The Development of Gender Stereotypes About STEM and Verbal Abilities: A Preregistered Meta-Analysis Protocol

David I. Miller
American Institutes for Research

Jillian E. Lauer
New York University

Ryan T. Williams
Courtney Tanenbaum
American Institutes for Research

Please direct correspondence to David I. Miller (dimiller@air.org). The following document accompanies our preregistration on the Open Science Framework website made on February 19, 2020 (https://osf.io/29egh/registrations).

This project is funded by the National Science Foundation (NSF Award No. DUE-1920401). Any opinions, findings, and conclusions or recommendations expressed in this project's publications, including in this document, are those of the research team and do not necessarily reflect the views of the NSF.

Abstract and Rationale

This project will study the origins of beliefs and motivational factors that could potentially limit girls' and women's full participation in science, technology, engineering, and mathematics (STEM) fields, as well as contribute to boys' underachievement in verbal domains (e.g., reading and writing). Specifically, this synthesis project aims to bring clarity to the mixed findings on (a) how gender stereotypes about STEM and verbal abilities first develop and (b) how they relate to gender gaps in STEM. Several studies of children have found the expected stereotype of superior male ability in mathematics and science, but others have found only in-group bias or even stereotypes of female superiority. Less research has focused on verbal ability stereotypes, but they are also critical because pro-female verbal stereotypes could potentially draw girls away from quantitative fields, contributing to gender gaps in STEM outcomes. Verbal stereotypes might also limit boys' academic success in reading, writing, and language domains, which generally show moderate to large gaps favoring girls in test performance.

This project consists of two meta-analyses that will analyze variation in (a) mean levels of children's gender stereotypes about STEM and verbal abilities and (b) these stereotypes' correlations with motivational STEM outcomes such as confidence and interests. In both meta-analyses, focal moderators will include child demographics, cultural contexts, and measurement characteristics. Knowledge from this project will help bring clarity to the mixed developmental literature on STEM ability stereotypes, in addition to synthesizing insights from the emerging literature on verbal ability stereotypes. Understanding why one study finds stereotypes strongly favoring males, whereas another study finds the opposite, will be critical to foster cumulative, replicable science and build integrative theories of stereotype development. Furthermore, synthesizing how ability stereotypes relate to outcomes such as confidence and interests can build fundamental knowledge on how these beliefs might relate to gender gaps in STEM participation.

Though both meta-analyses will examine STEM and verbal ability *stereotypes*, the second meta-analysis will focus on just STEM *outcomes* (i.e., relating STEM and verbal stereotypes to STEM outcomes, but not verbal stereotypes to verbal outcomes, for which there is likely limited research). We plan to present the two meta-analyses as two separate journal article manuscripts.

PRISMA-P 2015 Checklist

This checklist has been adapted for use with protocol submissions to *Systematic Reviews* from Table 3 in Moher D et al**:** Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews* 2015 **4**:1

Section/topic	#	Checklist item	Information reported		Page number(s)
			Yes	No	
ADMINISTRATIVE INFORMATION					
Title					
Identification	1a	Identify the report as a protocol of a systematic review	☒	☐	1
Update	1b	If the protocol is for an update of a previous systematic review, identify as such	☐	☐	N/A
Registration	2	If registered, provide the name of the registry (e.g., PROSPERO) and registration number in the Abstract	☒	☐	1
Authors					
Contact	3a	Provide name, institutional affiliation, and e-mail address of all protocol authors; provide physical mailing address of corresponding author	☒	☐	1
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review	☒	☐	5
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments	☒	☐	5
Support					
Sources	5a	Indicate sources of financial or other support for the review	☒	☐	1, 5
Sponsor	5b	Provide name for the review funder and/or sponsor	☒	☐	1, 5
Role of sponsor/funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol	☒	☐	5
INTRODUCTION					
Rationale	6	Describe the rationale for the review in the context of what is already known	☒	☐	1, 8, 9
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)	☒	☐	1, 6

Section/topic	#	Checklist item	Information reported		Page number(s)
			Yes	No	
METHODS					
Eligibility criteria	8	Specify the study characteristics (e.g., PICO, study design, setting, time frame) and report characteristics (e.g., years considered, language, publication status) to be used as criteria for eligibility for the review	☒	☐	12-13
Information sources	9	Describe all intended information sources (e.g., electronic databases, contact with study authors, trial registers, or other grey literature sources) with planned dates of coverage	☒	☐	9-11
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated	☒	☐	Appendix A
STUDY RECORDS					
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review	☒	☐	15
Selection process	11b	State the process that will be used for selecting studies (e.g., two independent reviewers) through each phase of the review (i.e., screening, eligibility, and inclusion in meta-analysis)	☒	☐	14, Appendix B
Data collection process	11c	Describe planned method of extracting data from reports (e.g., piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators	☒	☐	15-16, Appendix C
Data items	12	List and define all variables for which data will be sought (e.g., PICO items, funding sources), any pre-planned data assumptions and simplifications	☒	☐	15-16, Appendix C
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale	☒	☐	8, 16-18
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis	☒	☐	15
DATA					
Synthesis	15a	Describe criteria under which study data will be quantitatively synthesized	☒	☐	13
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data, and methods of combining data from studies,	☒	☐	16-24

Section/topic	#	Checklist item	Information reported		Page number(s)
			Yes	No	
		including any planned exploration of consistency (e.g., I^2, Kendall's tau)			
	15c	Describe any proposed additional analyses (e.g., sensitivity or subgroup analyses, meta-regression)	☒	☐	16-24
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned	☐	☐	N/A
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (e.g., publication bias across studies, selective reporting within studies)	☒	☐	21, Appendix D
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (e.g., GRADE)	☒	☐	23

Amendment Timeline

Project Status at Time of Initial Preregistration

We submitted an initial draft of the following protocol for an NSF grant proposal. We responded to NSF panel reviewers' comments, after which NSF notified us of winning the grant on July 9, 2019 (Award No. DUE-1920401). We then reviewed the proposed plan in more detail with an external advisory board of five experts (Andrei Cimpian, Beth Kurtz-Costes, Catherine Riegle-Crumb, Jo Boaler, Larry Hedges) during a 2-hour meeting on November 4, 2019. Lastly, our project's internal quality assurance review, Martyna Citkowicz, reviewed this document before we finalized and preregistered it on the Open Science Framework website on February 19, 2020 (https://osf.io/29egh/registrations).

Hence, this protocol collectively represents input from project team members, anonymous NSF grant panel reviewers, external advisory board members, and our project's designated internal quality assurance reviewer. As noted in the following section on "Explanation of Existing Data," the PI had analyzed a small subset of studies on children's ability stereotypes (14 studies) six years prior to this preregistration, but he did not reanalyze those data to form this current protocol, which greatly differs in scope and approach from that initial investigation.

We used the following standardized table from PROSPERO, an international prospective register of systematic reviews, to document the project status at the time of preregistration:

Project Status at Preregistration (February 19, 2020)

Stage	Started	Completed
Preliminary searches	Yes	Yes
Piloting the study selection process	Yes	Yes
Formal screening of search results against eligibility criteria	Yes	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

D. Miller drafted the protocol document and is the guarantor (i.e., responsible for the overall scientific integrity of the work). D. Miller, J. Lauer, and C. Tanenbaum contributed to the conceptual development of the protocol, including to its inclusion criteria, directional hypotheses, and coding schemes. D. Miller, J. Lauer, R. William contributed to the technical development of the protocol, including to its effect size metrics, analysis plans, and literature search. J. Lauer also drafted the coding protocol and codebook. All authors read, provided feedback and approved the final protocol.

Amendments to Review Protocol

If we need to amend this protocol, we will give the date of each amendment, describe the change, and give the rationale in this section. The text of the later sections will not be altered; the time-stamped versions on the Open Science Framework (https://osf.io/29egh/registrations) will also document and verify the timeline. The PI, David Miller, will be responsible for approving changes and updating this section at least every year, if not more frequently.

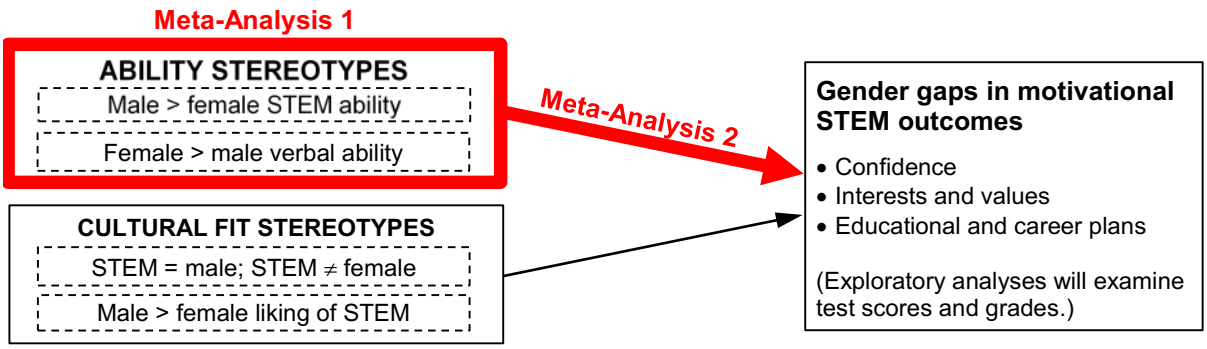
Initial Preregistration (February 19, 2020)

Research Questions

This synthesis project aims to bring clarity to the mixed findings on how gender stereotypes about STEM and verbal abilities first develop and relate to gender gaps in STEM outcomes. We will investigate two core research questions, which will be investigated across two sets of statistical analyses (see Figure 1):

1. How do children's gender stereotypes about STEM and verbal abilities vary across child demographics, cultural contexts, and measures?
2. Do children's gender stereotypes about STEM and verbal abilities correlate with motivational STEM outcomes? How do these stereotype-outcome correlations vary across key moderators?

