Writing Sample: ***Analyzing Associations Between Gender-Based Stereotypes and Educational Priorities***

Note: This writing sample is an excerpt from an unfinished research project and an expanded version of my NSF GRFP proposal. It includes a review of relevant literature, details of the proposed experimental design and survey, details of proposed techniques for model estimation, justification for the proposed sample size based on the statistical power of results in related research, and details of the proposed pilot version of the survey. Data has not been collected or analyzed yet. Some parts have been omitted for brevity.

Abstract: Gender gaps in educational performance can lead to stereotypes about certain genders being predisposed to perform better in particular subjects. This can create self-fulfilling prophecies, exacerbating those gender gaps. By administering a survey to elementary and middle school-age children and their parents, I will elicit *perceived gender stereotypes* of different subjects and preferences over which subjects the children and parents would like the children's performance to improve in. With educational priorities measured as *marginal utilities* over subjects, regressing those marginal utilities on interactions between perceived gender stereotypes and student gender will contribute to evidence on whether gender gaps in educational priorities are linked with stereotyping. This research will expand on existing work on educational priorities by allowing for greater granularity in measuring educational choices and by examining parental stereotyping and student self-stereotyping, two underexplored areas. Further work will be needed to determine causal relationships and examine potential policy interventions.

Table of Contents

Introduction

In historically male-dominated fields like STEM and business, there is a severe disparity in professional outcomes between men and women. In the 1960s, fewer than 25% of life science, physical science, and computer science jobs were done by women, including fewer than 2% of engineering jobs. [Sassler et al. (2016)] While those numbers have gone up in recent decades, with women comprising a majority of life sciences graduates since the 1990s, considerable wage gaps continue to exist. [Sassler et al. (2016)] Similar patterns are seen in business, where there is both a wage gap and a disparity in the rate at which managers of different sexes are promoted. [Bertrand et al. (2010)] [Elmins et al. (2016)] This problem extends to academia, with one paper finding that serving as a co-author on an academic paper increases tenure probability less for women economists than for male economists. [Sarsons et al. (2021)]

In educational settings, gender gaps in performance can lead to stereotypes about certain genders being predisposed to perform better in particular subjects. This can create self-fulfilling prophecies in educational performance, contributing to gender gaps in education. One study found that women and men with strong math backgrounds scored similarly well on an easy math exam, but men scored significantly higher on a more difficult version of the test. When participants were explicitly told that women had underperformed men on the difficult exam, the gap widened relative to the untreated group. However, when participants were told that there was no gender gap, the gap narrowed significantly. [Spencer et al. (1999)] An RCT in east China found similar results. [Details omitted] [Huang et al. (2023)]

One type of gender-based stereotyping that may contribute to disparities in performance is *self-stereotyping*, the phenomenon under which individuals deem themselves more or less likely to succeed in a particular field based on whether their gender matches the archetypal gender of that field. Experimental evidence suggests that self-stereotyping is associated with people's decisions about whether to take initiative in a group setting. In one study, paired participants were asked to answer trivia questions of six different categories. The interaction between whether a question's category was perceived to be "male-typed" or "female-typed" and the gender of each participant contributed to whether each participant believed themselves or their partner better equipped to answer each question, and ultimately influenced

willingness to contribute one's answer. [Coffman (2014)] Gender disparities in willingness to contribute led to disparities in performance in the game, as women were significantly more likely to fail to contribute a correct answer for questions from male-typed categories, while men and women failed to contribute correct answers to female-typed questions at roughly the same rate. [Coffman (2014)]

Besides performance and willingness to take initiative, negative self-stereotypes are also associated with a lower likelihood of participating in an academic or professional field, which can contribute to representation and achievement gaps. One survey of middle school-age students in Italy found that girls were less likely than boys to express interest in STEM fields, and more likely to cite their gender as a barrier to achieving their career goals. [Carlana and Fort (2022)] Stereotyping by others is also associated with educational choices. For instance, using class fixed effects to account for unobservable educational variables, a greater degree of gender bias among middle school math teachers in Italy was found to lead to worse mathematical performance by their female students, while increasing the probability that those students would select less academically rigorous high schools. [Carlana 2019]

The role of gender-based stereotyping by parents in influencing educational choices is underexplored in existing literature, but earlier work has shown associations between other educational attitudes of parents and children. For instance, higher levels of math anxiety in parents are associated with more negative attitudes toward math and, in some cases, worse performance in math classes. [Casad et al. (2015)] Gender also appears to play a role: higher math anxiety in fathers is associated with greater levels of devaluing the importance of math courses in daughters, but the relationship is not as strong for mothers and sons. [Casad et al. (2015)] Moreover, there is evidence that parents may harbor negative stereotypes about the mathematical ability of their daughters relative to their sons. [Furnham et al. (2002)]

Knowing that gender-based stereotypes are associated with people prioritizing certain academic and professional fields over others, which in turn can contribute to disparate career outcomes, I aim to investigate the relationship between gender stereotypes and self-stereotypes about performance in various subjects and children's decisions about whether to prioritize those subjects in school. This will build on existing literature, as most existing work on student self-stereotyping focuses on expressed interest in

certain fields [Carlana and Fort (2022)] or college major choice [De Gioannis 2022] rather than children's educational priorities in primary and secondary school. I also intend to investigate the relationship between parental stereotyping and parents' decisions on whether to prioritize certain subjects for their children.

I will field a survey to children and their parents, eliciting preferences over which subjects they would like either their own or their children's performance to improve in, respectively. Following the methodology of Joensen et al. (2023), I will administer the survey at schools in the Chicago metro area. One study on Chicago Public Schools during the 2016-17 and 2017-18 school years found gender disparities in performance, making this an appropriate setting to study gender differences in educational choices: ninth-grade girls earned higher mathematics grades than ninth-grade boys, a difference that was not driven by disparities in attendance, out-of-school suspensions, or prior experience. [Easton and Diaz (2023)] Moreover, Chicago exhibits considerable income inequality along geographic, racial, and educational lines, while its south suburbs have undergone a significant rise in poverty over the past two decades. [Melstrom and Redding (2020), Joensen et al. (2023), Kneebone and Berube (2013)]. Earlier research has shown family income to be a contributing factor to children's educational choices, further making the greater Chicago area a germane setting for this experiment. [Tamm (2008)]

Educational priorities will be measured in terms of marginal utilities over subjects. Next, by asking students and parents whether girls or boys tend to perform better in each subject, I will determine that subject's *perceived gender stereotypes*. Finally, following the specification in Table IV of Coffman (2014), but with educational priorities as the dependent variable, I will regress marginal utilities on interactions between perceived gender stereotypes of subjects and children's gender. This will contribute to evidence on whether there is a gender gap in which subjects children and their parents prioritize, and whether that is associated with stereotypes and self-stereotypes. Since I have not yet partnered with a school to field the survey to students, I will first field a pilot version of the survey to parents only. I will refer to the two versions of the survey as the *full version* and the *pilot version*.

