calculated as the highest value of the probability of a feature given the category divided by the lowest value of the probability of this same feature given the other category. So, if the word “wrong” is used more by women than men, the informativeness is calculated as  $\frac{P(\text{“wrong”}|\text{woman})}{P(\text{“wrong”}|\text{man})}$ .

We ran multiple classifiers on either the first open-ended response or concatenated versions of all of a participant’s open-ended responses. The concatenated versions provide more text for each participant, which improved the classifier performance slightly compared to taking into account only the first open-ended responses. However, since these questions target educational experiences, using only the first open-ended response gives a broader picture of what is contributing to math attitudes beyond experiences in school. We built classifiers to identify informative features that differed by gender, attitude rating, math anxiety rating, and childhood SES. We intended to additionally look at systematic differences that varied based on race, but because our sample did not have sufficient variability in the racial distribution of study participants, we were unable to perform any computational text analyses predicting race.

Topic modeling: We next employed topic modeling as a more exploratory analysis of the open-ended responses using Python’s gensim package (version 1.0.1). Topic models were designed to output five topics in this paper and included individual words as well as frequently co-occurring two-word phrases. We did not tune the number of topics.

## Results

We first report on quantitative analyses of the multiple choice questions about attitudes and personal characteristics, focusing on the overall numerical results and tests of preregistered hypotheses. We then turn to results of computational text analyses based exclusively on responses about overall math attitudes. Finally, we explore individuals’ attitudes over time by conducting text analyses on responses recounting experiences from the educational stages and examine changes in attitude ratings across each individual.

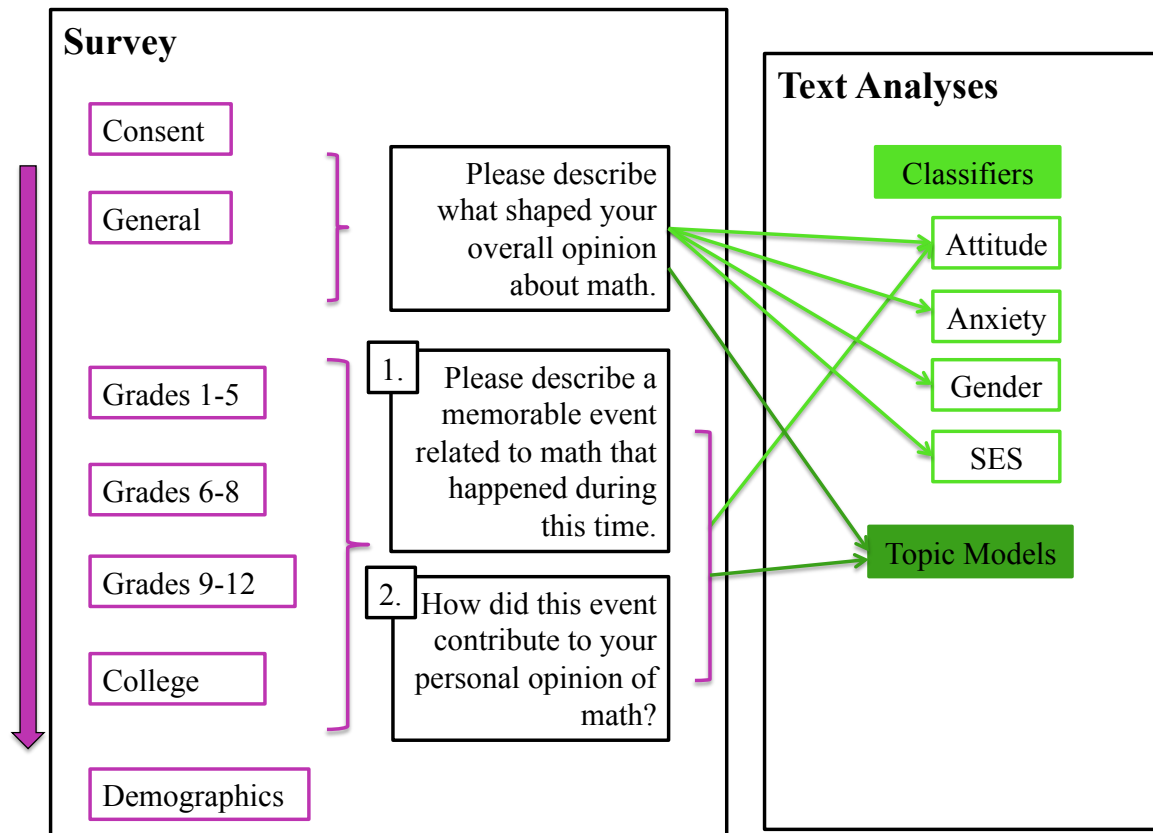


Figure 2.3: Diagram depicting open-ended survey questions for each section of the survey along with associated computational text analyses. Using the first general written response, we ran four classifiers distinguishing word usage between groups (positive and negative attitudes, high and low math anxiety, men and women, high and low SES), and one topic model. To explore narratives about individual time periods, we ran four classifiers separating those with positive and negative attitudes over each pair of responses for each time period. We then ran one topic model over these responses.

## Results: Numeric results and tests of preregistered hypotheses

Of the 942 participants included in analyses, 508 were men (54%), 429 were women (46%), and the rest responded “other” or did not specify. The mean age was 36 years (range: 18–79) and the average time spent on the survey was 21.6 minutes. There were 717 white participants, 79 black or African American, 52 Asian or Asian American, 40 Hispanic or Latino, and the rest other or mixed race.

Participants responded to a series of questions on a scale from 0 to 10 to rate their feelings

about doing math “in general” ( $M = 5.54$ ,  $sd = 2.88$ ), “in school” ( $M = 4.97$ ,  $sd = 3.20$ ), and “at work” ( $M = 5.52$ ,  $sd = 2.86$ ). As seen in Figure 2.4, there was a spike at 0 or “hate” when asked about doing math in school and a spike at the middle rating of 5 when asked about doing math at work. Participants were asked how often they used math at work and more than half ( $n=494$ ) stated that they used it hourly or daily. The more frequently they used math at work, the more positive their feelings were about doing math at work; those who used math less than once a year had an average work attitude of 3.98, while those who reportedly used math hourly gave math a rating of 6.81.

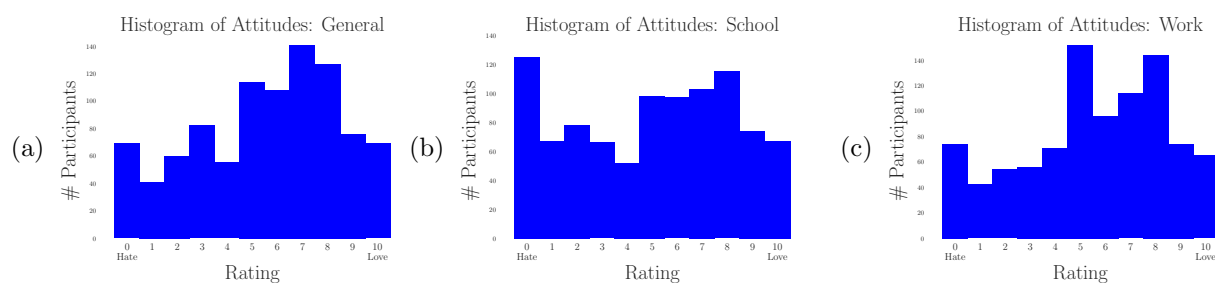


Figure 2.4: Histograms of ratings of (a) general feelings about math ( $M = 5.54$ ), (b) feelings about doing math in school ( $M = 4.97$ ), and (c) feelings about doing math at work ( $M = 5.52$ ).

	NUMBER OF PEOPLE	MEAN RATING	STD. DEV
GRADES 1-5	218 (23%)	5.94	3.37
GRADES 6-8	247 (26%)	5.29	2.85
GRADES 9-12	319 (34%)	5.03	2.66
COLLEGE	104 (11%)	6.66	2.13
AFTER FORMAL SCHOOLING	29 (3%)	7.00	2.04

Table 2.3: Number of people who said they formed their feelings about math in each time-period as well as mean ratings and standard deviations for those groups. Notably, relatively few made their decisions in college or after (83% of participants formed their feelings at some point in primary or secondary school, especially in high school). The 25 unaccounted for said they did not recall when, also a markedly small proportion.

Each of us narrates our life as it suits us.

---

Elena Ferrante, *Those Who Leave and Those Who Stay*

## Memory

At the end of the survey, participants were asked whether they had trouble remembering any particular time-periods. A total of 429 participants (45.5%) selected at least one time period (185 mentioned grades 1-5, 90 grades 6-8, 35 college, 34 grades 9-12, and the rest reported multiple). Only 8 participants selected all four time-periods of schooling, indicating that people generally believe they remember at least part of their educational experiences with math. They were given an opportunity to explain their responses and many mentioned remembering overall attitudes, but not specific details as much. For example, one respondent shared that “Since this was a long time ago I felt that I was digging deep to remember events. I remember an overall feeling about math at this time, but memory changes as we get older.” Some who claimed to remember well pointed to math being a particularly salient subject, e.g., “I was able to remember everything pretty well. If it was a different subject I would have had a harder time.” This seemed particularly common when memories were negative: “I had an easier time remembering bad memories than good;” “I had such terrible feelings all through school about math that it is very easy to recall;” “Given the traumatic nature of my educational arc, I have quite strong memories of all of these.” If they did not struggle with math early on, they seemed less likely to remember: “I don’t think I could remember anything significant from these grades because math wasn’t hard for me then.” These responses highlight the importance of math attitudes: early negative experiences remain salient and likely influence choices throughout life.

## Influences

We asked a multiple-choice question about what influenced participants’ attitudes before they were directed to write their first open-ended response. They could check all that applied to them. Teachers and parents were the most common cited influence, with over 85% (805) of participants identifying teachers and 25% (243) identifying parents as important influences on their opinions. Other influences were less common: 129 friend(s), 123 current job, 116 book(s), 106 game(s), 95 future job, and 12 listed clubs. The relative importance of parents and teachers shifted based on when participants formed their overall attitude about math.

The influence of teachers was highest for those who claimed to form their attitudes in high school (92% of this group listed teachers as influential compared to 87% in both elementary and middle school and 72% in college). In contrast, the influence of parents was highest for those who formed their attitudes in elementary school (39% of whom listed parents as opposed to 30% from middle school, 18% from high school, and 19% from college).

### Socioeconomic status

To calculate childhood socioeconomic status using the Hollingshead measure, we had to exclude a total of 95 participants (53 for leaving at least one necessary response blank, and 42 that could not be hand-coded because responses to the questions about parents' professions were too vague). This left 846 participants with a Hollingshead score for childhood SES. The average SES score was 39.28 (sd = 10.90), which is comparable to other studies with diverse samples (e.g., (Arentoft et al., 2015) with a mean of 36.84). SES calculated via the Hollingshead measure was correlated with mother's level of education ( $r(846) = 0.75$ ,  $p < .001$ ), but not perfectly so.

### Testing preregistered hypotheses

We use pre-registered analyses as described above to test both hypotheses.

Hypothesis 1: Men will have more positive attitudes about math than women and this will come across in the types of words and topics present in their narratives (women will use more negatively valenced words).

For nearly all models, main effects of gender and time when attitudes were formed (grade) were significant at  $p < 0.0036$ , adjusted via a Bonferroni correction (see Appendix for details about ANOVAs). The only model for which this was not the case was that predicting attitude in elementary school, though there was a significant interaction effect of gender and grade. For attitudes in middle school, there was a significant effect of gender, but not of grade. Race and SES were not significant factors at this alpha value in any of the models. We have evidence in favor of our first hypothesis: for nearly all attitude ratings, men rate math significantly more positively than women (see Table 2.4).

	WOMEN	MEN
GENERAL	5.07 (3.02)	5.95 (2.67)
SCHOOL	4.58 (3.29)	5.30 (3.07)
WORK	5.04 (2.96)	5.94 (2.70)
GRADES 1–5	6.04 (3.06)	6.53 (2.81)
GRADES 6–8	4.98 (3.09)	5.80 (2.92)
GRADES 9–12	4.40 (3.34)	5.06 (3.19)
COLLEGE	4.96 (3.32)	5.69 (3.06)

Table 2.4: Mean attitude ratings and standard deviations for women and men.

We discuss the types of words used by different groups in the next section about text analyses after addressing Hypothesis 2 (which, again, was motivated by results from Study 1).

Hypothesis 2: Women who report forming their attitudes in middle and high school will have more negative attitudes than the other groups.



As is apparent in Figure 2.5, though time when attitudes were formed was a significant predictor of rating, there were no significant differences between male and female respondents who formed their attitudes in middle and high school (looking only at the general math attitude rating). Additionally, women who reported forming their attitudes in elementary school had comparably negative attitudes about math to those who formed them in middle or high school, but not men. We also performed the same analysis with only the participants who attended college ( $n=745$ ). In this population, there are gender differences in attitude ratings for those who formed them in middle school, though still not high school – men continue to have feelings about math that are similar to women’s when formed in high school (see Appendix for similar histograms for all of the six other attitude ratings made).

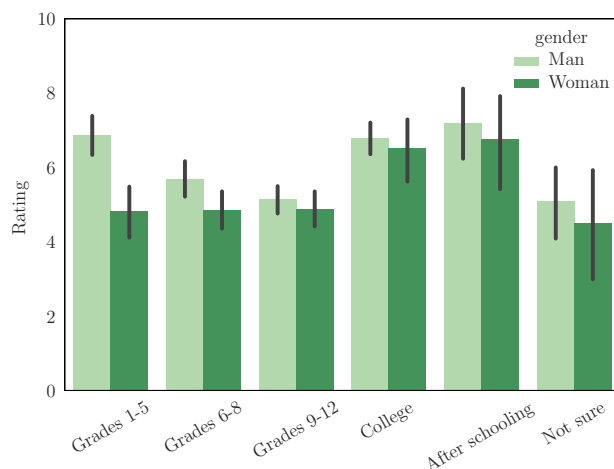


Figure 2.5: Differences in overall attitude ratings by gender and time when attitudes were reportedly formed. Error bars indicate standard deviations.

## Summary

The consistency between this representative sample of adults online and results from previous work (gender differences in attitudes) demonstrates that this population is well-suited for study. In addition to this proof of concept, inquiring about time when attitudes were formed and specific influences provides new observations including that, as expected, influential adults play a large role and attitudes can form at any educational stage.

## Results: Text analyses based on responses about overall math attitudes

Computational text analyses are methods for systematically detecting patterns in open-ended narratives. Simple word frequencies (see Figure 2.6) are a coarse lens into people’s narratives, but the Naïve Bayes classifiers provide a way to identify differences in word usage

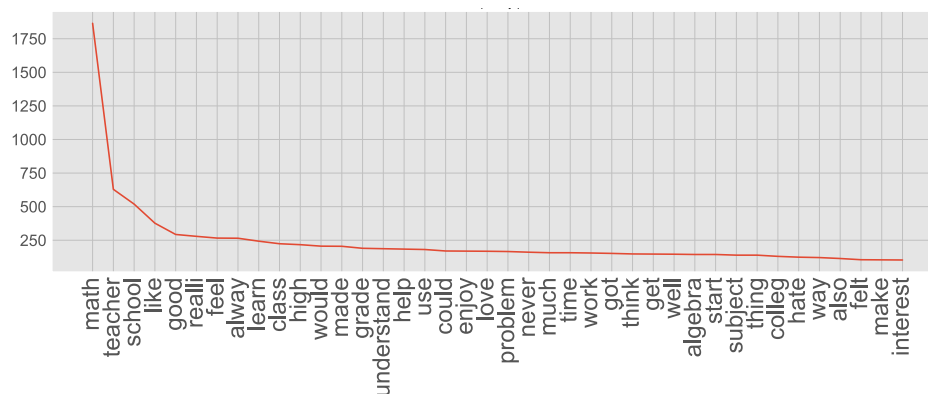


Figure 2.6: Frequency of 40 most frequent words from the general open-ended responses across all participants. Note that these words are generated after pre-processing of the text, including eliminating stopwords, stemming, and lemmatizing.

between various groups. We report the accuracy of each classifier, but what we find most valuable about this method is not model performance, but the informative features and their interpretability for each classifier. The main use here of the numerical indicators of classifier performance is to compare the classifiers we built to each other rather than in the hopes of constructing a classifier with the best performance possible in order to use these models predictively. The reason for this is that the open-ended question asked is specifically about feelings, so we can verify the validity of computational text analyses in this context if model performance is superior when comparing math attitude ratings versus other variables.

### Classifier: Positive vs. Negative Attitude

To classify positive versus negative attitudes, we take all responses below 5 on the question “how do you feel about math” (scale from 0–10) to indicate a negative attitude and those above 5 to indicate a positive attitude. This eliminated 114 responses that gave a 5. It was a slightly imbalanced sample containing 307 negative responses and 521 positive (63% positive). This classifier achieved 77% accuracy<sup>10</sup> on the test set (precision: 0.78; recall: 0.77;  $F1 = 0.76$ ) and people with negative attitudes were more likely to use words (informative features) such as “behind,” “confused,” “fraction,” and “avoid,” while those with positive attitudes were more likely to use the words “game” and “play” (see Table 2.5 for example usages by participants and Table A.8 in the Appendix for the top 25 most informative features and note that these terms all had particularly high informativeness ratios compared to the other informative features generated by this classifier and the others). Positive responses were classified more accurately than negative ones (see the confusion matrix located in the Appendix Table A.9), with the most common classifier error being actually negative responses classified as positive.

<sup>10</sup>Above chance is about 63% because that is the proportion of positive responses.

	WORD	EXAMPLE
NEGATIVE	behind	“I was way <u>behind</u> and frustrated”
	confused	“The way my professor and teachers of the past taught the subject is just unintuitive and just end up making me more <u>confused</u> ”
	fraction	“When I started learning <u>fractions</u> , I couldn’t understand what the teacher wanted”
	avoid	“I tried to <u>avoid</u> anything to do with math”
POSITIVE	game	“what mainly shaped the general love for math were <u>games</u> ”
	play	“ <u>played</u> a lot of <u>games</u> that also used a bit of math”

Table 2.5: Example uses of informative features from attitude rating classifier.

### Classifier: High vs. Low Math Anxiety

The math anxiety measure was, as expected, negatively correlated with general feelings (Pearson’s  $r(940) = -.56$ ,  $p = 3.20 \times 10^{-78}$ ), feelings about doing math in school ( $r(940) = -.55$ ,  $p = 2.89 \times 10^{-76}$ ), and feelings about doing math at work ( $r(940) = -.51$ ,  $p = 2.95 \times 10^{-63}$ ). Here, we identify how narratives are differently worded for those with high or low math anxiety compared to those with positive or negative feelings about math.