Figure 1. Focus of This Proposed Project (Regions Highlighted in Red)

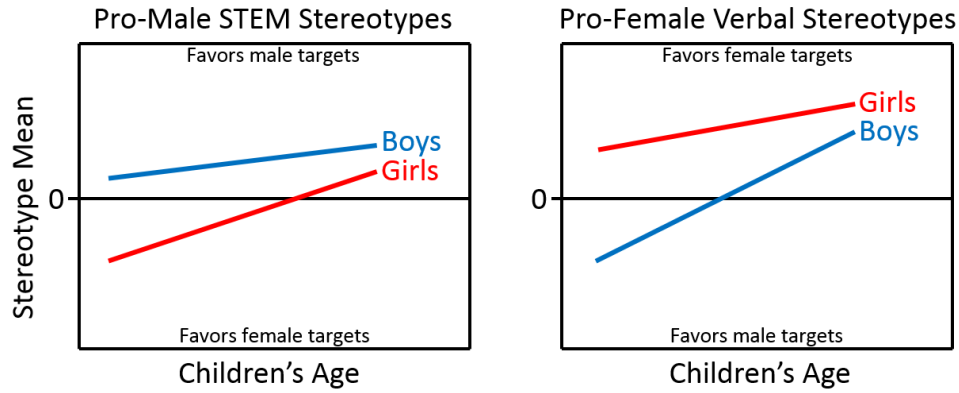


Directional Hypotheses

Based on theoretical frameworks from developmental science, social psychology, and educational research, we preregister the following directional confirmatory hypotheses. Consistent with community norms for preregistration, we do not detail these theoretical considerations in depth here. Rather, we simply note which predictions we will empirically test as preregistered confirmatory hypotheses. Exploratory analyses may examine other moderators and hypotheses, but we will clearly label results from such analyses as tentative and distinct from our confirmatory analyses.

Meta-Analysis 1: Figure 2 shows our core predictions for how stereotype means will vary with age, gender, and domain (STEM vs. verbal ability). For both domains, we predict that conventional stereotypic beliefs (pro-male STEM and pro-female verbal) will strengthen as children age. Children should also favor their own sex such that (a) pro-male STEM stereotypes will be stronger for boys than girls and (b) pro-female verbal stereotypes will be stronger for girls than boys. However, in-group preferences should decline with age (i.e., meaning there will be an Age × Gender interaction). In addition, we predict that the overall magnitude of pro-female verbal stereotypes will be stronger than pro-male STEM stereotypes.

Figure 2. Meta-Analysis 1 Predictions for Age, Gender, and Stereotype Domain



As also shown in Figure 2, another way to view the Age × Gender hypothesis is that developmental increases in conventional stereotypic beliefs should be stronger for (a) girls' than boys' STEM stereotypes and (b) boys' than girls' verbal stereotypes. Table 1 details these hypotheses, along with others for race/ethnicity, cultural context (cross-national and cross-temporal), and measurement characteristics.

Table 1. Directional Hypotheses for Meta-Analysis 1

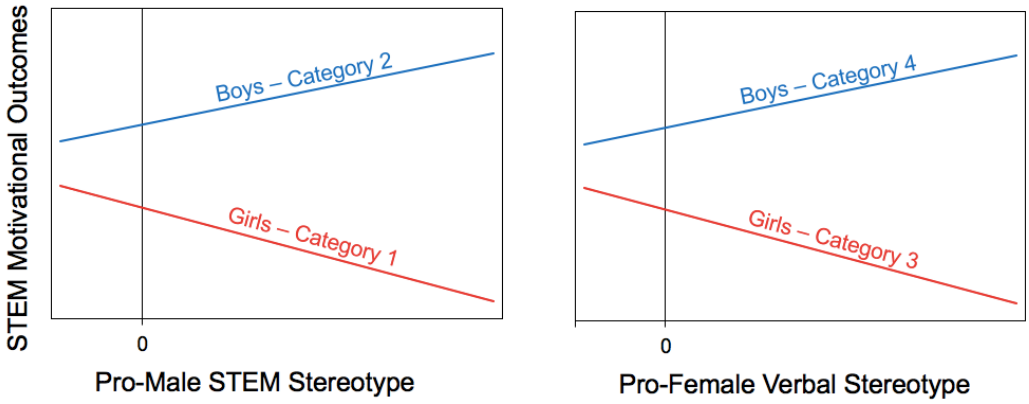
Moderator	Pro-M STEM	Pro-F Verbal
Child Demographics		
Children's age (older vs. younger)	+	+
Male (vs. female) child participant	+	-
Age × Male	-	+
Black versus White, non-Hispanic children (U.S. only)	-	
Cultural Context		
Data collection year (U.S. only)	-	
National % female among STEM majors	-	
National % female among employed researchers	-	
Stereotype Measurement Characteristics		
Indirect (vs. direct) measure	+	
Adult (vs. child) stereotype target	+	

Note. An empty cell means no specific directional hypothesis. The hypotheses for race/ethnicity and data collection year are specific to the United States. Indirect stereotype measures are those that do not mention gender explicitly (e.g., ask children to draw a student good at math), whereas direct measures explicitly probe about gender differences (e.g., "Do you think boys or girls are better at math?"). The last row reflects that the stereotype measure could use adult targets (e.g., "Are women or men better?") or child targets (e.g., "Are girls or boys better?"). In addition to these hypotheses, we predict that pro-female verbal stereotypes will be overall stronger in magnitude than pro-male STEM stereotypes, as reflected in Figure 2.

We expect there will be fewer studies of verbal ability stereotypes, which is why we make a more restricted set of hypotheses for the verbal than STEM domain. In addition, the relevant theory informing the STEM-related hypotheses may not directly extend to the verbal domain.

Meta-Analysis 2: For Meta-Analysis 2, we predict that children's pro-male STEM and pro-female verbal ability stereotypes will correlate with STEM motivational outcomes (e.g., confidence, interests) negatively for girls and positively for boys, as shown in Figure 3.

Figure 3. Different Categories of Stereotype-Outcome Correlations for Meta-Analysis 2



Hence, Meta-Analysis 2 will focus on four categories of correlations (i.e., how STEM and verbal stereotypes relate to STEM motivational outcomes, separately for boys and girls). For correlations between STEM stereotypes and STEM outcomes among girls (i.e., Category 1), we also predict those correlations will be stronger in magnitude (i.e., larger negative values) for:

1. Older than younger children
2. Indirect than direct stereotype measures
3. Stereotype measures with child than adult targets
4. Confidence than other outcomes

We restrict these confirmatory moderator hypotheses to Category 1 (i.e., STEM stereotype–STEM outcome correlations for girls) because the relevant theory most directly informs hypotheses for those correlations and we expect to gather the most primary data for that category. However, in exploratory analyses, we will examine variation within the other categories and consider other moderators. In addition, exploratory analyses will examine correlations with two categories of STEM performance outcomes (test scores and grades), but we do not have strong a priori predictions for these analyses given the lack of widespread gaps favoring boys in average STEM performance (Miller & Halpern, 2014).

Literature Search ("Sampling Plan")

Unlike traditional sampling plans for primary research studies, we aim to comprehensively synthesize all relevant prior literature using meta-analytic techniques. The following sections detail our plan for doing so.

Explanation of Existing Data

Our preliminary searches (i.e., a "scoping" review) examined citations to highly influential papers (e.g., Ambady, Shih, Kim, & Pittinsky, 2001) and found 58 potentially eligible studies with more than 25,000 children, demonstrating a voluminous literature. Some studies of children have found the expected stereotype of superior male ability in mathematics and science (e.g., Hargreaves, Homer, & Swinnerton, 2008), but others have found only in-group bias (e.g., Heyman & Legare, 2004) or even stereotypes of female superiority (e.g., Rowley, Kurtz-Costes, Mistry, & Feagans, 2007). The evidence is also mixed on how these stereotypes relate to gender gaps in motivational outcomes, such as confidence, interests, and future career plans in STEM (e.g., Evans, Copping, Rowley, & Kurtz-Costes, 2011). Several aspects of this scoping review informed the development of our review protocol (e.g., we used common phrases in the abstracts to develop our tentative list of keywords for the literature search). However, at the time of preregistration, we have not completed our formal screening process or coded studies (see Table 2). We have nearly finished our initial abstract screening but have not yet started full text screening.