Section 1 details the different questions asked on the survey and gives a brief overview of the sample. Section 2 outlines the headline regressions and their interpretations. Section 3 details the methodology for calculating marginal utilities. Section 4 examines the rationale behind the intended sample size and anticipates the statistical power of the results based on existing work on similar topics. Sections 5 and 6 are specific to the pilot version of the experiment. Section 5 explains the rationale for choosing Prolific as the intended online survey fielding platform for the pilot. Then, because I will not have access to student grades for the pilot, I will instead examine children's performance in each subject using *parents' subjective ratings of their children's performance*. Section 6 details a model for adjusting for biases due to *scale-use heterogeneity* in those subjective ratings, closely following the procedure established by Benjamin et al. (2023). [Section 6 is mostly omitted]

Section 1: Data and Survey

I plan to collect data on elementary and middle school-age (K-8) students' grades in five academic subjects, standardized within schools: *mathematics*, *English*, *art and music*, *social studies/history*, and *science*. In the pilot, I will instead use *subjective ratings of academic performance* in these subjects. I am developing a 10-minute-long online survey about educational performance and preferences. I will first field a pilot version of this survey to parents only using Prolific, asking each parent to answer the survey about one child in the household. After the pilot is fielded, the next step will be to partner with a school and administer the survey to students and their parents. Once I find a school to partner with, I will ask teachers to administer the survey to students whose parents consent in the classroom using school tablets or computers, as well as to those students' parents over email. The survey will have six main question types:

1. **Grade Verification Questions (full version only)**: students and their parents will be asked to verify the student's grades in each subject *s* from the most recent academic year. Each student's grade-point average in each subject over that year will represent their *performance level* in that subject. Parents will also be asked to verify the grading scale used: for instance, 0-100, 1-4, 1-5, or A-F.

In regressions, grades will be *standardized within schools* by subtracting the mean and dividing by the standard deviation to account for heterogeneities in school curricula: for instance, if some schools require students to take multiple math or science classes in a single semester while others do not.

2. **Rating Questions (pilot version only)**: parents will rate the performance of the child they have chosen to focus their responses on in each of the five subjects *s* relative to their peers. Each question is worded as follows:

   "Compared to other students your child's age, all across the United States, how would you rate your child's performance in [subject s]?"

   Ratings are on a sliding scale of 0-100, with 0 denoted *"Worst performance possible"* and 100 denoted *"Best performance possible."*

   These labels deliberately use extreme wording to reduce the risk of *bottom and top-coding*, which refers to when respondents' ratings disproportionately bunch near the bottom or top of the rating scale. [Benjamin et al. (2023)] Bottom and top-coding are potential problems, as they can bias the sample distribution of ratings upwards or downwards, respectively, since respondents may have chosen ratings outside the bounds of the scale were those options available. [Oganian et al. (2020)]

3. **Tradeoff questions**: to calculate parents' and students' *marginal utilities* of improvements in different subjects, students and their parents are asked to choose between two options: (1) a *magnitude-x improvement* in the student's performance in *subject a*, or (2) a *magnitude-y improvement* in the student's performance in *subject b*. Details on the calculation of marginal utilities are given in Section 3.

   On the pilot survey, magnitudes *x* and *y* are in units of points on the 0-100 rating scale: for instance, an 8-point improvement in mathematics versus a 2-point improvement in science. Following the design of Benjamin et al. (2014-B), Benjamin et al. (2023), and Kundu (2023), I limit the magnitudes of improvements to between 1 and 8 points. On the full survey, magnitudes *x* and *y* are in units of points corresponding to the student's GPA in a subject *s*. For consistency between different grading scales at different schools, I will use the equivalent of improvements between 1 and 8 points on a 0-

100 scale but translated to other scales. For example, if a school reports GPAs between 0 and 4.00, I will ask tradeoff questions with improvements of magnitudes between 0.04 and 0.32 points, in increments of 0.04.

An additional challenge with the full survey is that I standardize grades within schools when I run the headline regressions and calculate marginal utilities, but I ask respondents to choose between *unstandardized* grade improvements relative to an *unstandardized* GPA. For instance, suppose a student chooses an 8-point improvement in mathematics, which after standardization corresponds to a 2-standard deviation improvement relative to the rest of the school. The student would have been asked this tradeoff question in terms of an 8-point improvement, but the marginal utility calculations for that student and corresponding regressions would be based on the 2-standard deviation improvement following standardization.

To reconcile this, I assume that respondents are fully aware of this correspondence: so, a student faced with the option of an 8-point improvement in mathematics subconsciously understands that this corresponds to a 2-standard deviation improvement after standardization. However, this is an extremely strong assumption, as students and parents would have no realistic way of knowing such specificities of the distribution of students' grades. Moreover, this would contradict results from experiments on cognitive uncertainty and complexity, which show that respondents make errors on simple mathematical reasoning questions, such as "\$50 shrunk by a factor of $\delta = 0.96$ four times," so it would be unrealistic for a respondent to have this knowledge. [Enke et al. (2023)]

I will ask tradeoff questions for all 10 possible combinations of subjects to determine the degree to which children and parents prioritize one subject over others. The order of subjects shown is randomized. I will also ask five of the 10 combinations for a second time but in reverse order to test for a *left-right bias* in the respondent's choices, which existing literature has found to exist on other types of response scales. [Nicholls et al. 2006, Royer (2017)]

4. **Demographic questions**: [Details on demographic questions omitted] On the *full survey*, I will only administer *demographic questions* to parents, not children.