For this classifier, we excluded those who gave their anxiety a rating of 5 ( $n=75$ ) or 6 ( $n=85$ ) because the scale went from 1 to 10 rather than 0 to 10 as in the attitude ratings. This left 782 responses to use for this classifier, 431 low anxiety (55%, with a rating from 1 to 4) and 351 high anxiety (with a rating from 7 to 10). This classifier performed at 69% accuracy on the test set (precision = recall = 0.69;  $F1 = 0.69$ ) and those with high anxiety were more likely to use lemmas like “difficulty,” “confuse,” “lost,” and “frustrat” while those with low anxiety used words like “video,” “everyday,” “engin,” “application,” and “challenge” (see Table 2.6 and Table A.10 in the Appendix). The lemmas “confus” and “frustrat” often co-occurred for those with high math anxiety (“I just became more confused and frustrated”). Interestingly, “statistics” was a relevant feature for low math anxiety participants. Statistics anxiety has been known to be uncorrelated with math anxiety (Baloglu, 2004).

### Classifier: Men vs. Women

We generated a classifier to see what differences there were in the narrative responses of men and women. We expected women to use more negatively valenced language given that their general feelings about math were less positive ( $M = 5.06$ ) than those of the men ( $M = 5.95$ ). A total of 937 participants specified that they were a man ( $n = 508$ ) or a woman ( $n = 429$ ) so this analysis refers only to this population. The training set included 750 responses while the test set was composed of the remaining 187. This classifier performed only slightly above

	WORD	EXAMPLE
HIGH MA	difficulty	“remember having <u>difficulty</u> with math”
	confuse	“It doesn’t interest me, and it <u>confuses</u> me”
	lost	“I had an algebra teacher that wasn’t very good at teaching and then I felt that I got <u>lost</u> ”
	frustrate	“seeing complex problems usually <u>frustrates</u> me”
LOW MA	video	“ <u>video</u> games helped by increasing my interest in computing”
	everyday	“My math teacher in grade school made math seem fun. She related variables to <u>everyday</u> objects”
	engin	“My dad is an <u>engineer</u> and my mom is pretty handy with numbers”
	application	“learning real world <u>applications</u> ”
	challenge	“My teachers took extra time with me to try and <u>challenge</u> me”

Table 2.6: Example uses of informative features from math anxiety rating classifier.

chance at 60% on the test set (precision: 0.59; recall: 0.60;  $F1 = 0.59$ ). Women were more likely to use words like “wrong,” “sure,” “instead,” “okay,” “language,” and “father” while men used words like “video,” “reinforce,” “application,” and “fascinating” (see Table 2.7). Based on the confusion matrix in the Appendix (Table A.13), men’s narratives were classified accurately more frequently than women’s. See Table A.12 in the Appendix for more details about the classifier’s output.

	WORD	EXAMPLE
WOMEN	wrong	“you’re either right or <u>wrong</u> ”
	sure	“I’m not <u>sure</u> ”
	instead	“there was only one correct answer <u>instead</u> of with Literature”
	okay	“I’m <u>okay</u> at basic math”
	language	“always remain a different <u>language</u> for me”
	father	“my <u>father</u> would help me with my homework”
MEN	reinforce	“I got positive <u>reinforcement</u> from my teachers”
	fascinating	“the golden ratio for example is simply <u>fascinating</u> ”

Table 2.7: Example uses of informative features from gender classifier.

These results reinforce the fact that women rated their feelings about math as more

negative compared to men. They also show a pattern of more internalizing language used by women, meaning they refer more to features of themselves (e.g., “wrong” and “smart”) while men seem to refer to applications of math by using language external to the self. However, the lower accuracy of this classifier suggests that this pattern is not as strong, which is evident in the lower informativeness ratios.

We additionally explored what gender differences could exist when comparing men and women who have similar feelings. To do so, we constructed two additional classifiers to distinguish men and women, looking exclusively at those with positive feelings (so who gave a rating above 5,  $n=518$ ) or negative feelings (ratings less than 5,  $n=305$ ). The positive group consisted of 313 men (60%) and 205 women and the associated classifier performed at 70% accuracy on the test set. Interestingly, this group of women still used more negative language than men (e.g., “fail,” “anxiety”). The negative group consisted of 132 men (43%) and 173 women and this classifier performed with only 52% accuracy. While the lowered accuracy could be due to the smaller sample, an alternative possibility is that men and women with negative attitudes may have more comparable experiences and use more similar language when describing those experiences.

### **Classifier: low vs. high SES**

One of the benefits of a computational text approach is that it allows us how word usage shifts in a variety of personal characteristics. We thus used it to also examine low vs. high SES. We ran one more classifier where did a median split of Hollingshead socioeconomic status scores. Removing those who had exactly the median of 38.5, this left 827 responses for analyses (414 below this threshold and 413 above). This classifier performed at only 54% but yielded interesting features. In particular, those of higher SES appeared to use more valenced terminology about their experiences (e.g., “wors,” “easier,” “stress”) compared to those in the low childhood SES group. It is no surprise that this classifier is not performing well because the distribution of Hollingshead scores is approximately normal with a high concentration of scores toward the middle, so there is not a large distinction between these two groups.<sup>11</sup> Given that SES was not significantly related to attitude, as shown earlier, it is also no surprise that there were no visible differences in valence of the terms used (meaning those of higher SES did not necessarily use more positive language than those with lower SES).

### **Topic model**

We ran an LDA-based topic model to discover patterns in the contexts in which words were used by participants. We used only each participant’s first open-ended response to feed into this topic model and eliminated all responses containing fewer than 10 words, leaving 907 responses for this analysis. We constructed a topic model over 5 topics (see Figure 2.7a

---

<sup>11</sup>We did not exclude more participants near the median because this would reduce our sample size more than we should for a text analysis.

for word clouds depicting each topic and Table A.14 for exact weight values associated with each word). These topics give a clear picture of certain themes in math learning experiences. As shown in Figure 2.7b, the school and teacher topic (topic #4) was the most prevalent throughout responses, consistent with the fact that “teachers” was selected most frequently as an influential factor in what shaped participants’ feelings. Other topics highlighted applications of math (e.g., “real\_world” and “skill” in topic #2) and negative feelings (e.g., “anxiety” and “embarrassed” in topic #1).

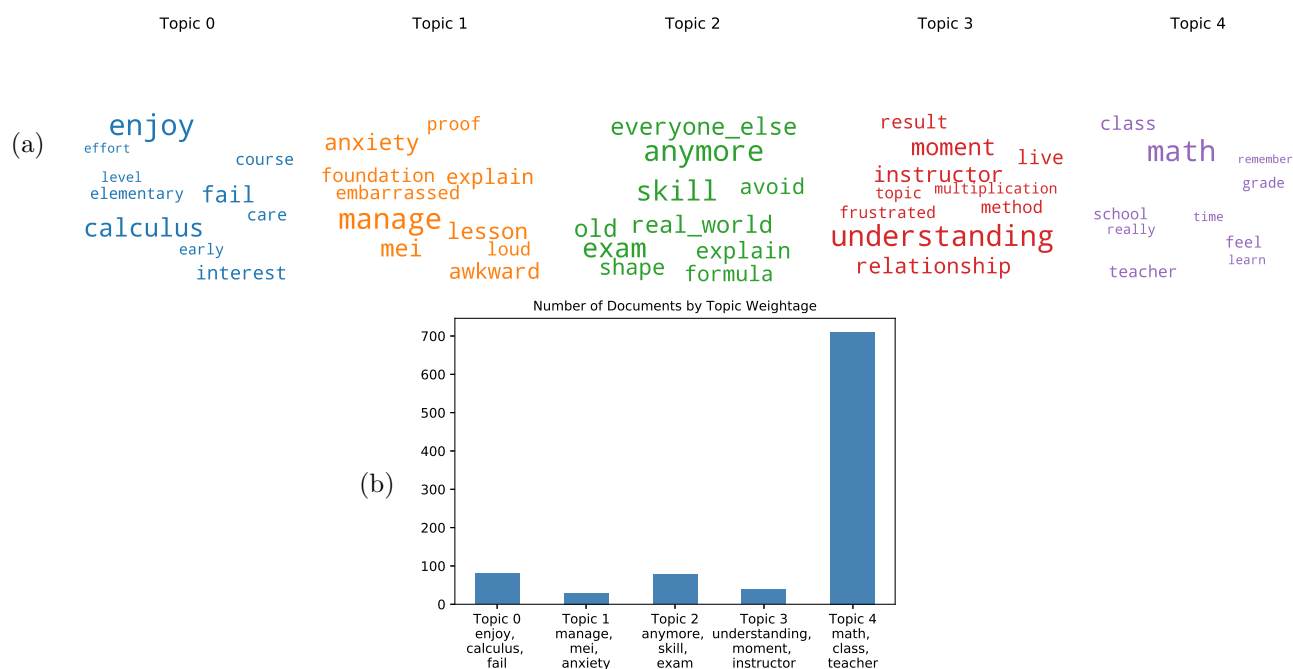


Figure 2.7: (a) Word clouds of top keywords for each topic. The size of the words is proportional to the weight and (b) topic weights over the entire set. Specifically, the sum of the actual weight contribution of each topic to respective documents.

## Summary

The text analyses exhibit great nuance in participants’ narrations about their math attitudes. The classifiers all performed above chance, but the best-performing classifier was the one comparing those with positive vs. negative attitudes. Given that this first open-ended question specifically asked people to write about what shaped their feelings, it is then unsurprising that the classifier predicting positive versus negative ratings performed better than the others we built. This shows that these text analysis methods may be able to reveal similar patterns to what would come out of quantitative analyses of a closed-form question which strengthens our case for the use of open-ended questions, but improves upon the value of closed-form responses because this method reveals more nuances in the types of emotions

experienced. To learn about other dimensions besides feelings about math, we need only rephrase the prompt to allude to experiences linked to gender or math anxiety, for example. Our results indicate that those with negative attitudes, high math anxiety, and women used more negative language and our topic model revealed that school is the most frequently discussed topic.

## Results: Text analyses based on specific time points

The previous text analyses demonstrate distinct patterns in narratives about overall math attitudes based on demographics. To better understand how narratives varied by time point, we performed text analyses of the narratives at each time point (see Table 2.8 for mean responses to the multiple-choice questions about math attitudes and Figure 2.8 for the distributions of ratings made about each time period.)

	MEAN RATING	STD. DEV
GRADES 1-5	6.30	2.94
GRADES 6-8	5.42	3.02
GRADES 9-12	4.76	3.27
COLLEGE	5.35	3.20

Table 2.8: Mean ratings standard deviations made in each time-period subsection of the survey (e.g., “in grades 1-5, how did you feel about math” (on a scale from 0 to 10).

## Classifiers

We first trained four classifiers to see whether word usage in written responses about a specific time period could predict attitude rating for the same time. As in previous classifiers, we excluded those who gave a rating of a 5 and categorized those with a rating under 5 as having a ‘negative’ attitude and those above 5 as ‘positive.’ Since there were two open-ended questions about each time period, we concatenated both responses for each participant to create one longer text for each. In elementary school, 136 (18%) gave ratings of 5, leaving 806 responses for this classifier, 594 (74%) of which were positive. This classifier performed with 80% accuracy on the test set. The most informative features were primarily used by those with negative attitudes, including, notably “hate,” “resent,” “father,” and “embarass” while those with positive feelings used the term “enjoy” frequently. In middle school, the number with neutral feelings was somewhat less (113 or 12%), leaving 829 for this classifier of which 491 had positive feelings (59%), also fewer than for elementary school. This classifier performed quite well on the test set (81% accuracy). Those with negative feelings used words or word roots like “hate,” “fail,” and “night” while those with positive feelings said “challeng,” “best,” “love,” and “posit.” There was some overlap between negative features at both of these time points (e.g., “hate,” “frustrat,” and “resent”) though the positive words

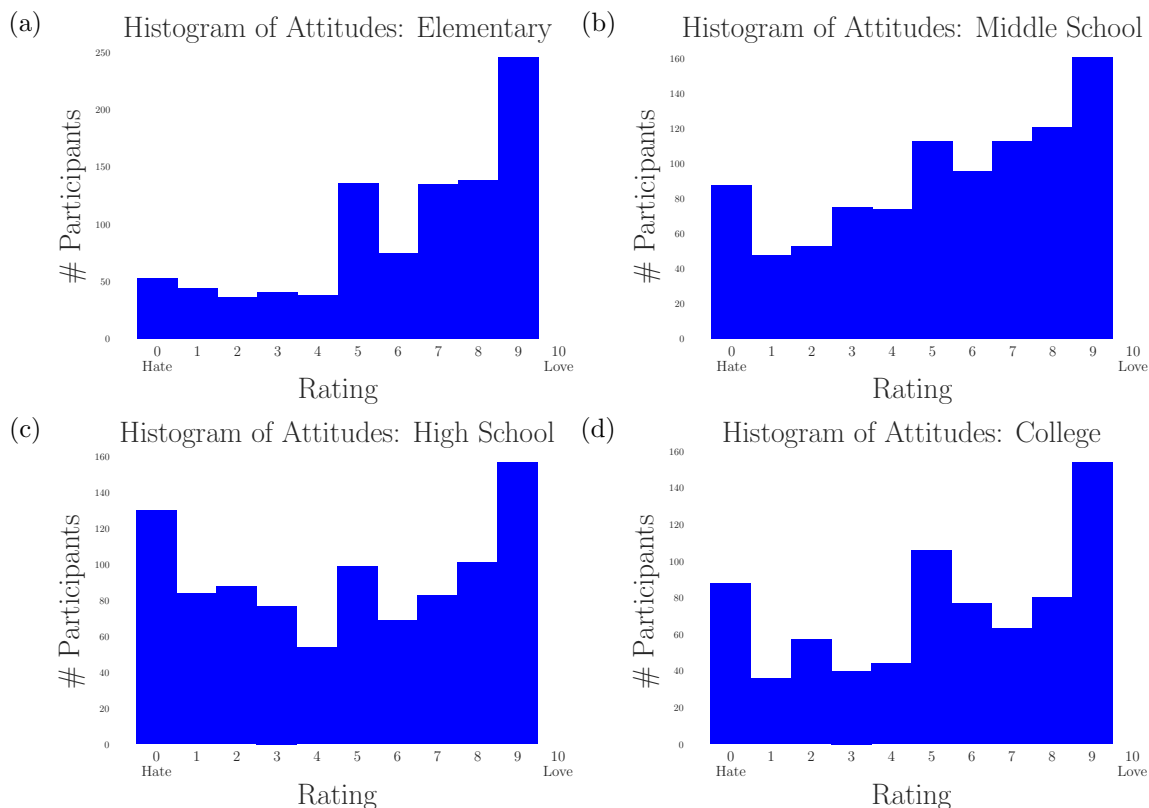


Figure 2.8: Histogram of attitude ratings from (a) Grades 1–5 ( $M = 6.30$ ), (b) Grades 6–8 ( $M = 5.42$ ), (c) Grades 9–12 ( $M = 4.76$ ) and (d) college ( $M = 5.35$ ).

seem to convey slightly different experiences (enjoyment vs. challenge and ability). In high school, even fewer gave neutral ratings (11%), so 843 responses were used in analyses, even fewer of which were positive (410 or 49%). This classifier performed with 72% accuracy on the test set. Those with negative attitudes referred to “hate” again, as well as “embarass” and “unabl.” There were again a larger amount of informative positive features than for earlier times including “proud,” “applic,” “compet,” “favorite,” “comput,” and “perfect.” Positive attitudes related even more to ability as well as seeing the applicability of the subject. Finally, we ran a classifier about the 750 participants who attended college, removing 108 who claimed neutral feelings (14%). A larger proportion of these 642 participants had positive feelings (376 or 59%) than did the high school group and this classifier performed at 71% accuracy on the test set. Again, those with negative attitudes used the word “hate” the most frequently as well as “annoy” and “horribl.” Those with positive attitudes said “love,” “import,” “science,” and “physic” showing that positive attitudes become even more associated with identifying relevant applications of math to other domains.



## Topic model

Next, we ran a topic model over all time-point specific responses where the two written responses referring to each educational stage were concatenated and linked no longer to the participant, but to the relevant period. This resulted in a corpus of 3576 responses (942 responses each from elementary school, middle school, and high school, plus 750 responses from college). Figure 2.9 again displays word clouds of primary keywords for each topic. Topic #2 (referring to “teacher,” “hard,” “math,” and “feel”) was the most dominant topic for responses about all stages, but this dominance became gradually less pronounced from grades 1–5 (dominant topic for 58% of responses) to college (dominant for 34% of responses). Topic #1 was dominant for about 17% of responses for all pre-college periods but for 24% of responses referring to college. Further, Topic #0 was more dominant for responses about grades 6–8 than for other periods and Topic #4 was much more dominant for grades 9–12 and college compared to the earliest years of schooling. Topic #3, which contains more negative language than the others was not nearly as prevalent as the others.

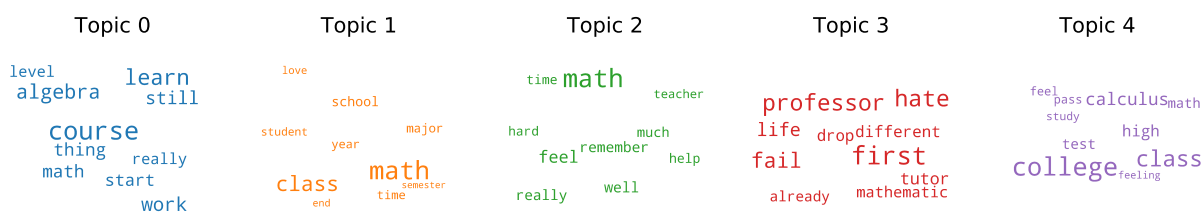


Figure 2.9: Word clouds of top keywords for each topic over the time-point-specific responses. The size of the words is proportional to their respective weights. These can be viewed alongside exact weight values in Table A.15 in the Appendix.

	TOPIC #0	TOPIC #1	TOPIC #2	TOPIC #3	TOPIC #4
GRADES 1-5	12%	17%	<b>58%</b>	6%	6%
GRADES 6-8	21%	17%	<b>53%</b>	3%	6%
GRADES 9-12	19%	18%	<b>47%</b>	4%	13%
COLLEGE	17%	24%	<b>34%</b>	5%	2%

Table 2.9: For responses from each time-point, proportion for which each topic was dominant (the topic a response is about).