The PI analyzed a small subset of studies on children's ability stereotypes (14 studies) more than six years ago in July 2013 but has not since reanalyzed those data (e.g., the PI did not reanalyze the data to prepare this protocol). This cursory analysis examined three moderators (gender, age, and stereotype domain for math vs. verbal). Some results were consistent with our current hypotheses (i.e., stronger pro-female verbal than pro-male math stereotypes, in-group preferences that declined with age), but other results were not consistent (i.e., no main effect of age). We still predict a main effect of age based on developmental theory and a related published meta-analysis (Miller, Nolla, Eagly, & Uttal, 2018), even though the earlier analysis of ability stereotype did not find such an effect. We consider our current protocol to be a preregistration because it greatly differs from this prior analysis both in terms of scope and approach. For example, in contrast to the 14 studies included in the earlier analysis, we expect to code and analyze roughly 80 to 120 studies for this project.

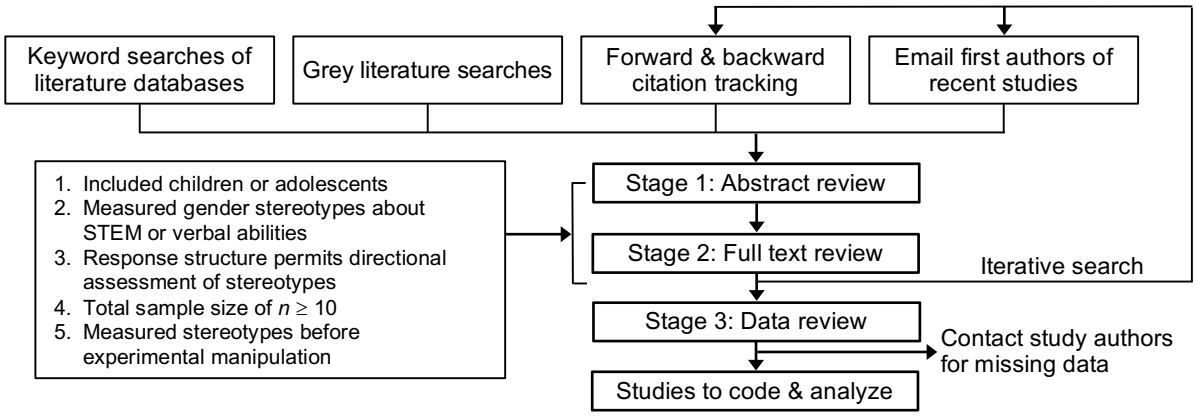
Table 2. Project Status at Preregistration

Stage	Started	Completed
Preliminary searches	Yes	Yes
Piloting the study selection process	Yes	Yes
Formal screening of search results against eligibility criteria	Yes	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

Search Strategy ("Data Collection Procedures")

We will gather citations to potentially relevant studies using (a) keyword searches of literature datasets, (b) grey literature searches, (c) citation tracking, and (d) emails to study authors (Figure 4). Each search strategy includes systematic methods for finding unpublished studies, such as using ProQuest to identify doctoral dissertations in Strategy 1. Finding grey literature is critical because reviews can otherwise produce distorted conclusions by focusing on well-known, easily available, published studies. Our four-pronged, multidisciplinary search process will help mitigate such biases.

Figure 4. Overview of Literature Search and Screening Process



Strategy 1 (Keyword Searches): We searched for studies whose titles, abstracts, or author-provided keywords included at least one keyword from each column in Table 3. In November 2019, we searched these 12 databases: Academic Search Complete, Education Full Text, Education Research Complete, Education Resources Information Center (ERIC), Education Source, GenderWatch, ProQuest Dissertations & Theses Global, PsycINFO, Scopus, Social Sciences Full Text, SocINDEX with Full Text, and Web of Science Core Collection. After removing duplicate citations using the *revtools* R package (Westgate, 2019), these keyword searches yielded 7,377 unique citations.

We identified the search terms through an iterative development and testing process that aimed to balance comprehensiveness and feasibility. First, we used the new *litsearchr* R package (Grames, Stillman, TIngley, & Elphick, 2019) to identify common terms from the titles and abstracts of eligible studies from our scoping review. Second, an experienced research librarian suggested additional terms and advised us on translating the search syntax across databases. Third, using test searches, we evaluated the usefulness of new terms by examining the number and relevance of additional search hits; general phrases that retrieved far too many irrelevant hits (e.g., "beliefs") were translated into more specific, relevant variants (e.g., "belie* about gender*"). Fourth, we used the *litsearchr* package again to consider more terms based on all retrieved titles and abstracts from a revised search. The file 00_keyword_search.R (https://osf.io/a3prx/) details this process, and Appendix A lists the exact search strings that can be directly copied and pasted into various literature databases search engines.

Table 3. Search Terms for Keyword Searches (Asterisks Denote Wildcard Characters)

Domain	Gender	Stereotype	Age
math*	gender*	stereotyp*	child*
science	sex	gender* perception*	adolescen*
technol*	boy*	gender* belief*	boy*
engineering	girl*	gender* bias	girl*
comput*	female*	male domain	grade*
STEM	male*	female domain	preschool*
spatial*	women*	belie* about [gender term]	pre-school*
mental rotation	men*	belie* that [gender term]	pre-kindergart*
quant* abilit*	son*	belie* [gender term]	prekindergart*
quant* achievement	daughter*	percei* about [gender term]	kindergart*
quant* performance		percei* that [gender term]	elementary school*
verbal abilit*		percei* [gender term]	elementary education
verbal achievement		percep* about [gender term]	middle school*
verbal performance		percep* that [gender term]	high school*
academic domain*		percep* [gender term]	highschool*
academic abilit*			junior high
academic achievement			primary school*
academic performance			primary education
cognitive abilit*			secondary school*
cognitive achievement			secondary education
cognitive performance			elementary secondary
intellectual abilit*			youth*
intellectual achievement			teen*
intellectual performance			K12*
reading			K-12*
writing			PK12*
intelligen*			PK-12*
			school-age*

Note. Keyword searches identified studies whose titles, abstracts, or author-provided keywords included at least one keyword from each column. We searched additional fields for the fourth column (age delimiters) to take advantage of searchable fields like ERIC's "Education Level" and PsycINFO's "Age Group" fields (see Appendix A for details and the exact search strings used across all databases). In addition, we searched the subject terms fields in PsycINFO for the first and third columns.

Strategy 2 (Grey Literature Searches): In addition to searching databases that index grey literature (e.g., ProQuest Dissertations & Theses), we will find grey literature by searching (a) conference programs, (b) federal grant abstracts, and (c) additional websites (e.g., Open Science Framework).

The search interfaces for most of these sources have limited functionality (e.g., not allowing for complex Boolean keyword searching), so we will generally use simple search phrases to index them. We will use the keyword "stereotypes" and its variants (e.g., stereotyped, stereotypic) as the default search phrase(s), but we will adapt the terms as needed for each source (e.g., "stereotypes" would index far too many hits

for social psychology conference programs, so we would instead use domain-related keywords such as "STEM" and "math" in that case).

Strategy 2A (Conference Programs): We will search the programs from the following conferences (the parentheses indicate the years for which the programs are publicly available):

- American Educational Research Association (2005 to 2019)
- Cognitive Development Society (1999 to 2019)
- Gender Development Research Conference (2018)
- Society for Personality and Social Psychology (2003 to 2019)
- Society for Research on Adolescence (2010 to 2018)
- Society for Research on Child Development (2015 to 2019)
- Society for the Psychological Study of Social Issues (2008 to 2019)
- Society of Experimental Social Psychology (2011 to 2019)

In total, these searches include 76 total conferences, counting multiple years as separate conferences. After identifying promising presentations based on the titles and abstracts, we will email the first author for the full presentation (e.g., slides, poster) or related reports (e.g., manuscripts) to assess eligibility.

Strategy 2B (Federal Grant Abstracts): We will identify relevant federal research grants based on keyword searches of project-level abstracts from these funding agencies:

- Institute for Education Sciences (https://ies.ed.gov/funding/grantsearch/)
- National Institutes for Health (https://projectreporter.nih.gov/)
- National Science Foundation (https://www.nsf.gov/awardsearch/advancedSearch.jsp)

We will identify publications resulting from those grants by entering the grant award number into three literature databases (ERIC, PsycINFO, Web of Science) that include funding information as a searchable field. We will also enter the award number into Google Scholar, which will search the publications' full text (e.g., acknowledgements section). Lastly, for grants that are the most promising (e.g., clearly relevant based on the award abstract), we will email the Principal Investigator for any other publications.