5. **Gender stereotyping questions**: these are intended to gauge students' and parents' *perceived gender stereotypes* – whether they believe each subject *s* is "male-typed" or "female-typed," in the style of Coffman (2014). Depending on whether children or their parents are being asked, these questions will be worded as follows: "Over the course of [your/your child's] education, have you found that girls or boys perform better in [subject *s*]?" with options for "girls" (-1), "boys" (1), or "about the same" (0).
6. **Calibration questions**: these are used to adjust for differences in how respondents use the 0-100 response scale, or *scale-use heterogeneities,* as defined in Section 6. [Most of Section 6 is omitted] Respondents will be asked to rate a series of stimuli with assumed objective "low," "medium," and "high" values on the 0-100 scale, where 0 represents the *"Lowest level possible"* and 100 represents the *"Highest level possible."* [Benjamin et al. (2023)] Calibration questions are based on a *visual* stimulus, such as "How dark is this circle?" or "How curved is this line?", following Benjamin et al. (2023).

Section 2: Headline Regressions

Estimates of each respondent's marginal utilities for each subject yield a representation of that respondent's educational priorities. In the intended full survey sample, in the survey administered to students, this represents the degree to which each student prioritizes *an improvement in their performance* in each subject, relative to the other subjects. In the survey administered to parents, this represents the degree to which each parent prioritizes *that their child's performance improves* in each subject, also relative to the other subjects. By regressing these marginal utility values on interactions between the *child's gender* and the *perceived gender stereotype* of each subject, I can observe whether there is an association between educational priorities and gender-based stereotypes.

In the following notation, I will index individual children as *i*. Then, to distinguish between the regression specifications for parents' and children's responses, I will index any data from the children's survey as i_c (c for "child's data") and any data from the survey administered to a parent of child *i* as i_p (p

for "parent's data", with i_p meaning "parent of child *i*"). If both parents of child *i* respond to the survey, I will still index each as i_p for simplicity, but that indexing can refer to either parent.

The headline regression using *parents' marginal utilities* emulates the specification of "Pooled" column in Table IV of Coffman (2014), but with marginal utilities as the left-hand side variable:

$$(2.1) \quad m_{i_p s} = \alpha_{i_p s} + \beta_1 G_{i_c} + \beta_2 g_{i_p s} + \beta_3 G_{i_c} * g_{i_p s} + \delta_1 w_{is} + \delta_2 x_i + \eta_{i_p s}$$

The dependent variable in this Ordinary Least Squares regression is *parents' marginal utilities* of an improvement by their children in subject *s*, representing the relative priority parents place on their children doing well in *s*. β_1 represents the difference in the relative priority placed by parents on subject *s* by children's gender – so, the difference in the relative priority placed on *s* between parents of boys and parents of girls. G_{i_c} is an indicator for the child's gender. $g_{i_p s}$ is parent i_p's response to the *gender stereotyping question* about subject *s*, as defined in Section 1. Thus, β_2 represents the difference in the degree of priority placed by parents on their child's performance in subject *s* by whether they believe girls or boys tend to do better in *s*. β_3 captures the interaction between the child's gender and their parents' beliefs about whether boys or girls do better in *s*. This is hypothesized to represent *gender-based stereotyping*: the differences in the extent to which parents prioritize subject *s* for their children when their child's gender matches the perceived gender of *s* versus when it does not.

w_{is} represents child *i*'s performance in subject *s*, as defined earlier. In the full survey, w_{is} will be in terms of the child's grade in *s*, standardized within schools. In the pilot, w_{is} will be the parent's scale-use corrected rating of their child's performance in *s* on a 0-100 scale. δ_1 is meant to capture differences in marginal utilities by level of performance, thus adjusting for diminishing or increasing marginal utility. x_i represents the vector of demographics of the child, their parent, and the rest of their family (excluding child's gender, G_{i_c}), while $\eta_{i_p s}$ is an error term. Units of observation are individual parents from whom data has been collected and for whom marginal utilities have been calculated. On the pilot survey, marginal utilities are in terms of utils per unit on the 0-100 scale. On the full survey, they are in terms of utils per GPA point *standardized within school*. In other words, marginal utilities are in terms of the

number of utils gained through a *one-standard-deviation* improvement in subject *s* relative to the grade distribution for each school.

The regression of *children's marginal utilities* is similar, but appears as follows:

$$(2.2) \quad m_{i_c s} = \alpha_{i_c s} + \beta_1 G_{i_c} + \beta_2 g_{i_c s} + \beta_3 G_{i_c} * g_{i_c s} + \delta_1 w_{is} + \delta_2 x_i + \eta_{i_c s}$$

β_1 captures the degree to which a child's prioritization of success in a subject *s*, $m_{i_c s}$, differs by their gender. Meanwhile, β_3 is hypothesized to capture *gender-based self-stereotyping*: whether the priority a child places on subject *s* varies by whether their own gender, G_{i_c}, matches the perceived gender of *s*. Units of observation are individual students, and the unit of measure of the dependent variable is the same as in the parents' survey.

Section 3: Model of Marginal Utilities

To determine the degree to which students and their parents emphasize certain subjects over others, I will adopt the theory developed in Benjamin et al. (2014-A). This paper proposes an approach for estimating preferences over market and non-market goods, which include intangible items like emotions, levels of satisfaction with various aspects of one's life, and, in the case of this study, academic performance [Benjamin et al. (2014-A)]. In the full implementation of this project, let w_{is} refer to student *i*'s standardized grade in subject *s*. On the pilot, let w_{is} denote student *i*'s performance in subject *s*, as rated by their parent and adjusted for *scale-use heterogeneities.* While both of a student's parents may take the survey, for simplicity, let *i* refer to either parent of student *i*. Moreover, in all notation going forward, let respondent *i* refer to either the student *i* answering the survey about themselves, or either one of student *i*'s parents answering questions about the student.