## Individuals over time

We were additionally interested in how much variability there was in ratings of feelings over time on an individual level. Overall feelings about math, as measured by the very first

survey question, were highly correlated with ratings made about each time point, though the least with feelings during grades 1–5 ( $r = .44$ ) and the most highly correlated with attitudes in grades 9–12 ( $r = .74$ ). Because not all participants attended college (and not everyone even partakes in math courses in college), high school was the final time when many completed a math course. As seen in Table 2.3, a mere 29 (3%) participants reportedly formed their attitudes after formal schooling. To probe individual rather than aggregate shifts in attitude, we calculated the variance over each participant’s set of ratings at each time point and found that variance was relatively high on average (3.97, range: 0–25).<sup>12</sup> This measure included four ratings for those who attended college and three for those who did not. Variance differed based on general feelings whereby those who rated math more negatively in their very first general attitude ratings had more variance in their ratings over time ( $M = 4.85$ ) compared to those with positive feelings ( $M = 3.30$ ), as confirmed by an independent t-test ( $t(941) = 5.11, p = 3.92 \times 10^{-7}$ ). It seemed that this pattern was due to having high variance being related to having more positive feelings in grades 1–5 and more negative feelings otherwise (meaning many of those who ultimately developed more negative attitudes began with positive attitudes which then declined).

We looked at each transition to understand when participants are more likely to have math attitudes that increase or decrease. We first identified what proportion of participants had a drastic shift or not: going from a positive (rating greater than a 5) to a negative attitude (less than 5) or vice versa. In all transitions, the majority either stayed negative or positive, though from grades 1–5 to grades 6–8, 121 participants transitioned from a positive attitude to a negative while 29 went from negative to positive. In the next transition, there were again far more who went from a positive attitude in middle school to a negative one in high school ( $n = 125$ ) than negative to positive ( $n = 54$ ). However, between high school and college, this pattern reversed: more participants went from negative to positive ( $n = 74$ ) than from positive to negative ( $n = 50$ ). Even fewer participants had two major shifts in their attitude ratings across periods (where, for example, they started with a positive attitude in elementary school, reported a negative attitude in middle school, then back to a positive attitude). However, the fact that any participants experienced multiple shifts is worth holding onto: attitudes can always change for the better.

## Summary

Participants’ math narratives from different time periods clearly show the variability of experiences over time and our results demonstrate the richness of emotional experiences that can be gleaned from an in-depth probing of experiences from multiple time-points. In general, remembered math feelings in elementary school are positive and present-day attitudes about math are most influenced by someone’s most recent experiences learning math.

---

<sup>12</sup>Note that we used biased sample variance, meaning we divided the sum by  $n$  rather than  $n - 1$  since the set of ratings used in these calculations are complete rather than a subset of possible responses.

## 2.5 Discussion

No feeling is final.

---

Rainer Maria Rilke, *Go to the  
Limits of Your Longing*

In this project, we explored connections between math narratives and attitudes, exposing associations between a person’s memories of their experiences, their demographic characteristics, and the attitudes that they formed about mathematics. We demonstrated the validity of our methodology by comparing open-ended to multiple-choice responses: we found, for example, that people who give more negative ratings about math on a scale also use more negatively valenced language. Responses on a fixed scale have the potential to be biased by social desirability or a respondent’s current state or mood. These issues can be lessened by posing open-ended questions and analyzing written responses that express more of the nuance involved in math attitudes. Computational text analyses make it possible to analyze open-ended narratives efficiently. Manual coding of open-ended responses has diminishing returns: there are strong relationships between hand-coding and computational text analysis output, but hand-coding becomes increasingly less reliable as data get larger (Roberts et al., 2014). The computational text analysis methods we used thus allow us to discover patterns beyond what hand-coding might provide and enable us to scale up with even more expansive datasets.

Leveraging these computational text analysis methods shows that we can identify systematic patterns across narratives and thus capture a fuller and more accurate picture of attitudes and how experiences vary. Through the elaborated survey we administered, we were able to see how word usage differs for groups based on socioeconomic status and other personal characteristics beyond what is typically focused on in studies of math attitudes. Our classifiers performed above chance, but more important than accuracy, they discovered meaningful differences in word usage across groups and provided a characterization of the relationship between word usage and attitude valence.

These particular computational text analysis methods help us describe trends in the data, some of which are consistent with previous work. We also observe new patterns that can lead to further research targeting specific groups. Consistent with previous work, we found gender differences in the valence of participants’ narratives, the frequent acknowledgement of parents and teachers as influences, and the mentioning of specific domains of math. Worth noting are the differences between informative words for overall attitude and level of math anxiety: a positive attitude seems to be associated with enjoying math (with words like “play” and “game”); while those with lower math anxiety tend to refer to math’s applications (e.g., “engineer” and “statistics”). Reducing math anxiety is a necessary area of work, but significant additional efforts are needed if we want to help people enjoy rather than simply tolerate math.

We went beyond previous work on math attitudes by inquiring about how attitudes change over time. In our sample, those with more negative current attitudes tended to report more variable attitudes over time, owing in many cases to starting out with positive attitudes that declined (while those with positive attitudes had maintained them throughout schooling). Identifying consistently as a so-called “math person” is consistent with these trends: having this identity would mean someone either maintains their positive feelings leading to less variance, or adjusts or selectively remembers their memories of past experiences to be less variable. At the very least, from these results we can see that attitudes are variable and can even change in adulthood for some, which should engender some optimism. We see in this data that people’s present-day attitudes are most closely related to their attitudes from their most recent school experiences, a finding consistent with the peak-end rule (Do et al., 2008) which suggests that how an experience ends dictates how we remember it. This could lead to interventions on attitudes about math with older groups that involve adding positive experiences with math late in schooling or in adulthood or helping recall or re-frame memories about encounters with math. Most of the research on math attitudes focuses on school-aged children because it is believed that this period is crucial for solidifying attitudes about math. In particular, middle school is treated as a time when girls are most at risk for developing negative attitudes that will persist. Our results show that people can adjust their attitudes at any point. Knowing that math attitudes are always malleable allows researchers to focus on how to readjust people’s math narratives which could lead to a populace with more positive attitudes that they can then transmit to children. And, generally making common the knowledge that math attitudes are flexible may have advantageous consequences for people without the need for a researcher-assisted intervention.

## Limitations

A limitation of this study is that it relies on peoples’ memories, which are inherently imperfect. We are not interested in an accurate retelling of experiences, but rather a lens into how people remember their experiences. This might not perfectly match how those experiences actually unfolded, but what people remember does dictate how they currently feel. The fact that we obtained findings consistent with previous research that does not rely on human memory lends credibility the use of our methodology. In future work, we hope to administer similar surveys to students at various stages in their education to compare to memories of experiences in a cross-sectional analysis: though there will likely be generational differences (e.g., use of technology, curricular changes), we would be able to explore the differences in recent versus more distant memories.

The particular choice of computational text analyses we used also comes with some caveats. The bag-of-words representation of Naïve Bayes classifier deals better with smaller datasets and does not identify bigrams (two-word phrases) that might be more informative than single words. More generally, neither computational text analysis approach that we used accounts for word order beyond two word sequences or the grammar or structure. The classifiers and topic models served well for our exploratory analyses and are relatively

simple to implement and interpret, but other textual analyses will likely yield even more observations than what we accomplished.

## 2.6 Conclusion

This research differs from earlier work on math attitudes by starting from a neutral vantage point in the wording of our questions and examining trends in people's experiences across time periods. By providing a fuller picture of people's narratives about their own experiences and how they vary over time, we can both address fundamental questions about what types of salient experiences are related to later positive or negative math attitudes and inform future applied research about educational interventions aimed at improving math attitudes. Applying computational text analyses is an efficient way of analyzing responses from open-ended questions and are especially useful when working with large datasets. One observation that came from deep engagement with these data were differences in the types of math participants chose to forefront in their narratives. Some mentioned early influences like fractions and algebra, while others mentioned applications of math in other scientific disciplines. The next chapter experimentally explores individuals' conceptions of math and the hypothesis that these differing notions of math may impact emotional experiences of the subject.

*Math city was a grid of hexagons.*

Nnedi Okorafor, *Binti*

## 3

# Math Conception

*What do you think math is?*

...then you just do some math.

---

Tom Griffiths

WHAT DO PEOPLE THINK OF WHEN THEY THINK OF “MATH?” My collaborators and I propose that individuals may have very different working definitions of the category of math, and that those with broader *math conceptions* may have less negative feelings about math or be more prone to partake in activities that they believe involve math. In a series of studies described here, we show that “math” is conceived of in very diverse forms by individuals regardless of their age, gender, and country of origin. We introduce a method for indexing the “breadth” of individuals’ math conceptions where participants rate whether various activities *could involve math*. Then we explore possible interventions to broaden conceptions of math and ultimately lead to more positive attitudes towards the subject.

## 3.1 Introduction

Recent U.S. initiatives in early science, technology, engineering, and math (STEM) highlight the growing importance of STEM education (e.g., White House Press Briefing, 2016), as well as the need for professionals in those fields to better represent the population. However, multiple barriers to an educated and diverse STEM workforce remain. One such barrier is psychological: an estimated 25–50% of U.S. college students are math anxious (Jones, 2001; Yeager, 2012), with women disproportionately affected (Hembree, 1990a), which I detailed in depth in Chapter 2. *Math anxiety* refers to the tension or fear associated with the prospect of doing math (Ashcraft, 2002). In addition to being associated with lower math performance,

---

This chapter is partially derived from Foushee et al. (2017).

math anxiety causes math-anxious individuals to generally avoid math. Given the national goal of broadening STEM participation, *math avoidance*, the deliberate avoidance of math-related endeavors (Choe et al., 2019), might be the most devastating byproduct of anxiety about math, as it implies that math-anxious individuals will choose to end their formal math training as soon as possible and avoid careers that involve math.

Here, we are interested in how individuals' ideas of what *math* is—i.e., their *math conceptions*—might be a factor in their math anxiety and avoidance. “Math” can be used to refer to a wide range of activities, involving diverse skill sets and forms of reasoning. Individuals may differ in how they implicitly define the category of math, however, and properties of those definitions may be linked to their math anxiety.

Of particular interest in the present studies is what we will call the “breadth” of an individual's math conception. Guided by the idea that category structures can differentially license inferences (Ross & Murphy, 1999), our studies test the hypothesis that having a working math conception that is narrow (i.e., limited to a few branches of the math taxonomy, like arithmetic operations and numeric notation) might facilitate generalization of negative associations across the category. If this makes individuals confident about disliking *math*, rather than disliking only *arithmetic* or *algebra*, it could make them wary of future topics labeled as “math” that might have otherwise been appealing. In contrast, anxiety about the math category, and any new topics that are labeled as “math,” might be harder to maintain if it encompasses many diverse subtopics and skills, ranging from the concrete (e.g., algebraic notation) to the abstract (thinking about infinity). In other words, insofar as math anxiety consists of anxiety generalized across the category of things construed as *math*, having a “broad” math conception may serve as a protective factor against the propagation of math anxiety.

In the next section, I give an overview of the stimuli used across our studies as we developed our own measure of math conception and explored multiple ways to access attitudes about math (anxiety and avoidance). Following this, I present three studies (Studies 1a, b, and c) evaluating whether adults and children have different conceptions of what counts as math, and whether individuals with broader math conceptions may be less susceptible to math anxiety, such that math conception breadth and math anxiety will be inversely related. Next, I describe two intervention studies (Studies 2 and 3) designed to test whether an intervention on conception of math could result in both broader conceptions and reduced feelings of math anxiety. This project began with the intention of relating math conceptions to math anxiety, but because anxiety is a narrow way of conceptualizing attitudes about math (see Chapter 2), we later turned to more behavioral measures (specifically math avoidance). This chapter is not presented in chronological order of data collection, so our measure of mathematics avoidance is used in Study 1c and otherwise math anxiety is the attitude-related measure of choice.

Could each activity or topic below involve math? Please drag each to the appropriate box:

	Not Math	Math	Unsure
Seeing			
Scheduling			
Physics			
Eating			
Playing the piano			
Playing soccer			
Programming			
Math			
Having a conversation			
Cleaning			
Geometry			
Thinking			

Figure 3.1: A screenshot of the math conception measure. Participants dragged each item into the “Math” box if they believed it could involve math.

## 3.2 Stimuli design

The focus of these studies was to compare math conception with math attitudes, though in many of the studies, we administered other surveys for exploratory work.

### Math conception measure

To measure breadth of conception, we developed a sorting task where participants were given a list of items, each of which they dragged into a box if they believed it could involve math or a separate box if they thought otherwise (Figure 3.1). The items were chosen to elicit different responses: for example, some we believed were “obviously math” like *geometry* or *finance*. We selected others that could be related to different mathematical principles, like counting, problem-solving, spatial reasoning. We believe that anything could involve math if looked at from specific perspectives, e.g., soccer could involve math when counting how many goals it will take to win or determining the best angle at which to kick the ball.

The list of items continually evolved to eliminate redundancies (e.g., cooking and baking were not both necessary in one survey) and adapt the study for different age groups (young children might not know what architecture is) and cultures (cricket is a more common sport in India). The task differed in format for the younger groups of participants who could not perform a computer-based task (described in more detail in the relevant sections).



## Attitudes about math

To study math attitudes, we began by surveying participants about their experience of math anxiety. We started with the single item math anxiety scale, or SIMA (Núñez-Peña et al., 2013) (“On a scale from 1 to 10, how math anxious are you?”) and later used a longer survey for adults, the abbreviated math anxiety rating scale, or AMARS (L. Alexander & Martray, 1989). Our results were not robust enough to continue down this line of inquiry and we also perceived math anxiety as capturing a narrow perspective of math attitudes, so we adjusted course to measure math avoidance. This, like the measure of conception, did not already exist, so we present one methodology in the study with young children (Study 1c).

## 3.3 Study 1: Math conceptions of adults and children

### Study 1a: Adult math conceptions

#### Participants

A total of 62 U.S. adults were recruited via Amazon’s Mechanical Turk (31 female, 19–74 years,  $M = 33.24$ ,  $SD = 10.25$ ). Participants were compensated for their participation, and the study took approximately 15 minutes to complete.

#### Stimuli & methods

##### Math conception

In one block, participants saw a randomized list of topics and activities (e.g., “architecture,” “cooking,” “exercising”). Participants were asked to *indicate whether... each activity or topic listed involves math or does not involve math*. Note that in this first study, we asked whether a specific item “involves math or does not,” though later revised our methodology to ask if an item “could involve math or not.” They responded by dragging each item into one of three boxes, labeled “Math,” “Not Math,” and “Not Sure.” The more items categorized as involving math, the broader we considered their math conception to be. We included the item “Math” as a control.

##### Activity skill

In another block, participants saw the same items in a new randomized order, and rated their skill at each item (*How good would you say you are at each of these things?*). They responded on a five-point Likert scale from ‘Not at all good’ to ‘Very good.’ We included a control item (*For this question, respond ‘Good’*), as well as an opt-out scale option (‘NA’) for participants who had no experience with the item.

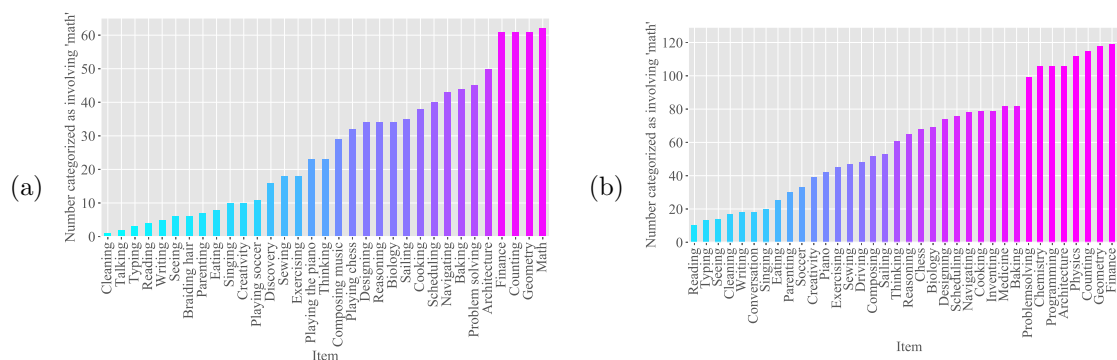


Figure 3.2: Number of adult participants in Study 1 who labeled each activity as involving math (a) and in a replication with more items (b).

## Math anxiety

We assessed participants’ math anxiety using the single item math anxiety scale (SIMA) (Núñez-Peña et al., 2013). The SIMA has been validated on a large sample of U.S. college students. It shows the expected negative correlation with math achievement measures, high test-retest reliability, and is consistent with lengthier, established measures of math anxiety, like the Shortened Math Anxiety Rating Scale (sMARS) (L. Alexander & Martray, 1989).

## Other measures

We collected several other measures of participants’ attitudes toward and history with math. One block assessed participants’ “math mindset:” an analogy to intelligence mindsets made specific to math (Yeager & Dweck, 2012). Five items probed participants’ beliefs about the fixedness of math ability (e.g., *Math is a gift: you either have it or you don’t.*), which they responded to using a five-point Likert scale of agreement. Two blocks consisted of a single, open-ended question, one asking participants for an informal definition of math (*Please describe what you think math is in the space below*), and one eliciting their personal math history (*Please write a brief summary of your experience with math from childhood until now*). In the final block, we collected demographic information, including the number of semesters of college they had completed, and a list of all math classes they had taken.

## Results & discussion

Descriptive statistics about the measures of conception, math anxiety, self-assessed skill, and math mindset can be found in table 3.1.

## Qualitative variation in math conceptions

There was substantial variation in the activities that participants categorized as involving math (Figure 3.2). All participants appropriately responded that “Math” involved math,

which we took as confirmation of their attention to the task. Items obviously involving math were categorized as such by the vast majority of participants (e.g., finance), while those representing related disciplines (e.g., biology), daily activities (e.g., cooking), and abstract, creative and language-related tasks (e.g., composing music, reading) received the fewest math-categorizations. In a separate unpublished pilot study, we elicited explanations for participants' categorizations of a similar list of items. In that study, both adults and children frequently used contrast categories (e.g., "No, that's *music!*"), often from the humanities, to explain why items could not involve math. This type of explanation implies that participants perceived the categories of music, art, and even science as exclusive with math. Such a picture of what math is (and isn't) is consistent with the idea of a narrow math conception, and echoes what mathematician Paul Lockhart famously lamented as the sorry byproduct of American math education:

The first thing to understand is that mathematics is an art. The difference between math and the other arts, such as music and painting, is that our culture does not recognize it as such. [...] Nevertheless, the fact is that there is nothing as dreamy and poetic, nothing as radical, subversive, and psychedelic as mathematics. It is every bit as mind-blowing as cosmology or physics (mathematicians conceived of black holes long before astronomers actually found any), and allows for more freedom of expression than poetry, art, or music (which depend heavily on properties of the physical universe). Mathematics is the purest of arts as well as the most misunderstood. (Lockhart, 2009).