Strategy 2C (Additional Websites): We will also search these following websites for relevant studies. We will use our full keyword search string (see Appendix A) for the Open Science Framework search interfaces because they allow for complex, nested Boolean search strings.

- EdWorkingPapers (https://edworkingpapers.com/)
- Open Science Framework website (https://osf.io/search/)
- Open Science Framework preprints (https://osf.io/preprints/)
    - Indexes multiple preprint servers such as PsyArXiv, SocArXiv, and EdArXiv.
- Think Tank Search (https://guides.library.harvard.edu/c.php?g=310680&p=2072552)

Strategy 3 (Forward and Backward Citation Tracking): As an iterative method, we will use two databases that track citation networks (Google Scholar and Scopus) to examine all citations to and from eligible studies. This strategy is promising because most eligible studies will likely have cited at least one other eligible study. Also, Google Scholar could help find unpublished studies because it indexes many grey literature sources, such as faculty websites, research organization websites, and preprint servers (Haddaway, Collins, Coughlin, & Kirk, 2015).

Strategy 4 (Emails to Study Authors): As another iterative method, we will invite the first author of any eligible study published in the last 10 years to send any other relevant studies and recommend other researchers who may have relevant unpublished data. We will focus on authors of recent studies to maximize the chances of receiving replies from people who still actively study ability stereotypes.

Expected Number of Citations to Examine: The keyword searches (Strategy 1) yielded 7,377 unique citations (after removing duplicates). We expect that Strategy 3 (citation tracking) will add many more unique citations, especially after merging citations from Google Scholar, probably at least by another 5,000 unique citations, based on citation searching using eligible studies from our scoping review. Strategies 2 and 4 will likely add another 100–200 citations. In total, we expect to screen approximately 12,000 to 13,000 unique citations. At this preregistration stage, we have not yet completed the iterative search methods (Strategies 3 and 4); in addition, we plan to update our literature searches before submitting the first manuscript to a peer-reviewed journal.

Inclusion Criteria ("Data Collection Procedures")

When screening citations, we will include studies meeting all the following criteria. We will place no restrictions on the publication year, study location, or language. Though our keyword searches are based in English, many literature databases include translated English abstracts for studies written in other languages, meaning our searches could still index non-English studies. In those cases, we will attempt to translate the study when making eligibility decisions.

Criterion 1 (Included Children or Adolescents): We will include samples with a mean age of less than 18 years (or PK–12 students) to focus on when ability stereotypes likely start to develop and solidify.

Criterion 2 (Measured Gender Stereotypes About STEM or Verbal Abilities): The term *ability stereotypes* will be broadly defined to encompass both beliefs about raw "innate" intelligence and performance considered more generally (e.g., in an academic, cognitive, or occupational domain). Studies must have included a stereotype measure about gender differences in STEM or verbal abilities, including (a) abilities in any specific STEM subject (e.g., mathematics) or verbal domain (e.g., reading, writing, language arts), or any group of subjects (e.g., STEM as an aggregate category); (b) academic or cognitive test performance; (c) spatial abilities such as mental rotation; and (d) job performance or abilities to pursue a STEM career (studies about occupational ability stereotypes will be limited to STEM jobs because "verbal jobs" is not a well-defined occupational category). Measures of perceptions of others' stereotypes (e.g., "Do adults think boys or girls are good at math?") will be included.

We will exclude measures strictly about representation in a field (e.g., "Are engineers typically men or women?") or gender role attitudes (e.g., "Is science more appropriate for boys or girls?") because they do not directly measure beliefs about abilities. Associative measures (e.g., Draw-A-Scientist Test, Implicit Association Test) will also be excluded, except if they directly related to abilities or performance (e.g., "Draw a student *who is good at* math"). For instance, the study on math-gender stereotypes by Cvencek, Meltzoff, and Greenwald (2011) will be excluded because the implicit measure was about associative stereotypes and the explicit measure was about who likes math (i.e., a cultural fit stereotype).

We will also exclude domain-general ability stereotypes such as beliefs about "brilliance" or "being smart" generally, not connected to any specific field such as math (e.g., Bian, Leslie, & Cimpian, 2017). We will flag these domain-general studies during screening and may later decide to include them in supplemental analyses; however, they will not be included in our focal confirmatory analyses.

Criterion 3 (Response Structure Permits Directional Assessment of Stereotypes): The measure's response structure must permit an unambiguous, directional assessment of stereotypes, which is needed to investigate our research hypotheses. The measure must allow respondents to express beliefs of male or female superiority in a symmetric way.

Acceptable response structures include, but are not necessarily limited to, (a) measures with direct comparisons of female and male targets (e.g., "Are women or men better at math?"), (b) measures with separate ratings of female and male targets (e.g., on a visual analog scale), (c) Likert measures based on agreement to male-biased items (e.g., "boys are better at math") subtracted by agreement to analogously worded female-biased items (e.g., "girls are better at math"), (d) indirect measures that present male and

female targets (e.g., ask children to choose the best student at science among pictures of boys and girls but never mention gender explicitly), and (e) measures with continuous responses (e.g., slider scale) or a discrete number of responses (e.g., two-alternative forced choice measures).

Some other measures are more ambiguous, such as those asking children to rate their agreement with statements about gender equality (e.g., "females are as good as males in geometry"; Fennema & Sherman, 1976). Disagreements with such statements are ambiguous because they could indicate beliefs of either male or female superiority (Forgasz, Leder, & Kloosterman, 2004). Even disagreements to directional statements (e.g., "boys are better than girls") are ambiguous if the scale lacked analogously worded statements in the opposite direction (e.g., "girls are better than boys"). Hence, our response structure criterion will maintain the minimum standards needed to investigate our research hypotheses and derive directional effect size estimates (i.e., stereotypes favoring male vs. female targets).

Criterion 4 (Sample Size Requirement): We will require a minimum sample size of at least 10 total children, summed across all subgroups, to exclude very small studies with questionable sampling bias (e.g., a researcher selects five students to interview who the researcher thought would yield interesting answers). However, we will still include and code subsamples with less than 10 children (e.g., 7 girls and 8 boys) if the total study-level sample size is 10 or more.

Criterion 5 (Measured Stereotypes Before an Experimental Manipulation): We will only include experimental or quasi-experimental studies if children's stereotypes were measured before administering the intervention or manipulation. Likewise, for Meta-Analysis 2, we will only include correlations with outcomes measured before an intervention or manipulation.

We impose these study design requirements because we aim to study naturalistic variation in children's stereotypes, meaning we want to estimate children's baseline stereotypes in the absence of study-specific educational interventions or experimental manipulations. For instance, a stereotype threat manipulation could either strengthen or weaken stereotypes (e.g., Galdi, Cadinu, & Tomasetto, 2014), creating variation in stereotypes that is extraneous to our project's research goals. As sensitivity analyses, we may include experiments under other certain circumstances (e.g., reported data separately for the control condition or found no significant differences across conditions), but our confirmatory analyses will only include studies that measured stereotypes before an experimental manipulation.

Criterion 6 (Ability Items in Multi-Item Scales): For stereotype scales with multiple items, we will require that at least half of the items are about ability stereotypes, if the item-level means cannot be obtained. We will conduct a sensitivity analysis to examine if any of our core results change with more stringent criteria (e.g., requiring that 100% of the items are ability-related). In general, we will first attempt to contact authors for data from only the ability stereotype items.

Contacting Study Authors for Needed Effect Size Data: To include studies in statistical analyses, we must obtain sufficient data to compute effect sizes (e.g., means in Meta-Analysis 1) for the overall sample or at least one subgroup (e.g., girls). We will email study authors if the study report lacks the needed information, starting with the first author, sending two reminders, and then trying other study authors. Because obtaining such missing information can be challenging, we will use established best practices for increasing author response rates, such as setting a response deadline (but still accept replies after then) and including a detailed data-sharing agreement (Polanin & Terzian, 2019; Polanin & Williams, 2016).

Additional Criteria for Meta-Analysis 2: Meta-Analysis 2 will include the subset of studies that reported how STEM ability stereotypes correlated with at least one eligible motivational STEM outcome (i.e., confidence, interests and values, or future educational and career plans in STEM subjects) or STEM performance outcome (i.e., test scores or grades). Eligible studies must report these stereotype-outcome correlations disaggregated by children's gender (or for only one gender group).

Screening Strategy ("Data Collection Procedures")

Three-Stage Screening: Using the above inclusion criteria, we will screen citations in three stages (see Appendix B for our preliminary screening manual). In Stage 1 (abstract review), screeners will examine titles and abstracts using the screening tool *Abstrackr* to help facilitate and manage this initial review (Rathbone, Hoffmann, & Glasziou, 2015); screeners will exclude only obviously ineligible articles at this stage. In Stage 2 (full text review), screeners will thoroughly examine the full text, including only articles meeting all the inclusion criteria. In Stage 3 (data review), we will contact study authors for missing information needed to compute effect sizes and then exclude studies in which no effect sizes can be computed, even after sending author queries. We separate Stages 2 and 3 to track otherwise eligible studies lacking needed effect size data (i.e., we will keep a list of such studies). In practice, Stage 3 will be combined with study coding; before coding a study, the coder will first determine whether effect sizes can be extracted and flag studies for author query as needed.