Then, for each student *i*, let $u_i(w_{i1}, \dots, w_{iS})$ be a utility function over student *i*'s performance in each subject *s*, where in this study, $S = 5$. In the full survey implementation, this will denote either student *i*'s preferences over grades in the five subjects, or a parent's preferences over their child's grades in the five subjects. As detailed in Section 1, I make the strong assumption that students and parents have

a full understanding of the correspondence between raw grades, which they see on the survey, and grades standardized within schools, which I use to calculate marginal utilities. Thus, I assume that the preferences represented by u_i are the same regardless of whether w_{is} represents raw or standardized grades. In the pilot, this utility function represents the preferences of a parent of student *i* over their child's performance in the five subjects, as rated by that parent. Then, the *marginal utility* of u_i with respect to a change in the child's performance in subject *s* is given by $M_{is} = \frac{\partial u_i}{\partial w_{is}}$. By the definition of the partial derivative, at very small changes in performance Δw_{is}, I can approximate the marginal utility as:

$$(3.1) \quad M_{is} = \frac{\partial u_i}{\partial w_{is}} \approx \frac{u_i(w_{is}+\Delta w_{is}, \boldsymbol{w_{i,-s}}) - u_i(w_{i1},\dots,w_{iS})}{\Delta w_{is}}$$

Where $\boldsymbol{w_{i,-s}}$ is the vector of child *i*'s performances in all subjects except *s*.

Thus, I have that at small Δw_{is}, the *change in utility brought on by* Δw_{is}, $u_i(w_{is} + \Delta w_{is}, \boldsymbol{w_{i,-s}}) - u_i(w_{i1}, \dots, w_{iS})$, is approximated by $M_{is}\Delta w_{is}$.

Consider a *tradeoff question*, as defined in Section 1. The respondent chooses between two options, displayed on the *left* and *right*, respectively: an increase in child *i*'s performance in subject *l* with magnitude Δw_{il}, and an increase in child *i*'s performance in subject *r* with magnitude Δw_{ir}. So, respondent *i* will choose the *left* option if and only if $u_i(w_{il} + \Delta w_{il}, \boldsymbol{w_{i,-l}}) - u_i(w_{i1}, \dots, w_{iS}) > u_i(w_{ir} + \Delta w_{ir}, \boldsymbol{w_{i,-r}}) - u_i(w_{i1}, \dots, w_{iS})$, or, equivalently, if and only if $u_i(w_{il} + \Delta w_{il}, \boldsymbol{w_{i,-l}}) > u_i(w_{ir} + \Delta w_{ir}, \boldsymbol{w_{i,-r}})$. Under the approximation for small Δw_{il} and Δw_{ir}, we equivalently have that the respondent *i* chooses the *left* option if and only if $M_{il}\Delta w_{il} > M_{ir}\Delta w_{ir}$. Taking logs and rearranging, respondent *i* chooses the *left* option if and only if: $m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) > 0$, where $m_{is} = log(M_{is})$.

I assume that choices on tradeoff questions are random under a standard normal error distribution. Moreover, I include a parameter λ_i to account for respondent *i*'s bias towards or away from left-side responses in tradeoff questions, as mentioned in Section 1. [Nicholls et al. (2006), Royer (2017)] So, respondent *i* chooses the *left* option if and only if: $m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i + \epsilon_{ilr} > 0$, where $\epsilon_{ilr} \sim N(0,1)$. Thus, I have the Probit model:

$$(3.2) \quad P(r_{ilr} = 1 | m_{il}, m_{ir}, \Delta w_{il}, \Delta w_{ir}, \lambda_i) = \Phi\left(m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i\right),$$

where r_{ilr} is an indicator for choosing the *left* option on the tradeoff described above.

The *log-likelihood* of observing tradeoff responses r_{ilr} given parameters $m_{il}, m_{ir}, \lambda_i$ for all $i \in \{1, \dots, I\}$ and $l, r \in \{1, \dots, S\}$ is given by:

$$(3.3) \quad l(\{r_{ilr}\}; \{m_{is}\}, \{\lambda_i\}) = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{r \in \{1,\dots,S\} \setminus \{l\}} r_{ilr} * log\left(\Phi\left(m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i\right)\right) +$$

$$\sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{r \in \{1,\dots,S\} \setminus \{l\}} (1 - r_{ilr}) * log(1 - \Phi\left(m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i\right))$$

I could try to estimate m_{is} for all $i \in \{1, \dots, I\}$, $s \in \{1, \dots, S\}$ as fixed effects by finding the values that maximize log-likelihood. However, this is problematic given the structure of the data. Since I am asking about only five subjects, it is realistic that some people may prioritize one subject *s* over the others by a considerable margin. In that case, they would choose that subject in every tradeoff question. This would mean that this respondent's likelihood-maximizing log marginal utility of subject *s*, m_{is}, would be infinite, so the parameter would not be identified. Instead, to estimate heterogeneities across *people* in their educational priorities, I can allow marginal utilities to vary by treating them as *random effects* under a hierarchical model, following Kundu (2023). Assume that the parameters of interest – m_{is} and λ_i – are normally distributed around their respective population means μ_s and λ, with additional hyperparameters σ_m and σ_λ representing variances:

$$(3.4) \quad m_{is} \sim N(\mu_s, \sigma_m)$$

$$(3.5) \quad \lambda_i \sim N(\lambda, \sigma_\lambda)$$

I also now assume that response error, ϵ_{ilr}, is normally distributed with its variance varying across *individuals* but not specific tradeoffs:

$$(3.6) \quad \epsilon_{ilr} \sim N\left(0, \sigma_{\epsilon_i}\right).$$

To ensure that response error variance is positive, we assume that σ_{ϵ_i} follows a *lognormal* distribution: $log\left(\sigma_{\epsilon_i}\right) \sim N(\mu_\epsilon, \sigma_\epsilon)$. I assume that any systematic bias in response error is captured by λ_i, so ϵ_{ilr} has mean zero. Next, all the hyperparameters are given uninformative *prior distributions*, following