### Math conception & anxiety

Variable	$M$	$SD$
Items Categorized as Math	13.10	5.35
Math Anxiety	4.44	3.04
Self-Assessed Skill	3.28	0.44
Math Mindset	2.13	0.99

Table 3.1: Descriptive statistics for four blocks in Study 1. 'Items Categorized as Math' is out of a total of 32, and was analyzed as a proxy for the breadth of participants' math conceptions. 'Math Anxiety' is on a 10-point self-report scale. 'Self-Assessed Skill' represents the mean skill rating on a 5-point Likert scale, across all items for all participants. 'Math Mindset' is coded to be on a 5-point scale indexing how fixed individuals believe math ability to be, with larger values indicating more fixed mindsets.

To answer whether breadth of math conception and math anxiety are related, we conducted a linear regression on individuals' math anxiety and the number of items they categorized as math, controlling for the number of semesters of college they had completed. In

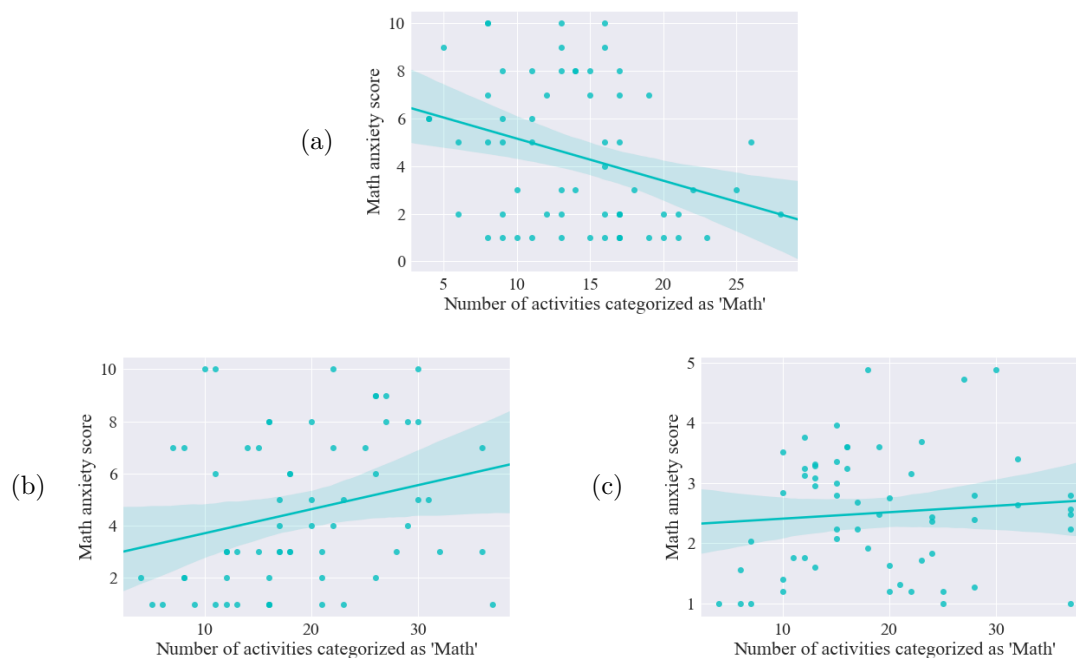


Figure 3.3: Plot of linear regression line showing relationship between breadth of conception and math anxiety in Study 1a, controlling for education ( $\alpha = 6.93$ ,  $\beta = -0.18$ ,  $p < .05$ ) and in a replication with the SIMA (b) and the AMARS (c) measures of math anxiety.

accordance with our predictions, math anxiety was negatively related to the number of items participants categorized as math, even controlling for education ( $F(1, 61) = 6.44$ ,  $p < .05$  with an  $R^2$  of .082; see Figure 3.3a). This appeared to support the idea that individuals with broader math conceptions are less likely to experience math anxiety, and that this relation may not be attributable to exposure to topics in math alone. However, we noted a very small effect size and found no relationship between conception and anxiety in a replication (Figures 3.3b and 3.3c). In this replication, we expanded the number of items slightly in the math conception measure (to 37 from 32) and randomly assigned participants to respond to the SIMA ( $n = 64$ , mean SIMA: 4.55, mean concept score: 19) or the AMARS ( $n = 59$ , mean AMARS: 2.50 out of 5, mean concept score: 19) to assess their math anxiety. In both conditions, there was no relation between concept and anxiety and we did not include other measures such as self-assessed skill as in the original study.

To address whether the relation between math conception and anxiety is due in part to individuals' perception of their own skill at things they think might involve "math," we analyzed self-assessed skill and anxiety. For each individual, we took the mean skill of the items they had categorized as involving math and those they had categorized as not involving math. We dropped items for which participants reported having had no experience. A linear regression on self-reported skill and math anxiety revealed a significant negative correlation between math anxiety and mean self-assessed skill for items the individual was able to relate

to math ( $\beta = -1.98$ ,  $SE = 0.60$ ,  $t = -3.29$ ,  $p < .01$ ), but no correlation between math anxiety and self-assessed skill for items judged to not involve math ( $\beta = 0.11$ ,  $SE = 0.69$ ,  $t = 0.154$ ,  $p = .88$ ). This asymmetry is important because it suggests that it is not just individuals who are less confident overall who suffer from math anxiety—if this were the case, we would have expected to find that lower skill related to higher anxiety for both items judged to involve math and items judged to not involve math.

The direct relationship between math conception and anxiety did not hold up in a replication, but participants' perceived skill at items they believed involved math was related to math anxiety (we did not ask about perceived skill in the replication because our goal was only to confirm whether the weak primary effect was replicable. We intend to replicate this study and look at both perceived skill and enjoyment.) As discussed above, one of the most dangerous features of math anxiety is its tendency to make individuals avoid math and thus fail to take advantage of opportunities to discover new aspects of mathematics they might excel at or appreciate. The fact that mean self-assessed skill at activities categorized as involving math was negatively related to math anxiety lends support to the idea that broad conceptions may be a protective factor in math anxiety, attenuating the impact of negative associations that individuals might have with activities they think could involve 'math.' Having a broad math conception does not mean that an individual has to feel confident and have positive associations with all activities that they think involve math, but it could mean that negative associations with specific topics (like geometry or algebraic notation) will have less of an impact on their associations with the category as a whole.

## Study 1b: Piloting child math conception

This pilot was intended to discover what the youngest age possible could be for running a comparable study with children and to determine the best way to adjust relevant stimuli. We specifically wished to test the youngest children who have the word "math" in their vocabulary to reveal whether this variability in breadth of math conception begins once people start having variation in their math-related educational experiences, i.e., throughout school, if children are being taught the same math content, there might be less variability in their conceptions compared to adults'. If, on the other hand, children who have just started learning about math have non-identical notions of what constitutes math, then very early input may already be influencing young children's ideas about the concept. With these next studies, we wondered whether we could make any preliminary conclusions about the developmental trajectory of math conceptions (barring the ability to conduct a longitudinal study).

### Participants

Nineteen children recruited from a local preschool and a local science museum (4.13-7.48 years,  $M = 5.64$ ) participated. Each participant was brought to a testing room and told they were going to answer questions and receive a sticker for each question.

### Stimuli & methods

In one block, we probed children’s beliefs and attitudes about math in a semi-structured interview to glean how much they might know about math with questions including “Have you heard the word math before?” and “Does math make you nervous/stress you out?” We also asked if their parents and siblings were good at math. In the other block, we showed participants images of children engaged in different activities, and asked: “Is this kid doing math?...Why/why not?”

### Results & discussion

Preliminary results indicate considerable variation in very young children’s math conceptions (Figure 3.4) and math anxiety levels (e.g., five children were ‘nervous’ about math). Children’s qualitative definitions of “math” ranged from broad ones invoking spatialization (e.g., “piano-playing is math because the keys are in a pattern”), to exclusively identifying math with symbolic numbers or traditional manipulatives.

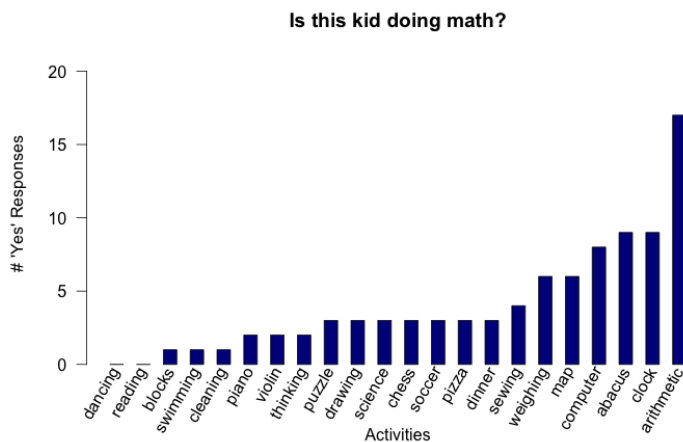


Figure 3.4: Number of participants in Study 1b who answered “yes” when asked whether each kid was doing math.

However, in our exploration of a range of ages, we discovered that many preschool-aged children did not know what the term “math” referred to (when it was unclear to them, we replaced “math” with “numbers”). We decided to continue with children who were at least five years of age, to ensure that all participants could understand the task.

I am interested in mathematics  
as a creative art.

---

G.H. Hardy,  
*A Mathematician's Apology*

## Study 1c: Child math conception

### Stimuli & methods

Studies 1a and 1b show that young children and adults have variable definitions of mathematics. In this study, we ask whether how broadly or narrowly a child defines math is related to their avoidance of mathematics. We additionally had the opportunity to probe participants' parents' beliefs about math because children were primarily recruited from a local science museum and had to be accompanied by a consenting guardian to proceed. As a bonus, assigning a separate task to parents prevented them from directly influencing their children's responses throughout the survey.

**Conception** We designed this measure as a tactile sorting task to make it both child-friendly and parallel to the sorting task performed by adults in Study 1a. We chose "art" as an equivalently vague and variably-defined domain to compare to math in a within-subjects design. Participants sorted images of children doing various activities as math (/art) or not. The experimenter then focused the child's attention on the items they did not initially describe as involving math (/art), asking "could this kid be doing math (/art)?" (see Figure 3.5 for examples). To generate internally consistent images for the sorting task, we photographed a pair of twins (one boy and one girl) performing all of the same activities and used the images of the child that corresponded to the gender of the participant (to avoid any potentially gender-influenced choices in the sorting task).

**Avoidance** Participants responded to a measure of avoidance wherein they were asked to choose between two games they would prefer to play based off of stimuli from Bian et al. (2017). One game was 'for children who are really, really good at math,' while the other was for children who excelled at art. Both games were described verbally and an image was shown.

**Parent survey** To see if parent conceptions and attitudes might be related to those of their children, we gave parents a quick survey that mimicked the task presented to child participants. They were asked to circle which items they believed involved math (/art) (the same items that their children saw) and then asked to rate their math (/art) anxiety on a single-item scale. We did not have an avoidance measure ready for adults and wanted the survey to be short so they would be willing to complete it.

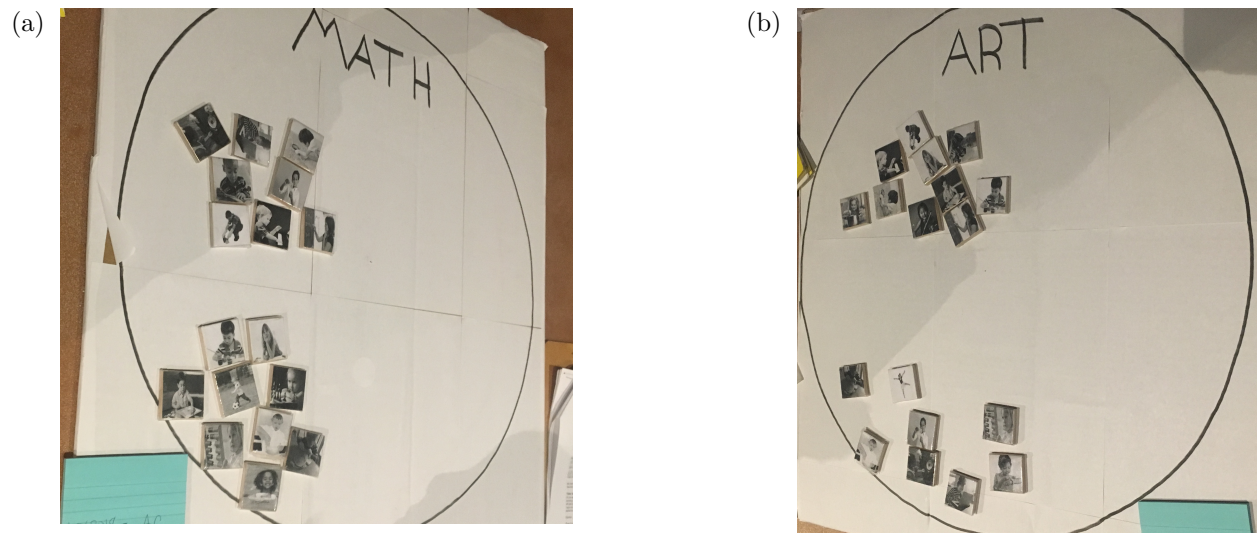


Figure 3.5: Example sorting tasks about math (a) and art (b). The items are separated into two clusters, one of which contains the activities participants believed involved math in the first round of sorting, while the second cluster includes activities the participant believed *could* involve math in a second round of sorting.

## Participants

Stimuli were presented with one task type first (conception or avoidance) and one domain first (math or art). For example, if a participant completed the math conception sorting task first, they next completed the art conception task, and in the avoidance task the first described game was for kids who were “really, really good at math.” Participants ranged from 5–7 years of age and we collected data such that there were 6 children (3 boys and 3 girls) tested from each of the three age groups (5, 6, and 7 years of age) for each of the four possible orderings of stimuli (conception vs. avoidance first and math vs. art first). The goal was therefore to collect data from 72 total participants.

We collected data from a total of 103 participants, ages 5–7 years old. We excluded 33 participants for failing an attention check where they were shown three images of kids playing and asked “is this kid playing outside?” to make sure they understood the premise of the sorting task. Out of the remaining 70 children, 2 of them did not categorize the image of a child writing a math equation as “math” (though 2 said it “could be math”) and 3 did not categorize the image of a child drawing as “art.”

## Results

Both parents and children had very diverse conceptions of math and of art (Figure 3.6). They all ranged from categorizing only one or two activities as math/art to categorizing all 18 activities as math/art (Figures 3.7 and 3.8). From Table 3.2, it is clear that parents categorized more activities as math/art on average. There was a substantial subset of parents



who categorized all activities as math/art (Figure 3.8). This suggests that both math and art conceptions may broaden with age as we are exposed to more applications of math in the world.

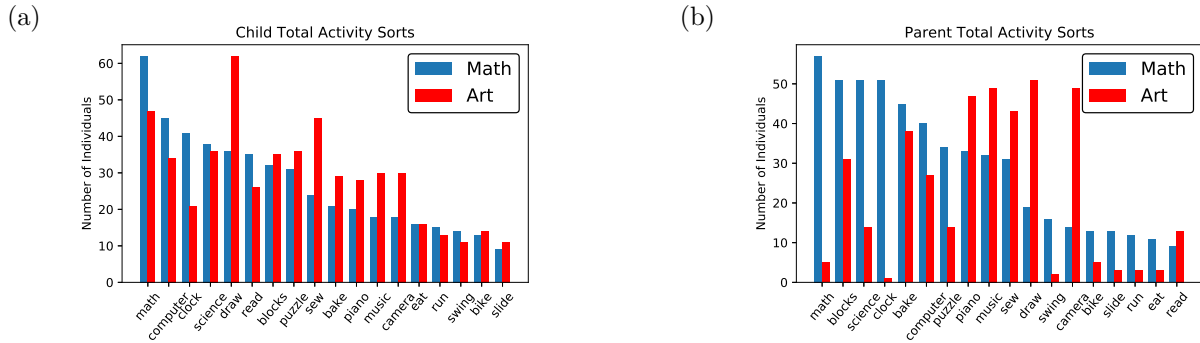


Figure 3.6: Histogram of (a) number of participants and (b) number of parents who categorized each activity as math and art. Items are sorted from most math sorts to least in each figure.

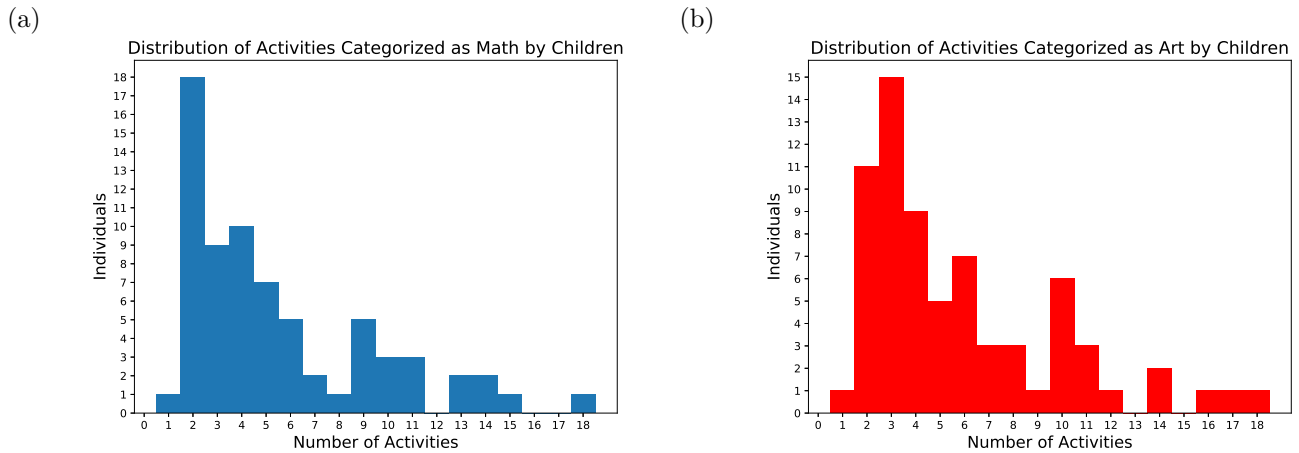


Figure 3.7: Histogram of number of activities categorized by children as (a) math and (b) art.