Screener Training: After developing a screening manual with specific examples of eligible and ineligible studies, Dr. Miller will lead the training to orient the screening team to the project's goals and review protocol. In the first training session, project staff will screen 50 titles and abstracts together. We will review any areas of disagreements and revise the screening manual for further clarity if needed. In each screening stage, the PI will meet weekly with screeners as well as dual screen 15% of the citations to minimize inclusion and exclusion errors. Two designated screeners will dual screen an additional 15% of the citations (i.e., at least 30% of the citations will be dual screened at each stage). The PI will help adjudicate any disagreements and provide further training if interrater reliability falls below standards (Fleiss' $\kappa < 0.80$).

Expected Sample Size

Our initial scoping review found 58 potentially eligible studies. Because our four-pronged systematic search process will likely find many more eligible studies, we expect to code approximately 80–120 primary studies. Studies from our scoping review generally had substantial sample sizes (average N = 481), which means our analyses likely will be well powered to detect even small effects. Using Hedges and Pigott's (2001) formulas and assuming 80 studies with N = 481, we estimate there would be sufficient power (i.e., 80% power) to detect a minimum mean effect size of 0.07 assuming moderate study heterogeneity ($\tau = 0.2$; $I^2 = 95\%$) and 0.10 for large heterogeneity ($\tau = 0.3$; $I^2 = 98\%$).

We will include all eligible studies that we find through search processes, regardless of the final number of studies to analyze. After completing our initial literature search and screening, we plan to update the literature search before submitting our first research manuscript (see Year 2, Quarter 3 in Table 4) but we do not plan further updates after the first time of submitting the manuscript for peer review.

Table 4. Projected Timeline of Major Research Tasks

	Year 1				Year 2				Year 3			
Project Task	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Finalize/preregister review protocol manuals												
Develop database and data entry forms												
Search literature and remove duplicate citations												
Train junior staff and screen citations												
Code studies and contact study authors												
Conduct statistical analyses/develop website												
Journal article 1 (report of Meta-Analysis 1)												
Journal article 2 (report of Meta-Analysis 2)												
Journal article 3 (narrative review of findings)												
Targeted outreach to teachers and nonprofits												
Prepare NSF annual reports; attend PI meeting												
Host advisory board meeting												

Variables

Coding Moderators ("Measured Variables")

After identifying eligible studies, trained coders will record the quantitative information needed to later compute effect sizes (e.g., sample sizes, means) and code three categories of moderators: demographic, contextual, and measurement (see Table 5 for a summary and Appendix C for more details).

Table 5. Study Characteristics to Be Coded as Moderators

Type	Variable	Additional Description
Demographic	Child age	Record both mean age and grade level
	Gender	% girls and boys
	Race/ethnicity	% White, Black, Hispanic, Asian, Multiracial, and Other
	Socioeconomic status	% low income or eligible for free/reduced-priced lunch
Contextual	Year tested	If missing, subtract constant from publication year at analysis stage
	Country	Country in which data were collected (select from list)
	U.S. region	West, Midwest, South, and/or Northeast (U.S. samples only)
	School locale	Includes students in rural, urban, and/or suburban schools
	School type	Includes students in public and/or private schools
Measurement	Stereotype domain	Mathematics, physical science, life science, spatial, etc.
	Stereotype task type	Direct or indirect
	Stereotype age target	Children or adults presented as stereotype targets
	Stereotype exceptionality	Measure is about exceptional giftedness (e.g., best student in the class) or average competence
	Stereotype reliability	Single item or multi-item scale. If multi-item, record internal consistency and type of reliability metric (e.g., Cronbach's α)
	Stereotype question wording	Better, ability/talent, skills, can, indirect (see Appendix C)
	Stereotype scale[a]	7-point scale, 3-point scale, forced choice, visual analog scale, etc.
	Outcome type	Motivational (confidence, interests/values, or future academic/career plans) or performance (test scores or grades)
	Outcome domain	Mathematics, physical science, life science, spatial, etc.

[a]As explained in the analysis plan, we will control for stereotype scale type in all moderator analyses.

Whenever possible, we will code the overall study sample and demographic subsamples (e.g., girls and boys; third and fifth graders). Hence, each study may include multiple samples. In addition, each sample may have completed multiple stereotype measures, and each stereotype measure may generate multiple effect sizes (e.g., means at different time points or correlations with multiple outcomes). We will use a relational database (Microsoft Access) to efficiently code and link these nested levels of information.

For pre-post and longitudinal designs, we will code data from the first occasion when stereotypes were measured (e.g., before an experimental manipulation, as noted in the earlier inclusion criteria section). For studies reported in multiple articles (e.g., a dissertation and journal article), we will examine all articles for relevant information and combine as needed when entering data into the Access database.

Study Quality: As noted earlier, Inclusion Criterion 3 will maintain the minimum standards needed to investigate our research hypotheses and derive directional effect size estimates. However, we will consider other methodological quality indicators, beyond an acceptable response structure, by systematically coding for them as moderators. For instance, Table 5 notes how we will code for measure reliability by (a) distinguishing whether the measure was a single item or multi-item scale and (b) recording the internal consistency and type of reliability metric (e.g., Cronbach's α) if the measure was a multi-item scale. In addition, we will code if the study's measures were created by the researchers for the specific study or if they were broader measures used by the research community (e.g., such as the *Who and Mathematics* instrument; Forgasz et al., 2004).

Transformations: Whenever possible, we will record both the average and range for children's age and grade level; children's average age will be the focal moderator in analyses. If a study reports an age range but not average age, we will impute children's age as the midpoint of the range (e.g., "four- and five-year-olds" would be coded as 5.0 years; "four-year-olds" would be 4.5 years). If the grade level but not age is reported, we will add 5.5 years to the grade level for U.S. samples (e.g., "third graders" would be 8.5 years) and adjust this transformation as necessary for nations with different grade bands (e.g., "Year 3" would be 9.5 years for U.K. samples). We will attempt to contact study authors for more information if neither age nor grade level information was reported.

Publication status will differentiate between (a) articles in academic journals and edited books versus (b) dissertations, Master's theses, conference papers, blog posts, and unpublished manuscripts. If a dissertation or other unpublished report was later published in a journal or edited book, we will consider that study to be published.

Coder Training: Dr. Miller will train junior staff on the coding protocol. Coders will first learn the Microsoft Access data entry forms, then independently code three studies, discuss disagreements at a following training session, and then code an additional 5–10 studies. Training sessions will continue until coders achieve adequate interrater reliability (Fleiss' $\kappa > .80$). Following successful training, 30% of the studies will be dual coded. The PI will meet weekly or biweekly with the coders, assist with dual coding of studies, assess interrater agreement throughout coding, and provide additional training if needed.

Computing Effect Sizes ("Indices")

Meta-Analysis 1: The primary effect size metric for Meta-Analysis 1 will be mean levels of ability stereotypes, after transforming the original stereotype scales (e.g., 1–5 range) onto a common scale ranging from -1 to 1. Positive values will indicate conventional stereotypes (favoring male STEM ability or female verbal ability), and a value of 1 will indicate the maximum possible stereotype mean in that direction (e.g., for STEM stereotypes, all children selected the most extreme pro-male endpoint). For instance, a value of +0.35 would indicate 35% of the maximum possible pro-male STEM stereotype. This rescaling is a variant of the proportion of maximum possible (POMP) scoring method proposed by Cohen, Cohen, Aiken, and West (1999), which has been applied in several meta-analyses when the effect metric is means rather than mean differences (e.g., Fischer & Chalmers, 2008; Fischer & Boer, 2011).

One key advantage of POMP scoring is allowing for separate analysis of means and sample variability. For instance, as children age, their ability beliefs might become more consensual (i.e., less variable), causing the standard deviation (SD) to decrease. An approach based on standardized means (i.e., dividing means by the sample SDs; see Metric 3 in Table 6) would conflate changes in means and SDs. As Viechtbauer (2007) cautioned, "the problem with standardized effect sizes is their dependence on the amount of variability in the population…two *d* or *g* values could be incommensurable if the samples were drawn from populations with unequal variances" (p. 59; see also Baguley, 2011; Bond, Wiitala, & Richard, 2003). In contrast, because POMP scoring uses known features of the scale range as the "standardizer," POMP means and SDs can be separately analyzed (see Metrics 1 and 4 in Table 6).