Betancourt and Girolami (2013) and Kundu (2023). The means are given normal priors with wide variances, and the variance hyperparameters are distributed according to a *half-Cauchy distribution,* or Cauchy distribution truncated at zero, to ensure positive values:

$$(3.7) \quad \mu_s, \lambda, \mu_\epsilon \sim N(0, 10)$$

$$(3.8) \quad \sigma_m, \sigma_\lambda, \sigma_\epsilon \sim HalfCauchy(0, 2)$$

Existing work has shown that at large sample sizes, the likelihood function will dominate the priors when estimating the posterior distribution. [Smid and Winter (2020)] So, given the intended full sample of $N = 2{,}200$ to $N = 3{,}700$ as determined in Section 4, it is reasonable to set uninformative priors. Then, the log-likelihood function to maximize, given the various hyperparameters, is:

$$(3.9) \quad l\left(\{r_{ilr}\};\ \{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\}, \{\mu_s\}, \lambda, \mu_\epsilon, \sigma_m, \sigma_\lambda, \sigma_\epsilon\right) = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{r \in \{1,\ldots,S\} \setminus \{l\}} r_{ilr} \ *$$

$$log\left(\Phi\left(\frac{m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i}{\sigma_{\epsilon_i}}\right)\right) + \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{r \in \{1,\ldots,S\} \setminus \{l\}} (1 - r_{ilr}) \ * \ log\left(1 - \Phi\left(\frac{m_{il} - m_{ir} + log\left(\frac{\Delta w_{il}}{\Delta w_{ir}}\right) + \lambda_i}{\sigma_{\epsilon_i}}\right)\right)$$

I can try to directly maximize the log-likelihood function over all the parameters and hyperparameters. However, with so many parameters to estimate, this is computationally infeasible. [Kundu (2023)] To work around this, I can estimate log marginal utilities and other parameters using *Hamiltonian Markov Chain Monte Carlo (HMCMC)*. Instead of estimating the parameters directly, I can study the *posterior distribution* of the parameters and hyperparameters given tradeoff data $\{r_{ilr}\}$. I will use Stan to sample draws from the posterior distribution to get point estimates and standard errors for each of the parameters and hyperparameters. The joint posterior probability density is defined as the following, with the vector of hyperparameters $\{\mu_s\}, \lambda, \mu_\epsilon, \sigma_\mu, \sigma_\lambda, \sigma_\epsilon$ written as ζ for ease of notation:

$$(3.10) \quad P(\{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\}, \zeta | \{r_{ilr}\}) = \frac{P(\{r_{ilr}\} | \{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\}, \zeta) * P(\{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\} | \zeta) * P(\zeta)}{P(\{r_{ilr}\})}$$

Although I cannot directly calculate the posterior distribution, I can still sample from it, as I have defined every component of the posterior density. By definition, $P(\{r_{ilr}\} | \{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\}, \zeta)$ is simply the likelihood function, $L\left(\{r_{ilr}\};\ \{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\}, \{\mu_s\}, \lambda, \mu_\epsilon, \sigma_\mu, \sigma_\lambda, \sigma_\epsilon\right)$. $P(\{m_{is}\}, \{\lambda_i\}, \{\sigma_{\epsilon_i}\} | \zeta)$ is the joint density of the parameters – log marginal utilities, left-right bias parameters, and response error variances

– conditional on the hyperparameters. This is simply the joint distribution combining $m_{is} \sim N(\mu_s, \sigma_m)$, $\lambda_i \sim N(\lambda, \sigma_\lambda)$ and $\sigma_{\epsilon_i} \sim LogNormal(\mu_\epsilon, \sigma_\epsilon)$. Finally, $P(\zeta)$ is the joint distribution of the hyperparameters – the joint distribution combining the uninformative priors.

I will use Stan to perform the HMCMC procedure, sampling draws of parameter and hyperparameter values from the joint posterior distribution. The goal is to sample from the region of the multidimensional distribution where probability is most densely concentrated, which Betancourt (2017) calls the "typical set." This is not the region directly at the mode of our posterior, as that region is far too narrow to give a representative distribution, and it is not the region on the fringes, as that is too far from the mode to give an accurate representation of the parameters' point estimates. [Betancourt (2017)] Rather, it is a region between the two, and the purpose of the procedure is to have the draws "orbit" around that region. This is done by using the *Hamiltonian equations* of mechanics to define subsequent draws using notions of kinetic and potential energy, thus simulating putting the samples into a physical orbit. [Betancourt (2017)] Sampling 1,000 draws would then give estimates of the distributions of all the parameters and hyperparameters.

Section 4: Sample and Statistical Power

On the full survey, I plan to collect a sample large enough for sufficient statistical power, with a baseline set by related literature. In calculating the intended sample size, I make the strong assumption that the effect sizes in my data will have similar magnitude to those in other studies linking student backgrounds and educational choices. I have examined five related papers, and the "headline" results from each, along with a description of each regression specification and estimate, are shown below in Table 1.

Table 1: Headline Results in Related Literature

Paper	Intent of the paper	Description of the headline result	Estimates (SE)	Sample sizes
Dustmann (2004)	Examining the association between parents' education and professions and children's educational choices in secondary school in Germany.	Marginal differential effects (MDE) of characteristics of parents' education on the probability of their child choosing the most rigorous form of secondary education under a probit model (Table 4). [Further details omitted]	a) Mother-daughter: 0.234 (0.095) b) Mother-son: 0.450 (0.133) c) Father-daughter: 0.297 (0.061) d) Father-son: 0.343 (0.090)	(a) and (c): N = 3,147 (b) and (d) N = 2,970
Tamm (2008)	Using a sibling fixed effects model to determine whether household income has a causal effect on children's educational choices in secondary school in Germany.	[Details on sibling FE omitted] I focus on the effects of household income at different ages on an indicator for choosing the most rigorous secondary education as the headline result (Table 4).	a) Effect of average income at ages 3-6: -0.038 (0.052) b) Effect of average income at ages 7-10: -0.009 (0.054) c) Effect of income at age 10: 0.002 (0.032)	(a): N = 160 (b): N = 233 (c): N = 322
Buser et al. (2014)	Measuring differences in competitiveness between boys and girls and assessing what fraction of the gender gap in educational choices in secondary school may be attributed to competitiveness.	Binary OLS regression, where the outcome variable is an indicator for choosing the most prestigious educational track. I focus on the coefficient on a *Female* gender indicator with and without a control for competitiveness (Table IX, Columns 1 and 2). [Details omitted]	a) No control for competitiveness: -0.195 (0.043) b) Control for competitiveness: -0.178 (0.044)	(a) and (b): N = 362
Dreber et al. (2014)	Measuring gender differences in competitiveness, altruism, and risk preferences, and how differences change based on whether experimental task is verbal or math.	Headline results: estimates from OLS regressions where the dependent variables are indicators for choosing to compete in a math (Table 4) and verbal (Table 5) task and the independent variable of focus is a *Female* indicator.	a) Choosing to compete in the math task: -0.058 (0.070) b) Choosing to compete in a verbal task: 0.066 (0.089)	(a): N = 169 (b): N = 167
Carlana (2019)	Using a class fixed effects regression to estimate the effects of gender and gender-based stereotyping by math and literature teachers on educational choices in high school in Italy.	[Details on class FE omitted] Headline regression (Table VII, Panel A) is an ordered logit where the dependent variable corresponds to three levels of academic rigor. I focus on the coefficient on the *Female* gender indicator in Column 7. [Explanation omitted]	Effect of gender on the decision to choose a more rigorous high school track, with controls including teacher stereotype and class fixed effects: -0.370 (0.048)	N = 25,395