It is also clear that the difference between the number of times an activity was categorized as math and the number of times an activity was categorized as art is much larger for parents than children. This could mean that adults are more likely to categorize activities into either math or art, and could also imply that there is more of a collective understanding with adults about which activities are math and which are not.

The number of children who chose the game for those who excelled at math was comparable to the number who chose the art game (math: 38; art: 32). However, we noticed a few participants listed irrelevant reasons for their decision, including the image provided of

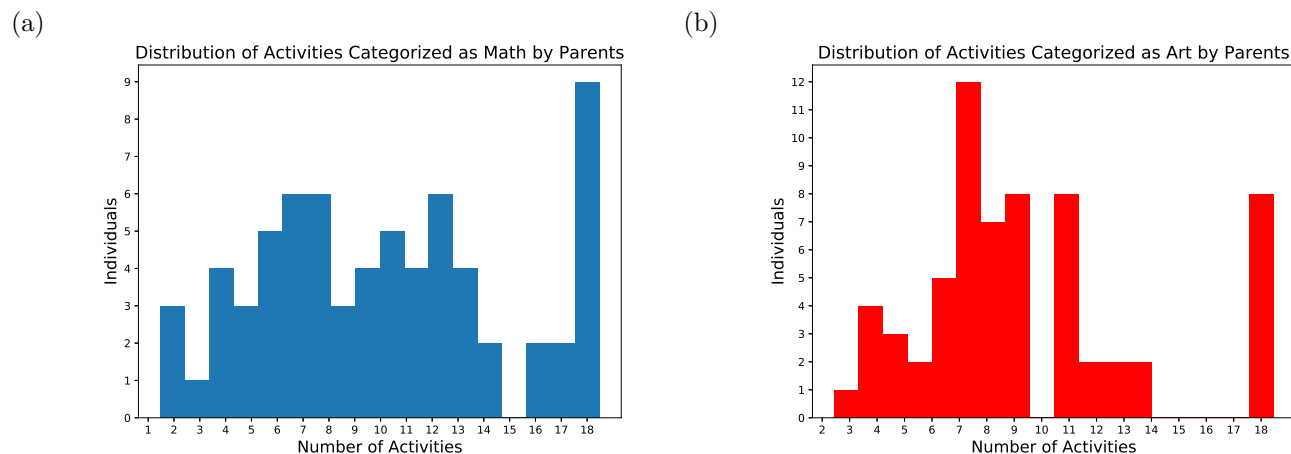


Figure 3.8: Histogram of number of activities categorized by parents as (a) math and (b) art.

	“IS MATH/ART”	“IS” + “COULD BE”	PARENTS
MATH	4.53 (3.95)	7.93 (5.18)	11.49 (5.09)
ART	4.86 (4.04)	8.30 (5.33)	10.86 (4.22)

Table 3.2: Mean number of activities categorized as math or art by children and parents (all out of 18).

the game looking more interesting. That said, the majority, if they provided an explanation, stated that their choice was based on their perceived skill in math or art.

We collected math and art anxiety scores from the parents, and found that, across all parents, the mean math anxiety score was 3.87(2.79), while the mean art anxiety was very comparable at 3.48(2.50) (out of 10). The parents of children who chose the math game had an average math anxiety score of 3.33(2.30) and art anxiety rating of 3.23(2.36). For parents of children who chose the art game the mean math and art anxieties were 4.48(3.18) and 3.77(2.67) respectively. These are such small differences though they do point in an interesting direction: the parents of children who chose the art game (so who avoided the math game) expressed higher levels of math anxiety compared to art anxiety and higher math anxiety compared to the other parents. This appears consistent with results that suggest that math anxious parents impart their math anxiety to their children (Maloney et al., 2015). As expected, children who chose the math game ( $n=38$ ) identified more activities as math than art, while those who chose the art game ( $n=32$ ) identified more activities as art than as math, indicating that preference for a domain may be related to having a broader conception of it.

## Discussion

We have introduced the idea of *math conceptions* to describe these qualitatively different definitions of the category of math, and focused especially on their “breadth” to explain why certain types of math conceptions might influence attitudes about math. This sequence of studies highlights that adults and young children both have diverse conceptions of “math.” What’s more, children’s conceptions of art are just as variable. Their parents have even broader conceptions of both domains (a number of adults categorized all activities as math/art), though they separated these concepts more than did the children (items frequently identified as involving math were not categorized as involving art and vice versa).

Our initial hypothesis connecting math conception directly to math anxiety was not supported, but we hoped that more behavioral measures (preference for or avoidance of math) would be related to breadth of conception, especially when these related to an individual’s skill in a given activity. In the next two studies, we explore whether it is possible to intervene on math conception and subsequently attitudes about math.

## 3.4 Study 2: Intervention on middle school students in India

We were interested in whether students in middle school would exhibit the same qualitative variation in math conceptions that we had seen with adults and children in Studies 1a, b, and c given that at this point students have been in the same schooling environment and exposed to an identical set of mathematics concepts, though their environments beyond school may differ. Additionally, as a first pass at investigating the causal relation between math conception and attitudes with this population, and potential educational implications, we designed a brief intervention intended to broaden students’ math conceptions consisting of a hands-on activity (origami) that we described as involving math.

We find evidence supporting the hypothesis that students in school exhibit significant variation in their math conceptions, but not quite for the effectiveness of our conception-based intervention, though the results are encouraging. We discuss why this intervention may have been ineffective.

## Stimuli & Methods

Study 2 consisted of an interactive origami activity followed by four measures administered to participants in two between-subjects conditions, BASELINE and BROAD. Only participants in the BROAD condition received an explanation for the ways in which the activity had involved math before completing the other assessments.

## Participants

A total of 80 6th, 7th, and 8th grade students at a school in Gujarat, India participated (33 6th-graders, 7 girls; 21 7th-graders, 9 girls; 26 8th-graders, 9 girls). We were interested in running this study with this population in India because previous work had been done there examining students' use of mental abacus (Barner et al., 2016a) and their are very strong cultural beliefs that math is very important. We wondered if we might see much broader conceptions compared to American children and we also desired a sample situated between kindergarten and adulthood. All 6th-grade participants were excluded for sharing answers ( $n = 33$ ), leaving 47 7th-8th grade students in our sample. Participants were tested in groups of 10–15 assigned to the BASELINE or BROAD conditions in a classroom at their school.

## Origami activity

Students sat in a circle on the floor around two experimenters who guided them through folding an origami crane. A third experimenter circulated to answer any questions, and students could also refer to printed, diagrammatic instructions distributed before the activity. All experimenters avoided using explicit math language during the folding instruction (e.g., reference to “angles,” “half,” “diagonal”), opting instead for generically narrated demonstration (e.g., “fold the paper like this”). Each student folded a paper crane, which they got to take home.

## Construal

Following the origami activity, students in both conditions answered whether the activity they just did *could involve math* (Yes/No/Not Sure), and to explain why. In addition, they rated how *enjoyable* and *difficult* they had found the activity, on a five-point Likert scale (from ‘Not at all—’ to ‘Extremely—’).

## Intervention

In the BROAD condition—but not in the BASELINE condition—an experimenter then gave a brief explanation of how the origami activity involved and related to math (e.g., . . . *you have to think about **spatial relations**, and things like **measurements** of the different **sides and angles**. When designing new pieces of origami, you have to think **creatively and flexibly**, and use what you already know to come to new conclusions, **like you have to do in math***). Following this, we asked students whether they thought the activity could involve math and had them write explanations for their answers. They also rated both how fun and how difficult they found the activity to be. We realized that a one-time quick intervention would likely not substantially and immediately influence such entrenched beliefs about a concept or anxiety, but if effective, it would allow for the development of simple and efficient interventions. Those in the baseline condition did not have anything read to them. Then,

all students completed three more surveys: a 6-item avoidance measure, a 16-question math anxiety survey, and a 40-item math conception questionnaire, explained in detail below. We hypothesized that those in the broad condition would have higher scores on the math conception questionnaire (indicating broader math concepts) and lower math anxiety scores, attesting to the possibility of broadening a child’s math conception and reducing their level of math anxiety.

### Avoidance

The next measure participants completed was intended to indirectly access their math avoidance. The survey consisted of 6 items, each asking about a different school subject (e.g., *How excited are you to learn a new topic in [math/Hindi] class?*). Participants responded on a 5-point scale (from ‘Not at all excited’ to ‘Extremely excited’).<sup>1</sup>

### Math anxiety

We administered a child math anxiety questionnaire adapted from Ramirez et al. (2013) by Barner et al. (2016a), for use in India. The questionnaire consisted of 16 questions regarding students’ experiences with math, which students responded to using a 5-point face scale (from ‘Not nervous at all’ to ‘Very, very nervous’). The experimenter explained the scale and completed three warm-up questions with the students beforehand to ensure understanding of the measure. The students responded to 16 more questions on a 5-point scale from “Not nervous at all” (1) to “Very, very nervous” (5), such that lower scores indicate less anxiety for a particular item.

### Math conception

The math conception measure was a variant of the one used in Study 1. We included 40 age- and place-appropriate items, and adjusted the wording used in the prompt from Study 1. Here, participants answered *Could this activity involve math?* (Yes/No/Not Sure), which we anticipated would encourage flexible thinking about the items and about math. Because of constraints of the testing site, we were not able to make this a sorting task as in the previous studies or as about enjoyment of or skill in each activity.

## Results

### Qualitative variation in math conceptions

Participants indicated that an average of 22.74 out of 40 items *could involve math*. As in Study 1, there was considerable variation across items in the proportion of participants who judged them as involving math (Figure 3.9).

---

<sup>1</sup>Because participants were on average enthusiastic to learn new topics in math ( $M = 4.29$ ,  $SD = 0.94$ ), more so even than other topics, we did not further analyze the results of this measure.

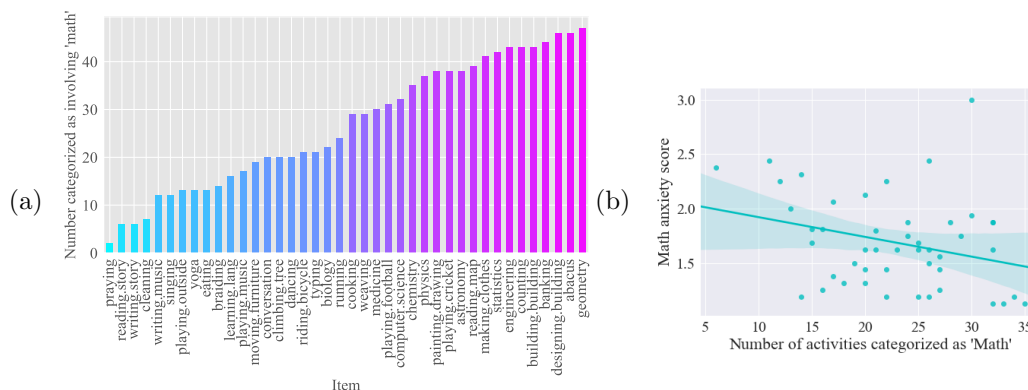


Figure 3.9: (a) Number of participants in Study 2 who answered “yes” when asked whether each item *could involve math* and (b) linear regression showing relationship between breadth of conception and math anxiety in Study 2 ( $\alpha = 2.10$ ,  $\beta = -0.02$ ,  $p < .05$ ).

Condition	Concept		Anxiety	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
BASELINE	20.96	7.24	1.78	0.47
BROAD	24.61	5.96	1.60	0.36

Table 3.3: Mean concept and anxiety scores and standard deviations for each condition.

### Math conception & anxiety

Participants received an average math anxiety score of 1.69 (out of 5). We were interested again in whether math anxiety scores were related to conception breadth, which we examined in our total sample, collapsing across condition. In middle-schoolers, as with adults in Study 1a, math anxiety was negatively related to the number of activities students categorized as *math*, ( $F(1, 44) = 4.15$ ,  $p < .05$  with an adjusted  $R^2$  of 0.07; see Figure 3.9b), though again this effect was small.

This provides further support to the idea that individuals have different conceptions of what counts as math and that individuals with broader math conceptions are less likely to experience math anxiety. It is also interesting that this is the case with children in middle school, demonstrating that even during school years, students do not have a consistent breadth in their conceptions of math.

### Conception intervention

We next analyzed math conception and math anxiety for our two conditions separately. If such a brief intervention were successful, we should expect conception scores to be higher in the BROAD condition, and anxiety scores to be lower. While conception and anxiety were

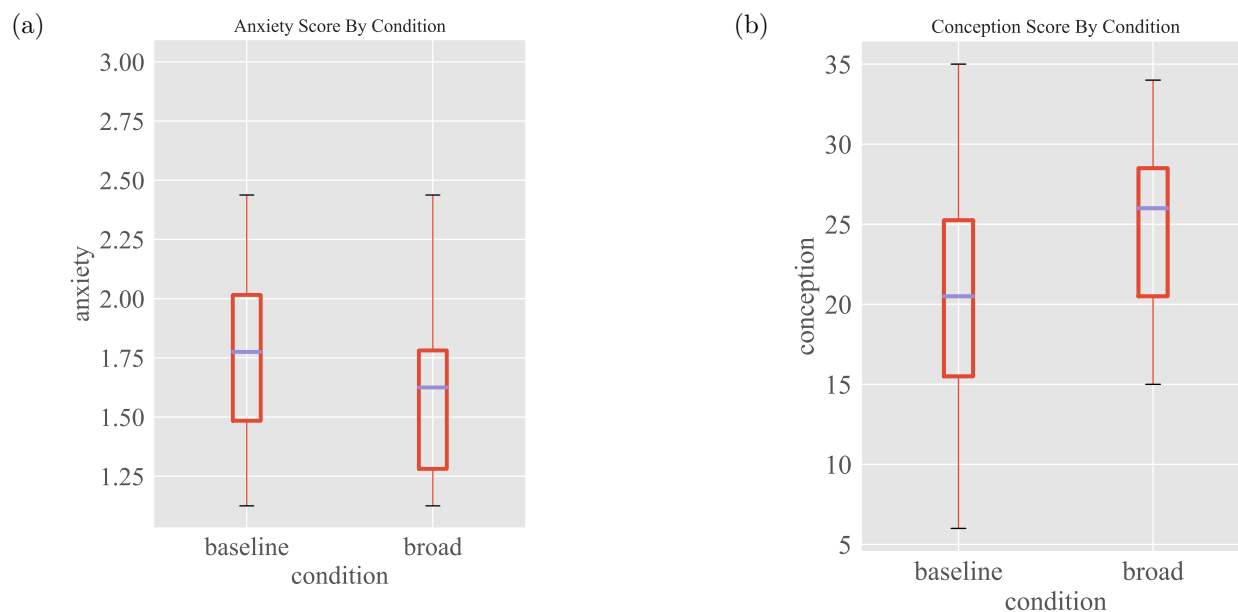


Figure 3.10: (a) Boxplot of anxiety scores by condition. (b) Boxplot of conception score by condition.

slightly different in the anticipated direction between the two conditions (in the baseline condition, the mean conception score was 20.96 and the mean anxiety score was 1.78 while in the broad condition, the mean conception score was 24.61 and the mean anxiety score was 1.60, see table 3.3), the differences between group means is not significant (as determined by one-way ANOVAs for math conception:  $F(1, 45) = 3.54$ ,  $p = .066$ , and anxiety:  $F(1, 45) = 2.31$ ,  $p = .14$ ). The trend for math conceptions in particular is promising (see Figure 3.10b): in the BASELINE condition, there was more spread in the magnitude of participants' conception scores, while those in the BROAD condition had generally 'broader' conceptions. Thus, it may be that with a different or merely more sustained intervention, students' math conceptions could be broadened.

### Influence of construal

Out of the 47 participants analyzed, 36 said that the origami activity could involve math (asked prior to the intervention in the BROAD condition). Participants on average enjoyed the activity ( $M = 4.43$ ,  $SD = 0.62$ ) and did not find it difficult ( $M = 2.28$ ,  $SD = 0.71$ ). This raises the possibility that we may not have found a robust intervention effect because our elicitation of construals of the origami activity as *math* itself served as an intervention on breadth of conception. In particular, given that all participants—including those in the BASELINE condition—were asked to consider whether an enjoyable and easy activity could involve math before completing any of the surveys, they may have been primed to think more broadly and favorably of math. Additionally, the anxiety measure was designed with

much younger children in mind, causing a floor effect for these older students.

## Discussion

The above study offers additional evidence for the intuition that individuals, regardless of age and even potentially country of residence, may have substantially different ideas of what constitutes *math*.

While the link between our measure of math conception breadth and math anxiety was found in this study, we do not seem this effect as particularly robust and would like to explore perceived skill and enjoyment again with a comparable population. Eliciting and analyzing participants' explanations for their categorization decisions may be one especially fruitful way to access other qualitative dimensions of math conceptions, alongside canonical methods to access category structure, like primed similarity judgments (Shepard et al., 1975). Size is not the only dimension along which a category may vary, as participants' explanations for their math categorizations in our small-scale pilot study reveals.

Without robust evidence for the efficacy of our intervention, we cannot speak to the potential directionality of the math conception-attitudes relationship. Our ongoing work is exploring this question through an interactive intervention on adults' math conceptions (described in the next section), as well as an adaptation of the math conception measure for use with bilingual children.

I think that mathematicians are  
actually artists.

---

Gregory Chaitin,  
*Conversations with a  
Mathematician*

## 3.5 Study 3: Intervention on adults in a science museum

In a second intervention study, we designed a intervention on math conception for adults attending evening events at the Exploratorium<sup>2</sup> of San Francisco using their exhibits in collaboration with members of the museum's visitor research group. The primary goal was to follow up on Study 2 and see if math conceptions can be intervened on in a more naturalistic setting.

---

<sup>2</sup><https://www.exploratorium.edu/visit/calendar/after-dark>



## Stimuli & methods

Participants were first directed to an intervention condition using one of two adjacent exhibits and then to a laptop to complete a series of surveys.

### Intervention

Participants were assigned to the intervention or control condition and to active or passive presentations of either (four total conditions). All signage on the exhibits was covered up. The passive presentations consisted of watching a video, either about the beauty of math<sup>3</sup> (intervention condition) or a control video of similar length (also without narration, a scene from a nature documentary). The active conditions involved playing with an interactive exhibit called Flock which simulates flocking patterns of birds in flight. After three minutes of exploratory time, participants in these conditions read a description of the exhibit that either described it in mathematical (“...mathematical model of birds in flight...”; intervention condition) or non-mathematical terms (“...social nature of bird flight...”; control condition).

### Math conception & anxiety

Next, participants completed a 36-item math conception questionnaire with sorting and the AMARS, a 24-item self-report measure of math anxiety. We chose this longer math anxiety measure rather than the SIMA to see if it related more robustly to math conception.