Table 6. Effect Metrics for Meta-Analysis 1 (Focal One Highlighted in Blue)

	1. Raw POMP	2. Log Transform	3. Standardized Mean	4. Log POMP SD
Effect size (ES)	$\frac{M - M_0}{M_{max} - M_0}$	$\frac{1}{2} ln\left(\frac{M - M_{min}}{M_{max} - M}\right)$	$\frac{M - M_0}{SD} J$	$ln\left(\frac{SD}{M_{max} - M_0}\right) + \frac{1}{2(n-1)}$
Standard error	$\frac{SD}{\sqrt{n}} \frac{1}{M_{max} - M_0}$	$\frac{SD}{2\sqrt{n}} \frac{M_{max} - M_{min}}{(M - M_{min})(M_{max} - M)}$	$\sqrt{\frac{1}{n} + \left(1 - \frac{n-3}{(n-1)J^2}\right) ES^2}$	$\sqrt{\frac{1}{2(n-1)}}$
Analytic role	Main metric	Sensitivity test	Sensitivity test	Supplemental

Note. M = raw mean, M_0 = scale's midpoint value indicating gender-neutral beliefs, M_{max} = maximum possible value for conventional stereotypes (e.g., strongest possible pro-male STEM stereotypes), M_{min} = minimum possible value (e.g., strongest possible pro-female STEM stereotypes), SD = raw sample standard deviation, n = sample size, and $J = 1 - 3/(4n - 5)$ = small-sample bias-correction term for standardized means. These equations assume symmetric scales (i.e., distance from neutrality is the same for both scale endpoints), which is a requirement for inclusion in our meta-analysis. For Metric 2, the $1/2$ divider term is added so that the metric is approximately equal to Metric 1 for small stereotype magnitudes (i.e., M is close to M_0) based on a first-order linear approximation. The standard error formula for Metric 2 was derived using the delta method (e.g., see Cheung, 2015, p. 61-62; mathematical derivation for this specific formula is available upon request), and its approximate accuracy was verified based on simulations in R. The standard error formula for Metric 3 is based on the unbiased estimator of the variance for one-sample standardized means, assuming underlying normally distributed individual-level responses (Viechtbauer, 2007, Equation 26). Lastly, the metric for the POMP SD is a variance-stabilizing log transformation plus a small-sample bias-correction term, as recommended by Nakagawa et al. (2015).

A notable limitation, however, of both POMP scores and standardized means is that these metrics may still not completely control for methods confounds due to different scale types (e.g., two-alternative forced choice measures versus continuous analog scales; Johnson & Eagly, 2014, p. 691; Simms, Zelazny, Williams, & Bernstein, 2019). Hence, as detailed later in the Analysis Plan section, we will include dummy codes for critical scale type features (e.g., 2 or 3 discrete response options) in all moderator analyses.

Additional Technical Justification for Meta-Analysis 1 Metric: Because POMP effect sizes are simple linear transformations of raw means, the central limit theorem ensures that they will be distributed approximately normal across repeated samples of reasonable size (i.e., asymptotically normal), even if the underlying individual-level responses are discrete or otherwise non-normal. Simulations run in R confirmed that the standard error formula for the POMP metric is accurate within ~1-3% on average for even small samples (e.g., n = 20) with underlying bounded responses that are discrete, skewed, leptokurtic, platykurtic, or otherwise non-normal (e.g., on a 5-point scale). Hence, this robustness to non-normal response distributions is therefore one attractive technical property of the POMP metric, especially because we will frequently encounter discrete response scales in our meta-analysis.

In contrast, traditional standard error formulas for standardized mean metrics (of non-dichotomous outcomes) almost always assume individual-level continuous, normal distributions (e.g., Hedges & Olkin, 1985, Chapter 5; Viechtbauer, 2007). Although approaches have been developed for estimating the variance of standardized effects for non-normal response distributions (e.g., Chen & Peng, 2015; Kelley, 2005), these approaches require information not typically reported in primary studies (e.g., bootstrapping, kurtosis values), limiting their practical utility for meta-analysis. When computing the variance of standardized effects, the meta-analyst therefore is usually forced to assume the individual-level responses are normally distributed (as we do for Metric 3 in Table 6). In contrast, the variance formula for POMP means avoids this distributional assumption because of the generality of the central limit theorem.

An additional technical advantage is that POMP scores will have greater statistical power than standardized means (e.g., t-values for differences from 0 will be larger) because the "standardizing" denominator term is a known constant (based on features of the response scale) rather than a sample estimate with noise (the sample SD). Though the sample SD is still needed to estimate the variance of POMP means (or of any sample mean), the effect size itself does not depend on the SD, contributing to relatively more precise effect estimates compared to standardized means.

Sensitivity Analyses for Meta-Analysis 1 Metric: One concern about the raw POMP metric is that the scores are bounded from -1 to 1, which might cause moderator analysis models to possibly make out-of-bounds predictions. To address this concern, we plan to conduct sensitivity analyses with Metric 2 in Table 6, which is a log transformation of the raw POMP score that can range from -∞ to ∞ (note that for proportions of dichotomous 0/1 responses, this formula reduces to a standard logistic transformation; see Cheung, 2015, Equation 3.27; Lesaffre, Rizopoulos, & Tsonaka, 2007). We do not expect any major conclusions will be substantially different between Metrics 1 and 2, so we favor Metric 1 as the "primary" metric because it is simpler to communicate and interpret. However, we will note in the main manuscript if

any major results are substantially different (e.g., different in statistical significance), but we otherwise plan to present the more detailed results for Metric 2 in supplemental tables and appendices.

Likewise, we will consider standardized means (Metric 3) as another sensitivity test, but it may yield more divergent conclusions because of the confound with differences in variability (beyond those simply due to different numeric ranges). Hence, we consider POMP scoring to provide stronger and purer tests of our moderator hypotheses about differences in stereotype means.

Lastly, we will include the log POMP SD (Metric 4) in supplemental, exploratory analyses. We do not have strong a priori predictions for analyses of sample variability, but they could nevertheless provide novel theoretical insights on how the *distributions* of children's stereotypes vary (not just their means). In addition, by examining how the "standard" for Metric 3 varies, such analyses could help explain any potential discrepancies between results based on POMP scoring versus standardized means.

Meta-Analysis 2: Meta-Analysis 2 will focus on the correlations between STEM and verbal ability stereotypes with three focal categories of STEM motivational outcomes: (a) confidence (e.g., expectancies for success, self-efficacy, perceived ability), (b) interests and values (e.g., intrinsic value, enjoyment, utility value), and (c) future educational and career plans (e.g., intended college major).

We will convert bivariate correlation coefficients (e.g., Pearson's *r*) into Fisher's *z*-scores for meta-analytic models (see Table 7; Borenstein et al., 2009, Chapter 6). If multiple regression coefficients are reported instead, we will send an author query for the bivariate correlations. Though we currently do not plan to include partial correlations or regression coefficients in our analyses, we may decide to include them if we would lose too many studies (e.g., more than half) if we excluded them. We plan to make this decision after coding studies, but before analyzing the data, so that our decision is based on data availability, but not the results. If we do include partial correlations (e.g., converted from regression coefficients), we would include a dummy code for them in all moderator analyses. In addition, we would conduct sensitivity analyses to examine if any results depend on including this other category of correlations.

Table 7. Correlation Effect Metrics for Meta-Analysis 2

	Fisher's *z*
Effect size (ES)	$\frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$
Standard error	$\sqrt{\frac{1}{n-3}}$

Note. *r* = raw bivariate correlation coefficient, *n* = sample size. Fisher *z* scores are recommended as a variance-stabilizing transformation that also allow the effect sizes to be unbounded (Hedges & Olkin, 1985, Chapter 11). The standard error formula assumes normally distributed individual-level responses. Although this normality assumption is not ideal, our plan to use robust variation estimation in analyses will help adjust for any misspecification of the effect size variances. After analysis, we will back transform point values (e.g., means) to the more familiar Pearson's *r* metric using the formula $r = (e^{2z} - 1)/(e^{2z} + 1)$.

Analysis Plan

Analytic Details Common Across Meta-Analysis 1 and 2

Several details about the analytic models for Meta-Analysis 1 (variation in stereotypes) and Meta-Analysis 2 (variation in stereotype-outcome correlations) will be the same, as detailed in the following sections.

Meta-Analytic Models: We will use mixed-effects meta-regression models to investigate how stereotypes and their correlation with STEM outcomes vary across child demographics, cultural contexts, and measures. These models, which will be estimated using restricted maximum likelihood, will assume that variation in effect sizes is due to fixed effects of moderators (e.g., age), random effects of residual between-study heterogeneity, and within-study sampling variance (Borenstein et al., 2009). Because many studies likely will generate multiple effect sizes, we will implement robust variance estimation (RVE) to account for effect size dependencies, using the small-sample correction based on the Satterthwaite approximation (Tipton, 2015) and the "CR2" bias-reduced linearization adjustment (Pustejovky & Tipton, 2018; Tipton & Pustejovky, 2018).

Inference Criteria: We will determine two-tailed *p*-values for regression coefficients using *t*-tests based the Satterthwaite degrees of freedom and apply a default alpha cut-off level of .05 when determining statistical significance. However, we will lower the alpha level to .025 when the degrees of freedom for individual coefficients are less than 4 because of inflated Type I error rates in those scenarios (Tipton, 2015). In addition, if any moderator level has 4 effect sizes or less, we will combine that moderator level with others prior to analysis (e.g., group engineering and computer science stereotypes measures as one category if there are not enough effects for disaggregated analysis of them), if possible.