I can calculate the *effect size* of each estimate, as defined as follows: $ES = \frac{|\mu_{treatment} - \mu_{control}|}{\sigma} = \frac{|\mu_{treatment} - \mu_{control}|}{SE\sqrt{N}}$, and the *ex-post statistical power* of each headline result. [Additional details and table of effect sizes omitted] [Details on calculating power omitted] To calculate *ex-post power* of estimate $\hat{\mu}$ for population parameter μ, I calculate $P(\text{reject } H_0 | \mu = \hat{\mu})$. With $X \sim N\left(\mu, \frac{\sigma^2}{N}\right) = N\left(\hat{\mu}, \frac{\sigma^2}{N}\right)$, this is $P\left(X \leq -1.96\frac{\sigma}{\sqrt{N}} + \mu_0\right) + P\left(X \geq 1.96\frac{\sigma}{\sqrt{N}} + \mu_0\right)$, or in terms of the standard normal CDF, Φ, with $\mu_L = -1.96\frac{\sigma}{\sqrt{N}} + \mu_0$ and $\mu_R = 1.96\frac{\sigma}{\sqrt{N}} + \mu_0$, this is given by $\Phi\left(\frac{\mu_L - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}\right) + 1 - \Phi\left(\frac{\mu_R - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}\right)$.

I assume that because our study is similar in subject matter to the five papers listed, its effect size will fall between 0.0035 and 0.2383. The statistical powers of these results are 0.05 and 0.99, respectively. Under the worst-case scenario effect size of 0.0035, I would like to achieve a power of at least 0.80, the standard in existing literature. [Serdar et al. (2021), Hintze (2008)] I can calculate the sample size needed to achieve 0.80 power using effect size. For simplicity, I assume that I am now performing a one-sided hypothesis test, where $H_0: \mu = \mu_0$ and, without loss of generality, $H_1: \mu < \mu_0$. This is a reasonable assumption, as for the papers listed above, I can intuitively hypothesize the sign of each headline result. For example, in Dustmann (2004), I can reasonably expect estimates to be positive. This is because it makes sense to expect children of more educated parents to enter more academically rigorous programs given that existing research has suggested that values and attitudes about education in children and parents tend to be associated with one another. [Casad et al. (2015)] Similarly, in this project, I can assume that negative stereotypes about one's own performance in a subject *s* will be associated with less priority given to subject *s*. Then, WLOG, I assume that the *ex-post power* is given by: $\Phi\left(\frac{\mu_L - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}\right) = \Phi\left(-1.96 + \frac{\mu_0 - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}\right)$. With Type-II error rate $1 - \beta = \Phi\left(-1.96 + \frac{\mu_0 - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}\right)$, we have that $\Phi^{-1}(1-\beta) = -1.96 + \frac{\mu_0 - \hat{\mu}}{\frac{\sigma}{\sqrt{N}}}$, or $N = \left[\frac{\sigma}{\mu_0 - \hat{\mu}}(\Phi^{-1}(1-\beta) + 1.96)\right]^2 = \left[\frac{\Phi^{-1}(1-\beta)+1.96}{ES}\right]^2$ with effect size $ES = \frac{|\hat{\mu} - \mu_0|}{\sigma}$.

Using the worst-case effect size of 0.0035, I would need sample sizes of $N = 640{,}702$ for both students and parents to achieve this desired power, which is unrealistic. However, if I make the less extreme assumption that the effect size will lie between 25th percentile, 0.0461, and the median, 0.0560, then the sample size required for 0.80 power falls into the more realistic range of $N \in [2{,}184, 3{,}686]$. Thus, for the full survey, I will recruit a sample of between 2,200 and 3,700 students and their parents.

Section 5: Prolific as an Online Survey Fielding Platform

For the pilot, I will field a sample of $N = 50$ to parents on *Prolific*, an online survey fielding platform. Although Prolific allows for a variety of demographic filters, such as respondent age and parenthood status, it does not allow me to request that the parents' children be of a certain age. So, I must instead choose the ages of the parents in the sample to maximize the probability that their children are in grades K-8. I will not advertise that the survey is specifically intended to be for parents of children in grades K-8, as that may incentivize respondents to lie about their children's age, hurting data quality. Instead, I assume K-8 students in the United States to be between ages 5 and 13, meaning that they were born between 2010 and 2018. Then, according to the U.S. Census Bureau, the median age of a mother at birth was 27 in 1990 and 30 in 2019 [Morse (2022), Figure 4]. Assuming a linear trend, this puts the median age of a mother with a child in grades K-8 between 35 and 42. Thus, to improve the probability of sampling parents of children in grades K-8, I will restrict the Prolific sample to parents of ages 35-42.

There are numerous options for deploying an online survey, including Amazon Mechanical Turk ("MTurk"), Prolific, and CloudResearch. I will field the pilot survey on Prolific, as existing literature suggests it to be the most cost-effective way of achieving a high-quality sample: two recent papers studying samples of respondents from five different survey platforms found Prolific to have the lowest cost per respondent passing a battery of quality checks. [Douglas et al. (2023)] [Peer et al. (2022)]

Section 6: Model of Scale-Use Heterogeneities

In conducting survey-based research, it is critical to adjust for differences in how people use the response scale, or *scale-use heterogeneity*. For example, when asked about their child's performance in mathematics, Parent 1 and Parent 2 might hold the same attitude about their child's performance, but Parent 1 might map that to a "60" on the 0-100 response scale, while Parent 2 maps it to an "80." Existing literature suggests that systematic scale-use differences may exist between demographic groups. For instance, Greenleaf (1992) found that a measure of a respondent's tendency to mark extreme options on a discrete response scale differed significantly by age, income, and education. More recent work has found that measures of how low or high respondents tend to rate on a scale, the *shifter* parameter, and how much of the scale respondents tend to use, the *stretcher* parameter, both differ significantly across demographic groups. [Benjamin et al. (2023)] This poses a problem for interpreting survey data in a policy analysis setting, as scale-use differences could misrepresent or obfuscate heterogeneities in policy impact. For example, if women, on average, tend to use higher response scale values than men, then using unadjusted survey data meant to measure the effect of a policy could hide an adverse impact on women.