## Participants

A total of 181 adults attending the weekly Thursday evening event at the Exploratorium consented to participate in our study (88 women, 88 men, 3 other, 2 declined to state; 104 white, 36 Asian, 5 Hispanic or Latino, 4 Native Hawaiian or Pacific Islander, 5 Other, 4 selected multiple, rest declined to state). After excluding 4 participants who failed control questions (didn’t categorize “math” as involving “math”), 10 who had previously interacted with the relevant exhibits, 51 who were not educated in the US, and 2 who left responses blank, we had 115 participants remaining for analyses. We decided to exclude those who were not educated K–12 in the US because all of our participant pools thus far were country-specific and we believe that educational experiences are likely to influence math conceptions. We were stationed at exhibits near the entrance of the museum so many adults participated prior to engaging with the rest of the museum.

## Results

As in the previous studies, math conceptions were diverse, though notably math conceptions were broader (mean conception was 27 out of 36 or 75% ( $SD = 6.66$ ) whereas in the original

---

<sup>3</sup><https://vimeo.com/77330591>

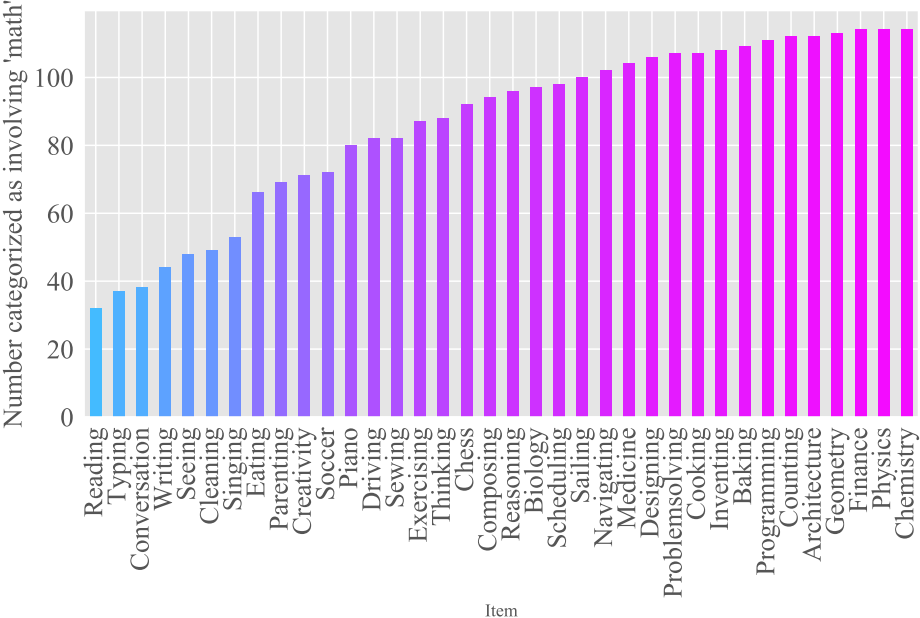


Figure 3.11: Number of participants who categorized each item as involving math in Study 3.

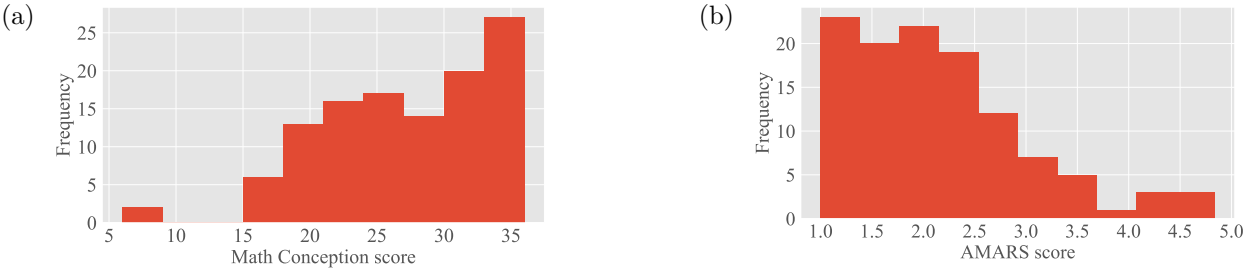


Figure 3.12: (a) Histograms of conception scores and (b) AMARS math anxiety scores.

study with adults online, participants sorted an average of 13.1 items out of 32 or 41% ( $SD = 5.35$ ) as math).

There were higher math conception scores in both intervention conditions compared to the control conditions (see Table 3.4), but there was no significant relationship between math anxiety and conception as tested with a one-way ANOVA and no significant differences between conditions as determined by independent samples t-tests comparing math conception scores and math anxiety scores across participants in the math or the control conditions.

Math anxiety was notably low for all participants, but there were 32 participants who had high math anxiety (greater than 2.5 out of 5). In this small group, there was an effect of condition on math anxiety ( $F(3, 28) = 4.41, p < .01$ ) as well as on math conception score ( $F(3, 28) = 4.42, p < .05$ ). In this high math anxiety group, post-hoc analyses using

Condition	Math Anxiety	Math Conception
Flock-Math	2.04 (0.62)	27.35 (7.94)
Flock-Control	2.01 (0.73)	25.45 (5.93)
Video-Math	2.34 (1.08)	28.17 (6.01)
Video-Control	2.31 (0.99)	27.24 (6.30)

Table 3.4: Means and standard deviations for math anxiety and math conception scores by condition.

Tukey’s honestly significant difference test revealed that math anxiety was lower and math conception broader for those in the flock-math ( $n = 8$ ) compared to those in the flock-control ( $n = 8$ ) condition, suggesting a possible benefit for the interactive intervention condition. Additionally, math conception was broader for those in the flock-math than in the video-math ( $n = 11$ ) condition. This might suggest a benefit of framing the Flock exhibit as being about math for those who have high math anxiety already. In debriefing discussions with participants, some who reported being math anxious assigned to the video-math condition claimed that the video heightened their feelings of anxiety. These conclusions are limited by the very small sample sizes, but make it seem worthwhile to explore naturalistic interventions with a more heterogeneous sample of adults.

## Discussion

This study revealed that adults in a science museum context still have diverse notions about the concept of math, but generally broader definitions compared to the other populations studied in this chapter. This could be a result of self-selection: adults choosing to spend their evenings in a science museum might represent a sliver of the population that already thinks broadly about math. Alternately, the environment of the museum itself might cause a broadening of math conceptions.

The results of our intervention are limited, but seem to point to further benefits of a more interactive form of intervention: giving participants an activity to engage with positively and then demonstrating or describing its math-relatedness may have a more positive effect than a video making similar claims. In the future, we would like to focus efforts on longer, multi-session interventions, track behaviors within a museum as measures of avoidance, and brainstorm new ways to conceive of the museum itself as an intervention on breadth of conception. From this collaboration, employees of the museum have been exploring how to adapt our conception measure to be about science and rethinking labeling and signage for new and existing exhibits.

## 3.6 General discussion

All of the studies accumulated in this chapter point to one main conclusion: math is conceived of in diverse and divergent ways by people of all ages and in varying contexts. We see the remarkable dissimilarity of the populations studied and their respective contexts as adding strength to this result. Our results suggest that math anxiety is related both to how expansive individuals perceive math to be, and how skillful they feel at the activities they think it could involve. Study 2 attempts an intervention on students' conceptions of math with a sample of middle school students and identifies a weak inverse relationship in students between math conception breadth and math anxiety. Study 3 pursues another type of intervention, this time with adults, and offers new insights into how we might go about refining our intervention design.

The main limitation of this work is in how we chose to measure feelings about math. Math anxiety is the most extensively studied attitude about math (see Chapter 2) and there are many tried and tested measures adapted for all ages. However, we did not find robust relationships between anxiety and breadth of conception and desired more measures (preferably behavioral) to understand what conception might be influencing. We adapted a measure of avoidance for young children in Study 1c, though it offers only a binary choice, which does not necessarily reveal a choice based on avoiding math (i.e., a child who chose to play the art game may be indicating a preference for art rather than an aversion to math). We feel that this was a necessary choice to start exploring more behavioral measures of math avoidance and hope to refine our ideas. A possible measure of math avoidance could make use of the idea of a museum where participants are shown multiple descriptions of exhibits and asked to choose a subset to visit where a select few explicitly involve math.

Exploring math conceptions in young children, as well as directly assessing math skill in future studies with adults, will also address the heretofore unconsidered possibility that a third variable (like actual proficiency in math) is responsible for both responses on our current conception measure and attitudes about math. The ultimate goal of these lines of research is to understand and describe the character of individuals' implicit math categories, and leverage this knowledge to inform interventions aimed at improving attitudes about math in adults and children.

*Life is a combination of logic, feelings, and accidents.*

Gad Hachlili

## 4 Conclusion

If there's one thing I know now,  
it's that anyone looking for  
order ought to steer clear of  
psychology altogether.

---

Olga Tokarczuk, *Flights*

IN THIS DISSERTATION, I explored distinct ways that math is perceived by humans of various ages alongside data that point out how math is discussed generally. I used a variety of methodologies, studied a diverse array of participants, integrated work from multiple domains, all to generally make the point that perceptions are complex and divergent. Learners differ in their prior beliefs about their ability, their competence at error detection, and their responsiveness to their own self-generated evidence (Chapter 1). Humans deviate in the emotional landscape associated with their math experiences and their narrative trajectory (Chapter 2). People of all ages hold variable ideas about where math exists in the world around them and why (Chapter 3). And the language used to discuss math differs by formality of the medium and is influenced by current events (Chapter 0). This story of vast and variable ideas is precisely what I have tried to convey: there is no correct way to perceive one's ability, no prescribed personal journey with a topic, and no exact definition of what we must believe math is.

All of these studies might be considered interventions themselves. Does asking a student to self-assess lead to more accurate future self-assessments? Anecdotally at least, students and educators seem to benefit from being given the opportunity to reflect in this way. Does writing emotional narratives of experiences with math help people process their feelings and move past them? Some research suggests that writing can alleviate negative feelings and allow for better performance on a math assessment (Park et al., 2014). Does probing what students think math is cause them to think more broadly about math? We can build interventions with these ideas in mind, but it might be that the simplest form of self-awareness through reflection already constitutes a subtle intervention (Collins & Brown, 1988).

To come to these conclusions, I demonstrated an assortment of methods for understanding math perceptions, including simple human judgments of a cognitive nature (Chapter 1), open-ended written narratives (Chapter 2), sorting tasks (Chapter 3), and self-report surveys of emotional experiences (Chapters 2 and 3). Analyzing this medley of data types required moving beyond statistical analyses and manual coding efforts and instead turning to probabilistic modeling to dissect simple human judgments of themselves (Chapter 1) and computational text analyses (Chapters 0 and 2) to thoroughly and systematically process vast text-based datasets. I explored contrasts between math and other subjects to further generate insights about what sets math apart from other areas of study, including trivia and reading (Chapter 1), science (Chapter 2), and art (Chapters 0 and 3), but all of these disciplines (and many others) deserve their own conceptual explorations. These comparisons reveal that science has an advantage over math in that references are easily accessible in popular culture (Chapter 2) and art is just as murky of a subject as math (Chapter 3). In the future, I would hope to explore a multitude of perception types across many more domains.

Alongside these math-perception-related conclusions, my study design and analysis choices have demonstrated other considerations generally in the field of cognitive science and research practice as a whole. First, the collection of large datasets as enabled by the existence of platforms such as Mechanical Turk provides much greater clarity in hypothesis testing (and also more diverse samples than simply testing undergraduates). Second, computational modeling can transform a seemingly simple cognitive measure into multiple component parts. Finally, combining experimental work with naturalistic data can expand our understanding of humans and their beliefs about the world.

There is a lot of bias against  
pigeons... that I carry as well.

---

Mahesh Srinivasan

## 4.1 Research into practice

One hope for this body of work is to inform educators, policy makers, parents, and others who might be interested in improving math's reputation. We must ask ourselves: who is in charge of (and most influential to) this discourse and the writing that narrates our story? Writers, journalists, researchers, and anyone disseminating their ideas to an audience have great power over the beliefs we hold. Professional writers may be people who were raised to think that their writing and math abilities were dichotomous and hated and/or struggled with math (another unfortunate TV trope<sup>1</sup>). I read extensively and am regularly disappointed in negative portrayals of math, from known authors like Margaret Atwood (in *Oryx and Crake*,

---

<sup>1</sup><https://tvtropes.org/pmwiki/pmwiki.php/Main/WritersCannotDoMath>

Snowman claims “I am not a math person”) to very recently published books that are helping us make great strides with LGBTQ+ and other forms of representation (for example, in T.J. Klune’s *The House in the Cerulean Sea*, everyone groans when they have to do algebra).<sup>2</sup> It is entirely reasonable to expect writers to share their feelings in this way, but if the population of writers is skewed towards *negative* portrayals of math, then the discourse is necessarily in math’s disfavor. The educated are also great influencers of our math narratives. If the so-called intellectual “elite” are the gatekeepers of advanced mathematical knowledge and make math success appear specific and difficult, how will people from all social strata advance in the subject? None of this is deliberate propaganda – no one decided it would be good for many to fear math (that I know of) but the pervasive kind where people express their feelings, which normalizes them for others and pretty soon we have an epidemic of math anxiety. Simply noticing that we’ve been conditioned to think about math in specific ways is a first step towards adjusting and seeking new narratives. This may be a very culturally-specific experience and I would hope to one day compare math discourse across different communities.

The smallest hint of an idea written down can unleash an overwhelming tide of misinformation. But there are plenty of examples of thorough exposure to ideas that have unleashed more positive thought amongst new generations of individuals: my generation grew up with *Harry Potter* which oriented us toward a social justice lens to the extent that the series’ legacy exists beyond the point of view of its author (Gierzynski & Eddy, 2013). Research has linked media representation and exposure to negative ideas surrounding math (Eccles & Jacobs, 1986), so why not think of reversing this?

So you must not judge what I  
know by what I find words for.

---

Marilynne Robinson, *Gilead*

## 4.2 Final thoughts

This journey began as a specifically targeted attempt to understand why people treat math with such vitriol. Why is math spoken about so differently compared to other areas of study? But my interests have broadened significantly past this one domain and into memory, media coverage, and perceptions of the self. It turns out that math is really a useful case study or comparison domain to understand how our world shapes our internal selves. Many things “start” from math – researchers studying the earliest stages of child development will focus on numeracy (S. Carey, 2000; Xu & Garcia, 2008) and a number of avenues for study rely

---

<sup>2</sup>Fun anecdote: I reached out to T.J. Klune to share my thoughts about his portrayal of math because I loved his book so much and he revealed that this was a part of him placed in the book, though he will try to keep this in mind in future writing!

on a mathematical foundation. But as discussed in this dissertation, conceptions of math are far-ranging and do not ever seem to resolve into one clean idea.

I have seen arguments for the “rebranding” of math<sup>3</sup> – that maybe we should put the word itself out of use and focus on “computational thinking” or another term that is not already tied up with fierce emotion. This, I feel, is a reasonable action in some cases as perhaps the word “math” is simply triggering. However, the beauty of this fraught term is its ambiguity, rich history, and comprehensiveness. I believe that broadening any discussion and presenting alternatives is a better way toward a world that ceases to condemn math so frequently, like how I broadened the measurement of math perceptions in this dissertation. We also must consider the need to neutralize dubious stereotypes, though this is a difficult task. Through the incorporation of inquiry learning into teaching practices and the generation of more opportunities to reflect on and learn about stereotypes, they might begin to become undone. The corrosive effects of anti-mathematics propaganda accumulate. This impudence may not outrage the insentient concept of math, but it certainly prevents many from realizing a latent mathematical ability.

---

<sup>3</sup><http://www.conradwolfram.com/home/toxicmathsbrand>



## Bibliography

- Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., & Gleicher, M. (2014). Serendip: Topic model-driven visual exploration of text corpora, In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE.
- Alexander, L., & Martray, C. (1989). The development of an abbreviated version of the Mathematics Anxiety Rating Scale. *Measurement and Evaluation in Counseling and Development, 22*, 143–150.
- Anderson, J. R. (1990). The adaptive character of thought. *Hillsdale, NJ: Earlbaum*.
- Andrews, R. (2018). This psychological effect explains why anti-vaxxers believe what they believe. *IFLScience*. <http://www.iflscience.com/health-and-medicine/antivaxxers-suffer-from-a-wellknown-cognitive-effect-according-to-study/>
- Arentoft, A., Byrd, D., Monzones, J., Coulehan, K., Fuentes, A., Rosario, A., Miranda, C., Morgello, S., & Rivera Mindt, M. (2015). Socioeconomic status and neuropsychological functioning: Associations in an ethnically diverse HIV+ cohort. *The Clinical Neuropsychologist, 29*(2), 232–254.
- Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science, 11*(5), 181–185.
- Baloğlu, M. (2004). Statistics anxiety and mathematics anxiety: Some interesting differences I. *Educational Research Quarterly, 27*(3), 38–48.
- Barbieri, C. A., & Miller-Cotto, D. (2021). The importance of adolescents' sense of belonging to mathematics for algebra learning. *Learning and Individual Differences, 87*, 101993.
- Barner, D., Alvarez, G., Sullivan, J., Brooks, N., Srinivasan, M., & Frank, M. C. (2016a). Learning mathematics in a visuospatial format: A randomized, controlled trial of mental abacus instruction. *Child Development, 1*–13.
- Barner, D., Alvarez, G., Sullivan, J., Brooks, N., Srinivasan, M., & Frank, M. C. (2016b). Learning mathematics in a visuospatial format: A randomized, controlled trial of mental abacus instruction. *Child Development, 87*(4), 1146–1158.
- Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*(4475), 1262–1264.
- Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391.
- Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist, 47*(1), 46–54.