Because we have prespecified our confirmatory moderator analyses, we will not apply corrections for multiple comparisons to them. However, for exploratory analyses of potentially many more comparisons, we will use the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to control false discovery rates, rather than the Bonferroni method which can too stringent in many situations (see Appendix F in What Works Clearinghouse, 2017).

Software Implementation of Robust Variance Estimation (RVE): We will first estimate mixed-effects meta-analysis model parameters using the *rma.mv()* function in the *metafor* R package (Viechtbauer, 2010) and then adjust the standard errors, degrees of freedom, and *p* values by implementing RVE using the *coef_test()* function in the *clubSandwich* package (Pustejovsky, 2018). Contrasting with the *robumeta* R package implementation of RVE, this combined "*metafor* + *clubSandwich*" implementation allows for greater flexibility in specifying the effect size covariance structure used to determine the model weights. Example code for this implementation is provided below.

Different Types of Effect Size Dependencies: Dependent effects can occur from both hierarchical, multilevel dependence structures (e.g., subsamples nested within studies) and correlated, multivariate structures (e.g., multiple measures for the same sample). Fortunately, "*metafor* + *clubSandwich*" implementation of RVE, unlike the *robumeta* implementation, allows for both types of dependence structures to be specified simultaneously when determining model weights. The example code below shows this specification in the following way:

1. **The nesting of effect sizes within samples is treated as correlated effects:** The *impute_covariance_matrix()* function imputes an assumed within-study correlation *r* to the effect size variance-covariance matrix based on matching sample ID's. We will assume a default $r = .5$ but will also test the sensitivity of our results to other values ($r = .2$ and $r = .8$). We will report when any major results differ depending on this parameter.

2. **The nesting of samples within studies is treated as hierarchical effects:** *The rma.mv()* function specifies the nesting of samples within studies (and effects within samples) through the random effects model specification.

3. **RVE is used to adjust standard errors based on study-level clustering:** The coef_test() from the clubSandwich package then adjusts the standard errors based on overall (study-level) clustering to ensure they are robust to deviations from the assumed dependence structure.

```
# This function was based on code that Beth Tipton (one of RVE's original co-creators)
provided at 2018 IES Meta-Analysis Training Institute on Day 4 (August 8, 2018). She
recommended this approach over the "old" robuMeta implementation because the model
weights will be more appropriate (she's working on a paper to explain this)

#Implement RVE using clubSandwich to adjust SEs
# Input - data: data frame of effect sizes nested within samples and studies
# Input - r: assumed correlation between items (0.5 is the default)
# Input - ...: other arguments to pass to rma.mv() such as moderators
# Output: output of coef_test from the clubSandwich package
rve = function(data, r = 0.5, ...) {

  #create covariance matrix based on known variances and assumed correlation
  #only treat effects nested within the same sample as correlated (multivariate, r=r)
  #samples nested within studies should be treated as hierarchical (multilevel, r=0)
  covM = impute_covariance_matrix(vi = data$vi, cluster = data$SampleID, r = r)

  #run multivariate regression model, accounting for effects nested within samples,
  #nested within studies
  m = rma.mv(yi, covM, random = ~1 | StudyID/SampleID/EffectID, data = data, ...)

  #adjust SEs using RVE based on the study-level clustering
  #Beth Tipton said CR2 estimation method is best
  coefs = coef_test(m, cluster = data$StudyID, vcov = "CR2")

  #output results as data frame
  data.frame(coef = rownames(coefs), coefs)
}
```

Addressing Missing Moderator Data: We will use multiple imputation to account for any missing moderator data (e.g., for racial/ethnic composition) that may remain missing even after attempting to contact study authors. Compared to common listwise deletion practices, multiple imputation preserves a larger set of studies and can often produce less biased estimates (Pigott, 2001, 2012). We will implement a joint modelling approach by using the *jomo* R package to also account for the multilevel structure of the data (i.e., nesting within studies) when imputing moderator values (Quartagno, Grund, & Carpenter, in press). We will allow the level-1 covariance structures to randomly vary across studies by setting the "meth" option in the *jomo()* function equal to "random." As a simplifying assumption, the imputation model will account for just the overall 2-level nesting structure (i.e., effects within studies). The effect sizes, but not their variances, will be included in the imputation models because omitting the outcome (i.e., effect sizes) when imputing covariates (i.e., moderators) can introduce bias due to a lack of correspondence between the imputation model and analysis model (e.g., Moons, Donders, Stijnen, & Harrell, 2006).

Data Exclusion and Outlier Detection: We will test the robustness of our findings to potential outliers by using the *rstudent()* function in the *metafor* package to identify effect sizes that have studentized deleted residuals exceeding 2.5 standard errors (Viechtbauer & Cheung, 2010). We will rerun meta-analytic models with and without such outliers, reporting when the conclusions are sensitive to such exclusions.

Selective Reporting Bias: Although we will use several systematic search methods to find unpublished studies, our conclusions may nevertheless be subject to selective reporting bias (e.g., publication bias or outcome reporting bias). Researchers may be cautious to report results conflicting with commonly accepted theory and research hypotheses (Flore & Wicherts, 2015). As detailed in Appendix D, we will use three approaches as sensitivity analyses to diagnose and adjust for such biases: (a) selection modeling, (b) comparison of unpublished versus published studies, and (c) meta-regression to assess small-study effects. If the three approaches yield diverging conclusions, we will place the greatest weight on selection models for reasons detailed in Appendix D. Simulation studies have shown superior performance for selection models, compared to other publication bias methods, in several conditions that may likely characterize our meta-analysis such as moderate to large between-study heterogeneity (Carter, Schönbrodt, Gervais, & Hilgard, 2019).

Meta-Analysis 1: Investigating Variation in Children's Stereotypes

Overall Meta-Analytic Averages and Heterogeneity: We will characterize the overall magnitude of STEM and verbal ability stereotypes using simple random-effects models of the POMP scores (Metric 1 in Table 6) and standardized means (Metric 3). These two metrics provide different ways of characterizing the average magnitude of stereotypes, though we favor the POMP scores for our moderator analyses, as noted earlier. Between-study heterogeneity will be quantified by presenting 90% prediction intervals, a measure of the estimated dispersion of true underlying effects. In contrast to some other heterogeneity metrics like I^2 statistics (percentage of total variation in effect size due to heterogeneity rather than chance), prediction intervals provide a direct, absolute measure of heterogeneity using the original units of the effect size metric (Borenstein, Higgins, Hedges, & Rothstein, 2017).

Planned Confirmatory Moderator Analyses: Confirmatory moderator analyses will test our directional hypotheses such as pro-male STEM ability stereotypes should strengthen with age. We will examine each moderator in separate models and in one multivariable model that will simultaneously adjust for all confirmatory moderators. To control for nuisance methods variance, all models will include four dummy-coded covariates for the scale type: (a) 2 or 3 response options, (b) forced-choice scale with no gender-neutral response option for individual items, (c) continuous scale (vs. discrete ratings), and (d) separate ratings (vs. comparative scales). Though effect sizes will be standardized, the original measurement scale still could contribute additional variance that should be adjusted for. STEM ability and verbal ability stereotypes will be analyzed in separate statistical models, except when directing comparing their overall magnitude (in that case, both will be included in the same model to account for their correlation).

Table 8 shows a concrete example of how we plan to report this information. We will report moderator results for the POMP scores (Metric 1) in the main manuscript and results for standardized means (Metric 3) in supplemental materials for reasons discussed earlier in the "Variables" section.

Table 8. Tests of Confirmatory Hypotheses for Meta-Analysis 1

	Simple				Multivariable			
	b	*SE*	*df*	*p*	*b*	*SE*	*df*	*p*
Pro-Male STEM Stereotypes								
Children's Age	+				+			
Proportion Male	+				+			
Age × Male[a]	-				-			
Proportion Black, Non-Hisp.[b] (U.S. only[c])	-				-			
Data Collection Year (U.S. only[c])	-				-			
National Prop. Female STEM Majors[d]	-				-			
National Prop. Female Researchers[d]	-				-			
Indirect Measure (vs. Direct Measure)	+				+			
Adult Targets (vs. Child Targets)	+				+			

Pro-Female Verbal Stereotypes[e]		
Children's Age	+	+
Proportion Female	+	+
Age × Female[a]	-	-
Cross-Domain Magnitude Comparison[e]		
Verbal (vs. STEM) Ability Stereotype	+	+

Note. The simple models (left-hand side) will test each confirmatory moderator one-by-one in separate models, but control for scale type as a potential methodological confounder. Scale type will be entered as four dummy codes for (a) 2 or 3 discrete response options, (b) forced-choice scale with no gender-neutral option, (c) continuous scale, and (d) separate ratings (vs. comparative scale). The multivariable models (right-hand side) will control for all confirmatory moderators and scale type simultaneously. Positive and negative signs indicate prediction directions.