[Rest of section omitted]

Next Steps and Possible Conclusions

Once the pilot is fielded, the next step will be to partner with a school in the Chicago metro area. This research will improve on current assessments of educational priorities because it uses marginal utilities, which allow for greater granularity in measuring educational priorities than indicator variables for educational choices, which are more typical in literature. [Dustmann (2004), Buser et al. (2014)] Moreover, existing work tends to focus on parental backgrounds or stereotyping by teachers, while factors like student self-stereotyping and parental stereotyping are underexplored. [Dustmann (2004), Carlana (2019)]

This project looks for associations between marginal utilities and student gender but does not purport a causal relationship. An association between gender and marginal utility may be due to a causal

link between gender stereotypes and educational priorities. It may also be a combination of stereotypes and value transfer between parents and children: for example, if gender stereotypes influence parents' priorities and those beliefs are passed on to their children. [Casad et al. (2015)] The causal pathway could also be indirect: for instance, stereotypes might affect the social desirability of subjects, such as through a desire to fit into perceived gender roles. [Hackett and Betz (1989)]

This study is meant to serve as the first step in a research program, highlighting gender stereotyping and self-stereotyping as possible factors contributing to gender disparities in educational and professional achievement in certain disciplines. Subsequent work will look for causal relationships, which is a challenging task, as sources of exogenous variation in gender-based stereotyping are not obvious. One possible strategy would be a randomized controlled trial following similar methodology to Spencer et al. (1999) and Huang at al. (2023) where the treatment group would be primed with a reference to gender differences in academic performance. Another option would be a fixed effects model in the style of Carlana (2019) or Tamm (2008). However, this would require several years of panel data with minimal attrition.

Subsequent research would also examine interventions to reduce stereotyping and gender disparities, such as by training teachers to reduce how often stereotypes are referenced and mitigate "stereotype threat," which is when students are afraid of performing poorly and reinforcing a negative stereotype. [Spencer et al. (1999)] Narrowing disparities could draw talent from underrepresented groups, while boosting the performance and willingness to take initiative of members of those groups. [Spencer et al. (1999), Coffman (2014), Carlana and Fort (2022), Bettinger and Long (2005)]

References

Benjamin, D.J. et al. (2014-A). **Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference.** *Am Econ Rev.,* 104 (9): 2698–2735. https://doi.org/10.1257/aer.104.9.2698.

Benjamin, D.J. et al. (2014-B). **Can Marginal Rates of Substitution Be Inferred From Happiness Data? Evidence from Residency Choices.** *Am Econ Rev.,* 104 (11): 3498–3528. https://doi.org/10.1257/aer.104.11.3498.

Benjamin, D.J. et al. (2023). **Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being.** *National Bureau of Economic Research*. http://www.nber.org/papers/w31728.

Bertrand, M., Goldin, C., and Katz, L.F. (2010). **Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors.** *AEJ: Appl. Econ*., 2 (3): 228-55. https://doi.org/10.1257/app.2.3.228.

Betancourt, M.J. (2017). **A Conceptual Introduction to Hamiltonian Monte Carlo.** Preprint, *arXiv*. https://arxiv.org/abs/1701.02434.

Betancourt, M.J. and Girolami, M. (2013). **Hamiltonian Monte Carlo for Hierarchical Models.** In *Current Trends in Bayesian Methodology with Applications*, 79–101. Chapman and Hall/CRC. https://arxiv.org/abs/1312.0906.

Bettinger, E. and Long, B.T. (2005). **Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students.** *Am. Econ. Rev.*, 95 (2): 2005, p. 152-157. https://doi.org/10.1257/000282805774670149.

Buser, T., Niederle, M., and Oosterbeek, H. (2014) **Gender, Competitiveness, and Career Choices.** *Q. J. Econ*., 129 (3): 1409-1447. https://doi.org/10.1093/qje/qju009.

Carlana, M. (2019). **Implicit Stereotypes: Evidence from Teachers' Gender Bias.** *QJE*, 134 (3): 1163–1224, https://doi.org/10.1093/qje/qjz008.

Carlana, M., and Fort, M. (2022). **Hacking Gender Stereotypes: Girls' Participation in Coding Clubs.** *AEA Papers and Proc.,* 112: 583-87. https://doi.org/10.1257/pandp.20221085.

Casad, B.L., Hale, P., and Wachs, J.L. (2015). **Parent-child math anxiety and math-gender stereotypes predict adolescents' math education outcomes.** *Front. Psychol.,* 6 (3) :1597. https://doi.org/10.3389/fpsyg.2015.01597.

Coffman, K. (2014). **Evidence on Self-Stereotyping and the Contribution of Ideas.** *QJE*, 129 (4): 1625–1660, https://doi.org/10.1093/qje/qju023.

Colley, A. and Comber, C. (2003). **School Subject Preferences: Age and gender differences revisited.** *Educ. Stud.,* 29 (1): 59-67. https://doi.org/10.1080/03055690303269.

De Gioannis, E. (2022) **Implicit gender-science stereotypes and college-major intentions of italian adolescents.** *Soc. Psychol. Educ.,* 25: 1093–1112. https://doi.org/10.1007/s11218-022-09709-3.

Douglas B.D., Ewell P.J., and Brauer M. (2023). **Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA.** *PLoS One.* 18 (3): e0279720. https://doi.org/10.1371/journal.pone.0279720.

Dreber, A., von Essen, E., and Ranehill, E. (2014). **Gender and competition in adolescence: task matters.** *Exp. Econ*., 17 (1): 154-172. https://doi.org/10.1007/s10683-013-9361-0.