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
- Boonen, A. J., Kolkman, M. E., & Kroesbergen, E. H. (2011). The relation between teachers' math talk and the acquisition of number sense within kindergarten classrooms. *Journal of School Psychology*, *49*(3), 281–299.
- Breda, T., & Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences*, *116*(31), 15435–15440.
- Burson, K. ., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77.
- Buurma, R. S. (2015). The fictionality of topic modeling: Machine reading Anthony Trollope's Barsetshire series. *Big Data & Society*, *2*(2). <https://doi.org/10.1177/2053951715610591>
- Carey, E., Hill, F., Devine, A., & Szücs, D. (2016). The chicken or the egg? The direction of the relationship between mathematics anxiety and mathematics performance. *Frontiers in Psychology*, *6*, 1–6.
- Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, *1*(1), 37–41.
- Chaitin, G. J. (2002). *Conversations with a mathematician: Math, art, science and the limits of reason*. Springer Science & Business Media.
- Chen, L., Bae, S. R., Battista, C., Qin, S., Chen, T., Evans, T. M., & Menon, V. (2018). Positive attitude toward math supports early academic success: Behavioral evidence and neurocognitive mechanisms. *Psychological Science*, *29*(3), 390–402.
- Chestnut, E. K., & Markman, E. M. (2018).  
Girls are as good as boys at math” implies that boys are probably better: A study of expressions of gender equality. *Cognitive science*, *42*(7), 2229–2249.
- Choe, K. W., Jenifer, J. B., Rozek, C. S., Berman, M. G., & Beilock, S. L. (2019). Calculated avoidance: Math anxiety predicts math avoidance in effort-based decision-making. *Science advances*, *5*(11), eaay1062.
- Collins, A., & Brown, J. S. (1988). The computer as a tool for learning through reflection, In *Learning issues for intelligent tutoring systems*. Springer.
- Correll, S. J. (2001). Gender and the career choice process: The role of biased self-assessments. *American journal of Sociology*, *106*(6), 1691–1730.
- Crossley, S., & Kostyuk, V. (2017). Letting the genie out of the lamp: Using natural language processing tools to predict math performance, In *International conference on language, data and knowledge*. Springer.
- Crossley, S., Ocumpaugh, J., Labrum, M., Bradfield, F., Dascalu, M., & Baker, R. S. (2018). Modeling math identity and math success through sentiment analysis and linguistic features. *International Educational Data Mining Society*.
- Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779.

- deBoer, F. (2020). *The cult of smart: How our broken education system perpetuates social injustice*. Macmillan Audio.
- Do, A. M., Rupert, A. V., & Wolford, G. (2008). Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review*, *15*(1), 96–98.
- Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, *12*(656-704), 3.
- Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*, 508.
- Dunning, D., Heath, C., & Suls, J. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106.
- Dunning, D., & Helzer, E. G. (2014). Beyond the correlation coefficient in studies of self-assessment accuracy: Commentary on Zell & Krizan (2014). *Perspectives on Psychological Science*, *9*(2), 126–130.
- Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, *99*(3), 597.
- Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society*, *11*(2), 367–380.
- Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*(21), 5–17.
- Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.
- Ehrlinger, J., Mitchum, A. L., & Dweck, C. S. (2016). Understanding overconfidence: Theories of intelligence, preferential attention, and distorted self-assessment. *Journal of Experimental Social Psychology*, *63*, 94–100.
- Elliott, L., Braham, E. J., & Libertus, M. E. (2017). Understanding sources of individual variability in parents' number talk with young children. *Journal of Experimental Child Psychology*, *159*, 1–15.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Epstein, E. J. (1982). Have you ever tried to sell a diamond? *Atlantic Monthly*, *23*, 363.
- Fan, J. E. (2015). Drawing to learn: How producing graphical representations enhances scientific thinking. *Translational Issues in Psychological Science*, *1*(2), 170.
- Feld, J., Sauermann, J., & De Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, *68*, 18–24.
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114.

- Foley, A. E., Herts, J. B., Borgonovi, F., Guerriero, S., Levine, S. C., & Beilock, S. L. (2017). The math anxiety-performance link: A global phenomenon. *Current Directions in Psychological Science*, *26*(1), 52–58.
- Foushee, R., Jansen, R. A., & Srinivasan, M. (2017). What counts as math? relating conceptions of math with anxiety about math. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.
- Geary, D. C., Hoard, M. K., Nugent, L., & Scofield, J. E. (2020). In-class attention, spatial ability, and mathematics anxiety predict across-grade gains in adolescents' mathematics achievement. *Journal of Educational Psychology*.
- Gierzynski, A., & Eddy, K. (2013). *Harry potter and the millennials: Research methods and the politics of the muggle generation*. JHU Press.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. Chapman; Hall/CRC.
- Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*(5880), 1164–1165.
- Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles*, *66*(3–4), 153–156.
- Gunderson, E. A., Hamdan, N., Sorhagen, N. S., & D'Esterre, A. P. (2017). Who needs innate ability to succeed in math and literacy? Academic-domain-specific theories of intelligence about peers versus adults. *Developmental psychology*, *53*(6), 1188.
- Hamblin, J. (2020). *Clean: The new science of skin*. Riverhead Books.
- Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of personality and social psychology*, *111*(5), 745.
- Healy, P. J., & Moore, D. A. (2007). Bayesian overconfidence. *Available at SSRN 1001820*.
- Hembree, R. (1990a). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*, 33–46
- Hembree R (1990) The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education* 21:33–46.
- Hembree, R. (1990b). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 33–46.
- Hollingshead, A. D. (1976). Four factor index of social status: Working paper.
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581.
- Jansen, R. A., & Foushee, R. (2020). How we talk about math: Leveraging naturalistic datasets to define the discourse of math in contrast to other domains., In *Proceedings of the 13th international conference on educational data mining*.
- Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2017). Algebra is not like trivia: Evaluating self-assessment in an online math tutor., In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

- Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2018). Modeling the Dunning-Kruger effect: A rational account of inaccurate self-assessment., In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the dunning-kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, 1–8.
- Jansen, R. A., Rafferty, A. N., & Griffiths, T. (2020). A rational model of sequential self-assessment., In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.
- Jeon, K.-N., Moon, S. M., & French, B. (2011). Differential effects of divergent thinking, domain knowledge, and interest on creative performance in art and math. *Creativity Research Journal*, 23(1), 60–71.
- John, J. E., Nelson, P. A., Klenczar, B., & Robnett, R. D. (2020). Memories of math: Narrative predictors of math affect, math motivation, and future math plans. *Contemporary Educational Psychology*, 60, 101838.
- Jones, W. G. (2001). Applying psychology to the teaching of basic math: A case study. *Inquiry*, 6(2), 60–65.
- Kalaycioglu, D. B. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory and Practice*, 15(5), 1391–1401.
- Klebanov, B. B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective writing about the utility value of science as a tool for increasing STEM motivation and retention—can AI help scale up? *International Journal of Artificial Intelligence in Education*, 27(4), 791–818.
- Krajč, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, 29(5), 724–738.
- Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2), 180–188.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Kruger, J., & Dunning, D. (2002). Unskilled and unaware—but why? A reply to Krueger and Mueller. *Journal of Personality and Social Psychology*, 82(2), 189–192.
- Larkin, K., & Jorgensen, R. (2016). 'i hate maths: Why do we need to do maths?' using ipad video diaries to investigate attitudes and emotions towards mathematics in year 3 and year 6 students. *International Journal of Science and Mathematics Education*, 14(5), 925–944.
- Lauermann, F., Tsai, Y.-M., & Eccles, J. S. (2017). Math-related career aspirations and choices within eccles et al.'s expectancy–value theory of achievement-related behaviors. *Developmental psychology*, 53(8), 1540.

- Lee, J. (2012). College for all: Gaps between desirable and actual P–12 math achievement trajectories for college readiness. *Educational Researcher*, *41*(2), 43–55.
- Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.
- Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., & Gunderson, E. A. (2010). What counts in the development of young children’s number knowledge? *Developmental psychology*, *46*(5), 1309.
- Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013). Scalable sentiment classification for big data analysis using Naïve Bayes classifier, In *2013 IEEE International Conference on Big Data*, IEEE.
- Lockhart, P. (2009). *A mathematician’s lament: How school cheats us out of our most fascinating and imaginative art form*. Bellevue Literary Press.
- Lopez, G. (2017). Why incompetent people often think they’re actually the best. *Vox*. <https://www.vox.com/science-and-health/2017/11/18/16670576/dunning-kruger-effect-video>
- Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., & McGrath, L. M. (2017). Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in psychology*, *8*, 1314.
- Maier, M., & Abdel Rahman, R. (2018). Native language promotes access to visual consciousness. *Psychological Science*, *29*(11), 1757–1772.
- Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2015). Intergenerational effects of parents’ math anxiety on children’s math achievement and anxiety. *Psychological Science*, *26*(9), 1480–1488.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.
- Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales, In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.
- Mohammad, S., & Yang, T. W. (2011). Tracking sentiment in mail: How genders differ on emotional axes, In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis at ACL-HLT*.
- Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naïve Bayes model, In *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Berlin, Heidelberg.
- Nasir, N. S., & Hand, V. (2008). From the court to the classroom: Opportunities for engagement, learning, and identity in basketball and classroom mathematics. *The Journal of the Learning Sciences*, *17*(2), 143–179.
- Nelson, T. O., & Dunlosky, J. (1991). When people’s judgments of learning (jols) are extremely accurate at predicting subsequent recall: The “delayed-jol effect”. *Psychological Science*, *2*(4), 267–271.

- Neuman, Y., Cohen, Y., Assaf, D., & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, *56*(1), 19–25.
- Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2013). The single-item math anxiety scale (SIMA): An alternative way of measuring mathematical anxiety. *Journal of Psychoeducational Assessment*, *20*(10), 1–12
- Chipman S-F, Krantz D-H, Silver R (1992) Mathematics anxiety and science careers among able college women. *Psychol Sci* 3:292–295. 6.
- Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2014). The single-item math anxiety scale: An alternative way of measuring mathematical anxiety. *Journal of Psychoeducational Assessment*, *32*(4), 306–317.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.
- O’Dea, R. E., Lagisz, M., Jennions, M. D., & Nakagawa, S. (2018). Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature communications*, *9*(1), 1–8.
- OECD. (2013). *Pisa 2012 results in focus: What 15-year-olds know and what they can do with what they know*. Author Paris, France.
- OECD. (2016). Ten questions for mathematics teachers ... and how pisa can help answer them. *PISA*.
- Park, D., Ramirez, G., & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied*, *20*(2), 103–111.
- Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638.
- Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men*. Abrams Press.
- Plante, I., O’Keefe, P. A., Aronson, J., Fréchette-Simard, C., & Goulet, M. (2019). The interest gap: How gender stereotype endorsement about abilities predicts differences in academic interests. *Social Psychology of Education*, *22*(1), 227–245.
- Prezioso, M. (2020). Enchantment and understanding in philip pullman’s his dark materials: Advancing cognition through literature. *Children’s Literature in Education*, 1–12.
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.
- Purtill, C. (2018). This psychological quirk could explain why Trump’s least experienced lawyer feels so confident. *Quartz*. <https://work.qz.com/1240245/the-dunning-kruger-effect-what-trumps-legal-team-and-the-russia-probe-have-to-do-with-it/>
- Rafferty, A. N., Jansen, R. A., & Griffiths, T. L. (2020). Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive science*, *44*(10), e12900.
- Ramirez, G., Gunderson, E. A., Levine, S., & Beilock, S. (2013). Math anxiety, working memory, and math achievement in early elementary school. *Journal of Cognition and Development*, *14*, 187–202.

- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.
- Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.
- Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10.
- Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the “signal extraction” counterexplanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85–100.
- Schmidt, W., Houang, R., & Cogan, L. (2002). A coherent curriculum. *American Education*, 1–17.
- Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2014). The reciprocal relations between self-concept, motivation and achievement: Juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational psychology*, *34*(1), 49–72.
- Shepard, R. N., Kilpatrick, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive psychology*, *7*(1), 82–138.
- Steele, J. R., & Ambady, N. (2006).  
Math is hard!” The effect of gender priming on women’s attitudes. *Journal of Experimental Social Psychology*, *42*(4), 428–436.
- Stephens, S. L., Westerling, A. L., Hurteau, M. D., Peery, M. Z., Schultz, C. A., & Thompson, S. (2020). Fire and climate change: Conserving seasonally dry forests is still possible. *Frontiers in Ecology and the Environment*, *18*(6), 354–360.
- Stephens-Davidowitz, S., & Varian, H. (2014). A hands-on guide to google data.
- Ting, S. L., Ip, W. J., & Tsang, A. H. (2011). Is Naïve Bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, *5*(3), 37–46.
- Towers, J., Hall, J., Rapke, T., Martin, L. C., & Andrews, H. (2017). Autobiographical accounts of students’ experiences learning mathematics: A review. *Canadian Journal of Science, Mathematics and Technology Education*, *17*(3), 152–164.
- Tyszka, T., & Zielonka, P. (2002). Expert judgments: Financial analysts versus weather forecasters. *The Journal of Psychology and Financial Markets*, *3*(3), 152–160.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.
- Vreeman, R. C., & Carroll, A. E. (2007). Medical myths. *Bmj*, *335*(7633), 1288–1289.
- Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles, In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.



- Wang, M. T., & Degol, J. (2013). Motivational pathways to stem career choices: Using expectancy–value perspective to understand individual and gender differences in stem fields. *Developmental Review, 33*(4), 304–340.
- Wang, W.-C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2018). Knowledge supports memory retrieval through familiarity, not recollection. *Neuropsychologia, 113*, 14–21.
- Wolf, N. (2002). *The beauty myth*. Harper Perennial.
- Wolraich, M. L., Wilson, D. B., & White, J. W. (1995). The effect of sugar on behavior or cognition in children: A meta-analysis. *Jama, 274*(20), 1617–1621.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-johnson III*. Itasca, IL: Riverside.
- Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, 105*(13), 5012–5015.
- Yeager, D. S., & Dweck, C. . (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47*, 302–314.
- Yeager, D. S. (2012). Productive persistence: A practical theory of community college student success. *Vancouver, Canada*.
- Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science, 9*(2), 111–125.
- Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology, 107*(4), 954.

# A

## Appendix: Chapter 2

### A.1 Participants

A sample size of 1,000 was selected via a set of power analyses performed on the data from a survey posted by Quanta magazine.<sup>1</sup> The first power analysis was based on the result in pilot data showing that men had more positive attitudes about math than women ( $t(3028) = 8.77$ ,  $p < .001$ ,  $d = .36$ ). We set a significance level of  $\alpha = 0.05$  and power of 0.8. Our pilot sample was skewed (73% of respondents were male), so assuming that there would be twice as many men than women to complete the survey, it was determined that we would require a sample of 183 participants to detect this gender difference.

We looked at other meaningful differences in the pilot data, specifically differences in mean attitude rating based on reported time when attitudes were formed. While effect sizes may be somewhat larger in the proposed study on MTurk because the range of attitudes is likely to be larger, the pilot data provides a conservative estimate for the power analysis. Because we are doing multiple tests, we adjusted the  $\alpha$  value via a Bonferroni correction by dividing  $\alpha$  by four (giving  $\alpha = .0125$ ). We found that those who formed their attitudes in college had more positive attitudes than those who formed them in middle school ( $d = .31$ ), requiring a sample size of 350 and those who formed their attitudes in elementary school had more positive attitudes than those who formed them in middle school ( $d = .25$ ), requiring a sample size of 537. Differences between those who formed their attitudes in college vs. high school and elementary vs. high school yielded effect sizes of less than 0.2 ( $d = .17$  and  $d = .13$ , respectively), which is already a small effect.

Thus, achieving .8 power for the relationships with effect sizes greater than 0.2 requires data from 537 participants. In a small pilot ( $n = 29$ ), data from 10% of participants met exclusion criteria due to giving responses that were not relevant to the questions being asked. Assuming this rate of exclusion in the full sample, we determined that 600 participants would be needed in order to collect sufficient usable data for the statistical analyses we wish to perform. While there is not the equivalent of a power analysis for the computational text analyses, most studies using these methods make use of large datasets and more data tends

---

<sup>1</sup><https://www.quantamagazine.org/20161020-science-math-education-survey/>

to yield more robust results for these methods. Given this, we increased the sample to have enough data for computational text analyses and aimed for 1,000 responses in total.

## A.2 Hollingshead (SES)

To calculate the Hollingshead measure of SES, two research assistants independently coded 20% of the open-ended responses about each participant’s mother’s and father’s occupations. According to Hollingshead, occupations need to be categorized into 9 different levels for calculation. The ambiguity of some occupation categorization in Hollingshead impacted the initial interrater reliability ( $\kappa = 0.63$ ). After discussion, it was determined that a systematic disagreement was causing the poor reliability, so this was resolved and redone on the same set of responses, achieving much higher reliability ( $\kappa = 0.93$ ). The two reviewers resolved remaining disagreements and each coded half of the remaining responses.

We then combined parental occupation scores ( $MomOcc$  and  $DadOcc$ ) with responses to a multiple-choice question asking parent levels of education ( $MomEd$  and  $DadEd$ ) via Equation A.1 as detailed by hollingshead to compute a score for childhood SES:

$$SES = \frac{5(MomOcc + DadOcc) + 3(MomEd + DadEd)}{2}. \quad (A.1)$$

### A.3 ANOVA tables

	MOTHER'S EDUCATION					HOLLINGSHEAD						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	188.9	188.9	24.7	8.27e-7*	0.03	1	175.9	175.9	22.8	2.17e-6*	0.03
GRADE	5	333.1	66.6	8.7	4.69e-8*	0.05	5	322.7	64.5	8.4	1.04e-7*	0.05
RACE	4	49.3	12.3	1.6	0.17	0.007	4	57.8	14.4	1.9	0.11	0.01
SES	8	112.6	14.1	1.8	0.07	0.02	1	2.7	2.7	0.4	0.55	0.0004
GENDER:GRADE	5	97.7	19.5	2.6	0.03	0.01	5	86.0	17.2	2.2	0.05	0.01

Table A.1: ANOVA results predicting general attitudes about math using Mother's Education as the measure of SES on the left (n=879) and using Hollingshead on the right (n=799). Because of the Bonferroni correction, we note with an asterisk (\*) the instances when  $p < 0.0036$ .

## A.4 Mean ratings

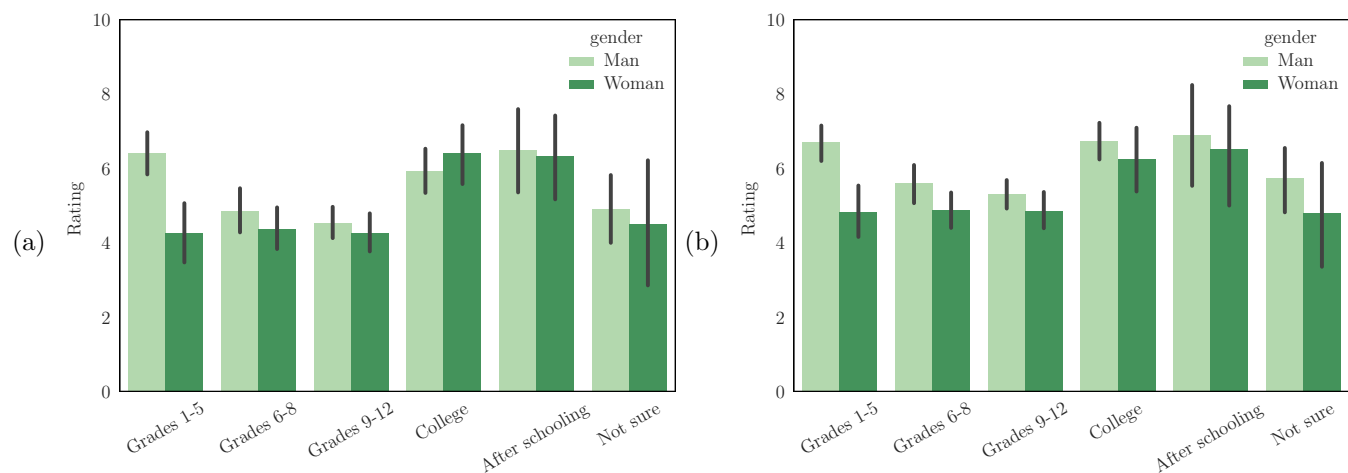


Figure A.1: Differences in ratings by gender and time when attitudes were formed (a) for attitudes about math in school and (b) at work.