Dustmann, C. (2004). **Parental background, secondary school track choice, and wages.** *Oxf. Econ. Pap.,* 56 (2): 209–230, https://doi.org/10.1093/oep/gpf048.

Easton, J.Q. and Diaz, B. (2023). **Lasting differences: Math grades and gender.** *Chicago, IL: University of Chicago Consortium on School Research.* https://consortium.uchicago.edu/publications/lasting-differences.

Elmins, W., Joyce, R., and Costa Bias, M. (2016). **The Gender Wage Gap**. Institute for Fiscal Studies Briefing Note 186.

Enke, B., Graeber, T., and Oprea, R. (2023). **Complexity and Time.** *National Bureau of Economic Research.* http://www.nber.org/papers/w31047.

Fabian, M. (2022). **Scale Norming Undermines the Use of Life Satisfaction Scale Data for Welfare Analysis.** *Journal of happiness studies*, 23 (4), 1509–1541. https://doi.org/10.1007/s10902-021-00460-8.

Furnham, A., Reeves, E., and Budhani, S. (2002). **Parents think their sons are brighter than their daughters: sex differences in parental self-estimations and estimations of their children's multiple intelligences.** *J. Genet. Psychol.,* 163: 24–39. https://doi.org/10.1080/00221320209597966.

Greenleaf, E.A. (1992**). Measuring Extreme Response Style.** *Public Opinion Quarterly*, 56 (3): 328–351, https://doi.org/10.1086/269326.

Hackett, G. and Betz, N.J. (1989). **An Exploration of the Mathematics Self-Efficacy/Mathematics Performance Correspondence.** *Res. Math. Educ.*, 20 (3): 261-273. https://doi.org/10.2307/749515.

Hintze, J.L. (2008). **Power analysis and sample size system (PASS) for windows User`s Guide I.** *NCSS.* Kaysville, Utah, USA.

Huang, M., Yi, H., and Rozelle, S. (2023). **On the Origins of Gender Gaps in Education: Stereotype as a Self-Fulfilling Prophecy.** *Preprint*, available at SSRN: https://ssrn.com/abstract=4325965.

Isaksson, S. (2019). **It Takes Two: Gender differences in in group work.** *Working paper*, Norwegian School of Economics. https://www.dropbox.com/s/hafhh70p3li9tij/1120jmp.pdf?dl=0.

Joensen, J.S. et al. (2023). **Using a Field Experiment to Understand Skill Formation During Adolescence.** *Working paper*. https://drive.google.com/file/d/1Opzz1LAzu3UuPKelLg7__ud_w4Am_qac/view.

Kneebone, E. and Berube, A. (2013). **Confronting suburban poverty in America.** *Brookings Institution Press.* https://www.jstor.org/stable/10.7864/j.ctt4cg88q.

Kundu, T. (2023). **Estimating our preferences model using HMC: preliminary results.** *Unpublished.*

Lakens, D. (2021). **Sample Size Justification.** *PsyArXiv*. https://doi.org/10.1525/collabra.33267.

Melstrom, R. and Redding, C. (2020). **Racial Inequality in Chicago: Income and Education**. *Loyola eCommons, School of Environmental Sustainability, Loyola University Chicago: Faculty Publications and Other Works,* 1. https://ecommons.luc.edu/ies_facpubs/38/.

Merchel, S. and Altinsoy, M.E. (2020) **Psychophysical comparison of the auditory and tactile perception: a survey.** *J. Multimodal User Interfaces*, 14: 271–283. https://doi.org/10.1007/s12193-020-00333-z.

Morse, A. (2022). **Stable Fertility Rates 1990-2019 Mask Distinct Variations by Age.** *U.S. Census Bureau.* https://www.census.gov/library/stories/2022/04/fertility-rates-declined-for-younger-women-increased-for-older-women.html.

Michelmore, K. and Sassler, S. (2016). **Explaining the gender wage gap in STEM: Does field sex composition matter?** *RSF*, 2 (4): 194-215. https://doi.org/10.7758/rsf.2016.2.4.07.

Nicholls, M. E. R., Orr, C. A., Okubo, M., & Loftus, A. (2006). **Satisfaction Guaranteed: The Effect of Spatial Biases on Responses to Likert Scales**. *Psychological Science.*, 17 (12): 1027–1028. https://doi.org/10.1111/j.1467-9280.2006.01822.x.

Oganian, A., Iacob, I., and Lesaja, G. (2020). **Multivariate Top-Coding for Statistical Disclosure Limitation.** *International Conference on Privacy in Statistical Databases*, 136-148. https://link.springer.com/chapter/10.1007/978-3-030-57521-2_10.

Oswald, A.J. (2008). **On the curvature of the reporting function from objective reality to subjective feelings.** *Economics Letters,* 100: 369–372. https://doi.org/10.1016/j.econlet.2008.02.032.

Peer, E. et al. (2022). **Data quality of platforms and panels for online behavioral research.** *Behav. Res.,* 54: 1643–1662. https://doi.org/10.3758/s13428-021-01694-3.

Royer, R. (2017). **Probit Model.** *Unpublished.*

Sarsons, H. et al. (2021). **Gender differences in recognition for group work.** *J. Political Econ.,* 129 (1): 101–47. https://doi.org/10.1086/711401.

Serdar, C. C. et al. (2021). **Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies.** *Biochemia medica,* 31 (1): 010502. https://doi.org/10.11613/BM.2021.010502.

Smid, S.S. and Winter, S.D. (2020). **Dangers of the Defaults: A Tutorial on the Impact of Default Priors When Using Bayesian SEM With Small Samples.** *Front. Psychol.*, 11. https://doi.org/10.3389/fpsyg.2020.611963.

Spencer, S.J., Steele, C.M., and Quinn, D.M., 1999. **Stereotype threat and women's math performance**. *J. Exp. Soc. Psychol.,* 35 (1), 4–28. https://doi.org/10.1006/jesp.1998.1373.

Tamm, M. (2008). **Does money buy higher schooling?: Evidence from secondary school track choice in Germany.** *Econ. Educ. Rev.,* 27(5): 536-545. https://doi.org/10.1016/j.econedurev.2007.10.005.

U.S. Census Bureau, American Community Survey (ACS). Updated annually. **American Community Survey.** https://www.census.gov/quickfacts/fact/note/US/RHI625222.