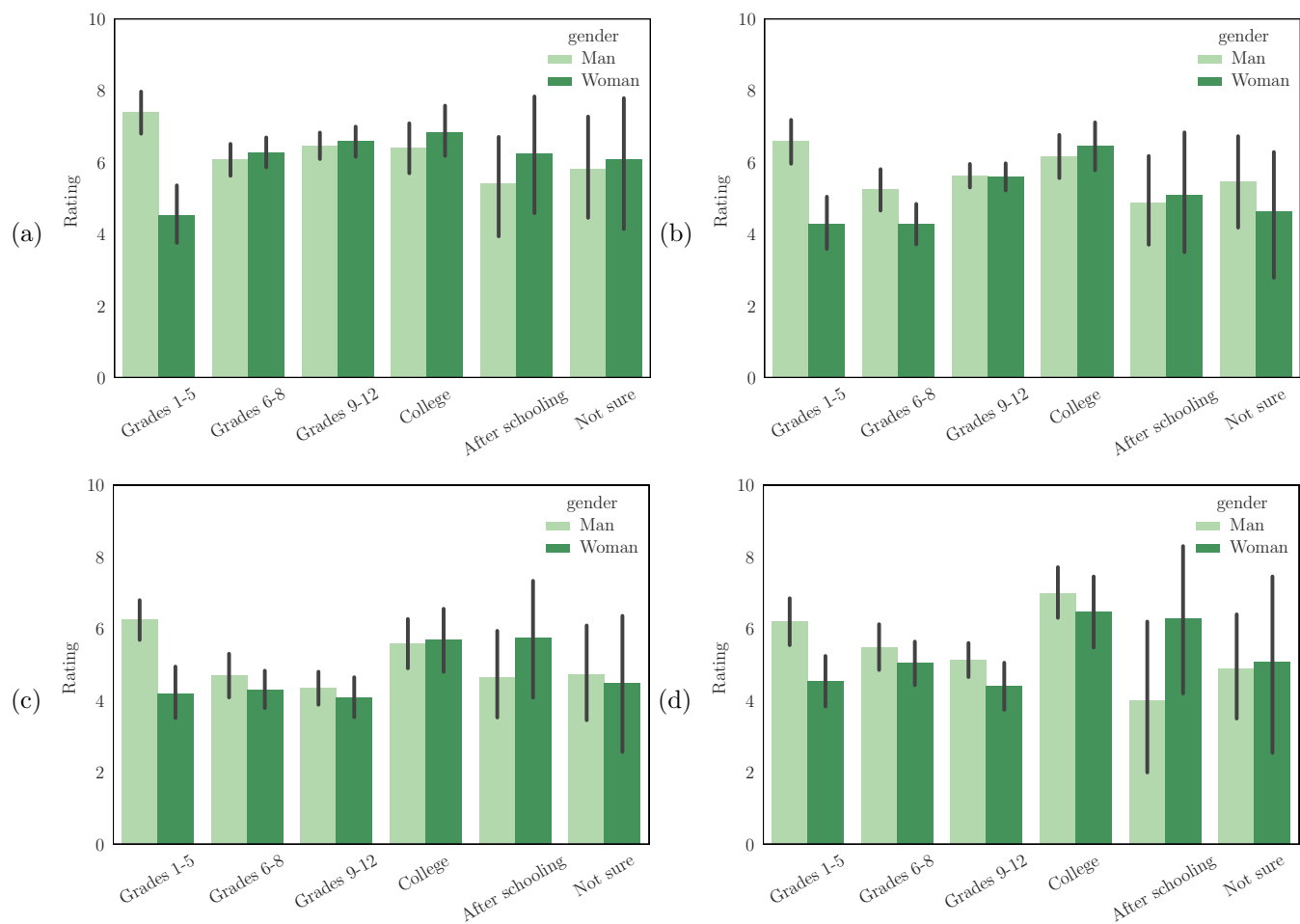


Figure A.2: Differences in ratings by gender and time when attitudes were formed (a) for attitudes in Grades 1–5, (b) in grades 6–8, (c) in grades 9–12, and (d) in college.

	MOTHER'S EDUCATION					HOLLINGSHEAD						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	142.1	142.1	14.9	0.0001*	0.02	1	125.4	125.4	13.1	0.0003*	0.02
GRADE	5	367.8	73.6	7.7	4.30e-07*	0.04	5	360.4	72.1	7.5	6.45e-7*	0.05
RACE	4	88.2	22.0	2.3	0.06	0.01	4	104.0	26.0	2.7	0.03	0.01
SES	8	149.9	18.7	2.0	0.05	0.02	1	0.2	0.2	0.02	0.89	0
GENDER:GRADE	5	137.2	27.4	2.9	0.01	0.02	5	117.4	23.5	2.5	0.03	0.2

Table A.2: ANOVA results predicting attitudes about math in school.



	MOTHER'S EDUCATION						HOLLINGSHEAD					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	189.8	189.8	24.7	8.21e-7*	0.03	1	175.4	175.4	22.7	2.29e-6*	0.03
GRADE	5	230.0	46.0	6.0	1.91e-5*	0.03	5	223.1	44.6	5.8	3.09e-5*	0.04
RACE	4	35.7	8.9	1.2	0.33	0.005	4	44.6	11.2	1.4	0.22	0.007
SES	8	85.2	10.6	1.4	0.20	0.01	1	2.6	2.6	0.3	0.56	0
GENDER:GRADE	5	62.3	12.5	1.6	0.15	0.009	5	47.3	9.5	1.2	0.30	0.008

Table A.3: ANOVA results predicting attitudes about math at work.

	MOTHER'S EDUCATION						HOLLINGSHEAD					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	52.5	52.536	6.3268	0.01	0.007	1	30.1	30.1	3.7	0.06	0.005
GRADE	5	52.2	10.4	1.3	0.28	0.007	5	49.0	9.8	1.2	0.31	0.008
RACE	4	17.2	4.3	0.5	0.72	0.002	4	25.5	6.4	0.8	0.54	0.004
SES	8	84.5	10.6	1.3	0.25	0.01	1	0.3	0.3	0.03	0.86	0
GENDER:GRADE	5	412.9	82.6	9.9	2.92e-9*	0.05	5	354.0	70.8	8.6	5.84e-8*	0.05

Table A.4: ANOVA results predicting attitudes about math in grades 1–5.

	MOTHER'S EDUCATION						HOLLINGSHEAD					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	163.8	163.8	18.7	1.71e-5*	0.02	1	156.2	156.2	17.8	2.69e-5*	0.02
GRADE	5	146.9	29.4	3.4	0.005	0.02	5	133.8	26.8	3.1	0.01	0.02
RACE	4	37.7	9.4	1.1	0.37	0.005	4	36.2	9.0	1.0	0.39	0.005
SES	8	132.1	16.5	1.9	0.06	0.02	1	4.7	4.7	0.5	0.47	0
GENDER:GRADE	5	178.0	35.6	4.1	0.001*	0.02	5	151.8	30.4	3.5	0.004	0.02

Table A.5: ANOVA results predicting attitudes about math in grades 6–8.

	MOTHER'S EDUCATION					HOLLINGSHEAD						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	100.7	100.7	9.8	0.002*	0.01	1	112.9	112.9	11.0	0.001*	0.01
GRADE	5	242.9	48.6	4.8	0.0003*	0.03	5	241.5	48.3	4.7	0.0003*	0.03
RACE	4	94.8	23.7	2.3	0.06	0.01	4	118.2	29.5	2.9	0.02*	0.01
SES	8	200.6	25.1	2.5	0.01	0.02	1	38.7	38.7	3.8	0.05	0.005
GENDER:GRADE	5	128.8	25.8	2.5	0.03	0.01	5	107.0	21.4	2.1	0.07	0.01

Table A.6: ANOVA results predicting attitudes about math in grades 9–12.

	MOTHER'S EDUCATION					HOLLINGSHEAD						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p</i>	$\eta_p^2$
GENDER	1	93.2	93.2	9.5	0.002*	0.01	1	83.4	83.4	8.4	0.004	0.01
GRADE	5	274.6	54.9	5.6	4.64e-5*	0.04	5	270.5	54.1	5.4	6.77e-5*	0.04
RACE	4	51.7	12.9	1.3	0.26	0.008	4	46.9	11.7	1.2	0.32	0.008
SES	8	140.0	17.5	1.8	0.08	0.02	1	82.0	82.0	8.2	0.004	0.01
GENDER:GRADE	5	87.5	17.5	1.8	0.11	0.01	5	79.2	15.8	1.6	0.16	0.01

Table A.7: ANOVA results predicting attitudes about math in college.

NEGATIVE			POSITIVE		
word	ratio	example	word	ratio	example
behind	11.4:1	“I was way <u>behind</u> and frustrated”	game	8.2:1	“what mainly shaped the general love for math were <u>games</u> ”
confus	11.0:1	“The way my professor and teachers of the past taught the subject is just unintuitive and just end up making me more <u>confused</u> ”	play	6.8:1	“ <u>played</u> a lot of <u>games</u> that also used a bit of math”
bare	8.7:1	“I barely did enough to get by”			
comprehend	8.7:1	“Math has always been difficult for me to <u>comprehend</u> ”			
fraction	7.5:1	“When I started learning <u>fractions</u> , I couldn’t understand what the teacher wanted”			
dislike	6.8:1	“when I got to high school and it got more complicated I began to really <u>dislike</u> math”			
negat	6.7:1	“I had some pretty <u>negative</u> experiences with math growing up”			
avoid	6.6:1	“I tried to <u>avoid</u> anything to do with math”			
lost	6.6:1	“I felt so unmotivated and <u>lost</u> ”			
difficulti	6.4:1	“I was usually able to overcome my <u>difficulties</u> , I often found my self unable to keep up”			
despis	6.4:1	“I began to despise math”			
along	6.4:1	“Neither my parents nor my teachers taught me the fundamental principles of mathematics, they just told me to follow <u>along</u> ”			
frustrat	5.7:1	“Some teachers were patient and others would get <u>frustrated</u> easily”			
remedi	5.2:1	“several semesters of <u>remedial</u> math”			
pre-algebra	5.2:1	“Math became really confusing for me when I started doing things like long division and <u>pre-algebra</u> ”			
terribl	5.2:1	“I had a <u>terrible</u> Geometry teacher”			
past	5.2:1	“I never got <u>past</u> the negative experiences”			

Table A.8: Classifier results for responses of positive vs. negative ratings (accuracy on the test set: 77%).

	PREDICTED: Negative	PREDICTED: Positive	F1-SCORE
ACTUAL: Negative	34	31	.64
ACTUAL: Positive	7	94	.83

Table A.9: Confusion matrix for Naïve Bayes Classifier for attitude ratings on the test set. It shows 34 negative responses and 94 positive responses were correctly predicted.

HIGH ANXIETY			LOW ANXIETY		
word	ratio	example	word	ratio	example
difficulti	6.9:1	“remember having <u>difficulty</u> with math”	video	6.3:1	“ <u>video</u> games helped by increasing my interest in computing”
confus	6.9:1	‘It doesn’t interest me, and it <u>confuses</u> me”	everyday	6.3:1	“My math teacher in grade school made math seem fun. She related variables to <u>everyday</u> objects”
given	6.9:1	“I had issues following the <u>given</u> steps”	excit	5.7:1	“the teacher was a bit more <u>excited</u> ”
lost	6.1:1	“I had an algebra teacher that wasn’t very good at teaching and then I felt that I got <u>lost</u> ”	engin	5.2:1	“My dad is an engineer and my mom is pretty handy with numbers”
begin	6.1:1	“I was at the <u>beginning</u> of the ‘new’ math revolution and it was awful”	applic	5.2:1	“learning real world <u>applications</u> ”
catch	6.1:1	“hard time <u>catching</u> up”	challeng	4.8:1	“My teachers took extra time with me to try and <u>challenge</u> me”
frustrat	5.8:1	“seeing complex problems usually <u>frustrates</u> me”	field	4.6:1	“career in the tech <u>field</u> ”
behind	5.7:1	“always felt like I was <u>behind</u> ”	statist	4.6:1	“I didn’t really like math much until some college classes in probability and <u>statistics</u> , which I loved”
order	5.3:1	“I think in <u>order</u> to be good in math, you have to have good memory”			
semest	4.5:1	“after a <u>semester</u> of struggling and feeling like I was stupid”			
horribl	4.5:1	“I am <u>horrible</u> at math”			
improv	4.5:1	“never did <u>improve</u> in it”			
write	4.5:1	“teachers expected me to <u>write</u> coherent sentences”			
pas	4.2:1	“I couldn’t really do better than just an average <u>passing</u> grade”			
materi	4.1:1	“5 years of going over the same <u>material</u> ”			
poor	4.1:1	“I had <u>poor</u> teachers”			

Table A.10: Classifier results for responses of high vs. low anxiety ratings (accuracy on the test set: 69%).

	PREDICTED: High Anxiety	PREDICTED: Low Anxiety	F1-SCORE
ACTUAL: High Anxiety	43	26	.64
ACTUAL: Low Anxiety	22	65	.73

Table A.11: Confusion matrix for Naïve Bayes Classifier for anxiety ratings on the test set. It shows 43 high anxiety responses and 65 low anxiety responses were correctly predicted.

WOMAN			MAN		
word	ratio	example	word	ratio	example
wrong	8.1:1	“you’re either right or <u>wrong</u> ”	video	6.1:1	“What shaped my views the most may have been <u>video games</u> ”
sure	5.4:1	“I’m not <u>sure</u> ”	reinforce	5.5:1	“I got positive <u>reinforcement</u> from my teachers”
instead	5.0:1	“there was only one correct answer <u>instead</u> of with Literature”	appli	5.5:1	“I could see the practical <u>applications</u> ”
extra	4.8:1	“had to work <u>extra</u> hard in and out of class to fully understand the material”	fascin	5.5:1	“the golden ratio for example is simply <u>fascinating</u> ”
okay	4.4:1	“I’m <u>okay</u> at basic math”	somewhat	4.9:1	“I could <u>tell</u> that while others struggled, I was able to do math <u>somewhat</u> well”
language	4.4:1	“always remain a different <u>language</u> for me”	program	4.9:1	“learning how to <u>program</u> ”
allow	4.2:1	“he did not <u>allow</u> me to come to the same answer in a different way”	materi	3.3:1	“my mother is an educator, and she always exposed me to a lot of school <u>material</u> and books while I as growing up.”
let	4.2:1	“the world is <u>letting</u> people down”			
drop	4.2:1	“I had to switch teachers a few times and <u>drop</u> courses”			
mark	3.5:1	“enough for me to get good <u>marks</u> ”			
father	3.5:1	“my <u>father</u> would help me with my homework”			
attent	3.5:1	“didn’t seem to give as much individual <u>attention</u> as we needed”			
brought forward	3.5:1	“the bad grade <u>brought</u> down my GPA”			
happi	3.5:1	“more complex as I move <u>forward</u> in school”			
happi	3.5:1	“The people around me like my parents, friends and teachers all were generally <u>happy</u> about my love for Math”			
instruct	3.5:1	“like a drill <u>instructor</u> in the army”			
smart	3.5:1	“he was also very <u>smart</u> and of course arrogant”			

Table A.12: Classifier results for responses of men vs. women (accuracy on the test set: 60%).

	PREDICTED: Man	PREDICTED: Woman	F1-SCORE
ACTUAL: Man	75	31	.67
ACTUAL: Woman	44	37	.50

Table A.13: Confusion matrix for Naïve Bayes Classifier for gender on the test set. It shows 37 responses by women and 75 responses by men were correctly predicted.



TOPIC #0		TOPIC #1		TOPIC #2		TOPIC #3		TOPIC #4	
prob.	word	prob.	word	prob.	word	prob.	word	prob.	word
.061	enjoy	.022	manage	.020	anymore	.034	understanding	.113	math
.046	calculus	.014	mei	.019	skill	.022	moment	.035	class
.036	fail	.013	anxiety	.018	exam	.019	instructor	.026	teacher
.025	interest	.013	lesson	.014	old	.018	relationship	.024	feel
.019	care	.012	awkward	.014	everyone_else	.015	live	.023	school
.018	course	.011	explain	.014	real_world	.014	result	.020	grade
.017	elementary	.009	foundation	.013	explain	.011	method	.017	really
.015	early	.009	embarrassed	.012	shape	.010	frustrated	.016	time
.013	level	.008	loud	.011	avoid	.009	topic	.015	learn
.011	effort	.008	proof	.011	formula	.009	multiplication	.013	remember

Table A.14: Topic modeling output with 5 topics. This table shows the 10 words (or word roots) with highest probability under each topic and each word's associated probability under that topic.

TOPIC #0		TOPIC #1		TOPIC #2		TOPIC #3		TOPIC #4	
prob.	word	prob.	word	prob.	word	prob.	word	prob.	word
.062	course	.156	math	.126	math	.054	first	.115	college
.045	learn	.104	class	.042	feel	.044	hate	.088	class
.038	algebra	.027	school	.030	well	.044	professor	.049	calculus
.034	work	.023	major	.030	remember	.038	fail	.040	high
.030	still	.022	time	.029	really	.028	life	.032	test
.029	math	.022	year	.027	much	.022	tutor	.030	math
.029	thing	.019	student	.024	time	.020	different	.022	feel
.026	start	.017	love	.024	help	.020	drop	.022	pass
.022	level	.015	end	.023	hard	.019	mathematic	.019	study
.021	really	.013	semester	.021	teacher	.016	already	.015	feeling

Table A.15: Topic modeling output over all time-points with 5 topics. This table shows the 10 words (or word roots) with highest probability under each topic and each word's associated probability under that topic.