[a] Age and gender will be grand mean centered so that the age and gender effect in the multivariable models can be interpreted as overall average effects.

[b] When testing for racial/ethnic differences, the omitted reference category will be White, non-Hispanic children. Confirmatory models will include two racial/ethnic moderators: (a) proportion Black (non-Hispanic) and (b) proportion of children who were neither mono-racial Black nor White (but Hispanic, Asian, multi-racial, etc. instead). Only the first moderator (proportion Black) will be reported here because our hypothesis centered on Black-White differences.

[c] Results for these moderators will come from models restricted to U.S. samples because the relevant hypotheses were U.S.-specific.

[d] These national-level moderators will be omitted in the U.S.-specific multivariable models.

[e] We will analyze STEM and verbal ability stereotypes in separate statistical models, except when testing the bottom hypothesis about overall magnitude difference; in that one case, we will include the stereotype domain as a dummy code (1 = verbal; 0 = STEM) in a combined model.

Additional Analysis of Children's Age as a Moderator: We will elaborate on age-related analyses given their central theoretical importance. First, we will determine the youngest age when STEM ability stereotypes significantly favor male targets and, separately, when verbal ability stereotypes significantly favor female targets. Second, as a robustness check of age-related effects, we will separately examine between-study comparisons (i.e., studies with older vs. younger samples) and within-study comparisons (i.e., older vs. younger children in the same study); we will do so by group-mean centering age for studies with multiple age groups (Tanner-Smith, Tipton, & Polanin, 2016).

Within-study comparisons are especially valuable because they control for between-study features such as geographic location and scale type. However, we anticipate that most of the variation in children's age may lie between (rather than within) studies, meaning that within-study age comparisons may be underpowered. For this reason, we consider the group-mean centering approach to be a robustness check, rather than a confirmatory model specification. Our confirmatory models (e.g., what we would report in Table 8) would not apply group-mean centering, meaning that the regression coefficient for age could be interpreted as a weighted average of within-study and between-study effects in that case.

Exploratory Analyses: Exploratory analyses will examine additional moderators such as question wording and stereotype domain (e.g., math vs. spatial ability; Table 2) that may yield further insight. We also will explore interactions with age to advance understanding of possible developmental mechanisms (e.g., indirect and direct measures might show different age-related trends, especially if they capture distinct theoretical constructs). Last, we will repeat analyses separated by children's gender to see if stereotypes vary in consistent ways for boys and girls. These analyses could help build fundamental theory about stereotype development and suggest promising novel directions for future research. However, we will interpret these exploratory analyses cautiously, given possible inflation of false-positive error rates.

Exploring Measure Quality: We will also use the coded methodological quality indicators (e.g., stereotype reliability; see Table 3) to empirically examine how the various psychometric properties of the

measures relate to differences in effect sizes across studies. As exploratory analyses, we will conduct robustness checks to examine if our central results remain when restricted to different combinations of measure quality thresholds (e.g., multi-item scales with an internal consistency greater than .70).

Confidence in Cumulative Estimate: Though our research questions largely center on variation in stereotypes, interpreting the overall mean estimate is important as well (e.g., broadly speaking, do STEM ability stereotypes favor males?). A random-effects meta-analysis without any moderators can provide a simple mean estimate, but we will keep several factors in mind when interpreting that mean:

- ***Typical Sample Age:*** If most studies included younger children, then the overall stereotype mean might be smaller than otherwise expected. Extrapolating to 18 years of age would be important to understand the magnitude of stereotypes as students enter adulthood.
- ***Typical Stereotype Measure:*** As one example, if nearly all measures were direct and "explicit," then social desirability might downwardly bias mean estimates.
- ***Typical Data Collection Year:*** If stereotypes have weakened over time, then the overall mean estimate might be higher than expected for stereotypes in recent years.
- ***Selective Reporting Bias:*** The overall mean estimate might be higher than its true value if results favoring accepted theory (e.g., that STEM ability stereotypes disadvantage girls) are more likely to be reported and published.

Hence, as these considerations show, moderator analyses and sensitivity analyses (e.g., regarding selective reporting bias) are important even when interpreting a "simple" mean estimate.

Meta-Analysis 2: Examining Relations With Key STEM Outcomes

Because many analytic details will match those in Meta-Analysis 1 (e.g., plans for robust variance estimation, publication bias analyses, multiple imputation), we detail below only the differences in analytic plans for Meta-Analysis 2.

Overall Meta-Analytic Averages and Heterogeneity: Our core prediction for Meta-Analysis 2 is that pro-male STEM and pro-male female ability stereotypes should relate to motivational STEM outcomes negatively for girls and positively for boys. We will test these basic predictions with four separate random-effects models for the four broad categories of correlations (see Table 9), along with characterizing their overall heterogeneity. Two additional mixed-effects models (bottom rows of Table 9) will directly examine average differences between boys' and girls' correlations by including one moderator for gender.

Table 9. Average Correlations With STEM Motivational Outcomes

	Sample			Mean				Heterogeneity	
	m	*k*	*n*	*b*	*SE*	*df*	*p*	τ	90% PI
Girls									
Pro-Male STEM Stereotypes				-					
Pro-Female Verbal Stereotypes				-					
Boys									
Pro-Male STEM Stereotypes				+					
Pro-Female Verbal Stereotypes				+					
Boys – Girls (Difference)									
Pro-Male STEM Stereotypes				+					
Pro-Female Verbal Stereotypes				+					

Note. *m* = number of studies, *k* = number of effect sizes, *n* = number of study participants, *b* = average correlation, *SE* = standard error of average correlation, *df* = RVE-adjusted Satterthwaite degrees of freedom, *p* = significance level for difference from 0, τ = estimated total effect heterogeneity, 90% PI = estimated middle 90% of true effects.

Results are based on four separate random-effects meta-analytic models. Positive and negative signs indicate predicted directions. Heterogeneity statistics for the boy-girl differences (bottom two rows) are omitted because the reported regression coefficients are drawn from the fixed-effects part of mixed-effects models.

Planned Confirmatory Moderator Analyses: For the correlations between girls' STEM ability stereotypes and STEM motivational outcomes (i.e., top row of Table 9), we will conduct additional moderator analyses using mixed-effects models (see Table 10). We focus on this category of correlations in part because we expect to obtain the most primary data for it (relative to other categories).

Table 10. Tests of Confirmatory Hypotheses for Meta-Analysis 2 (Correlations Between Girls' STEM Ability Stereotypes and STEM Motivational Outcomes)

	Simple				Multivariable			
Moderator	*b*	*SE*	*df*	*p*	*b*	*SE*	*df*	*p*
Children's Age	+				+			
Indirect Measure (vs. Direct Measure)	+				+			
Adult Targets (vs. Child Targets)	-				-			
Confidence (vs. Other Motivational Outcomes)	+				+			

Note. The simple models (left-hand side) tested each confirmatory moderator one-by-one in separate models but controlled for scale type as a potential methodological confounder. Scale type was entered as four dummy codes for (a) 2 or 3 discrete response options, (b) forced-choice scale with no gender-neutral option, (c) continuous scale, and (d) separate ratings (vs. comparative scale). The multivariable models (right-hand side) controlled for all confirmatory moderators and scale type simultaneously. Positive and negative signs indicate predicted directions.

Exploratory Analyses: Exploratory analyses will extend moderator analyses to the other effect size categories (e.g., correlations for boys), though the statistical power may be lower. We also will explore other moderators (e.g., socioeconomic status, stereotype or outcome STEM domain) and correlations with two categories of performance outcomes (i.e., test scores, grades).

References

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12(5), 385–390. https://doi.org/10.1111/1467-9280.00371

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608x377117

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*(4), 406–418. https://doi.org/10.1037/1082-989x.8.4.406

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115-144. https://doi.org/10.1177/2515245919847196

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods*, *22*(1), 28–41. https://doi.org/10.1037/met0000119

Chen, L.-T., & Peng, C.-Y. J. (2014). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavior Research Methods*, *47*(1), 107–126. https://doi.org/10.3758/s13428-014-0461-3

Cheung, M. W. L. (2015). *Meta-analysis: A structural equation modeling approach*. John Wiley & Sons.

Coburn, K. M. (2018). A weight-function model for moderators of publication bias [Doctoral dissertation]. Retrieved from https://escholarship.org/uc/item/3t6993k2

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, *20*(3), 310–330. https://doi.org/10.1037/met0000046

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. https://doi.org/10.1207/s15327906mbr3403_2

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, *145*(6), 610–651. https://doi.org/10.1037/bul0000194

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Del Re, A. C., & Hoyt, W. T. (2014). MAd: Meta-analysis with mean differences. R Package Version 0.8–2. Available at http://cran.r-project .org/web/packages/Mad

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Evans, A. B., Copping, K. E., Rowley, S. J., & Kurtz-Costes, B. (2011). Academic self-concept in Black adolescents: Do race and gender stereotypes matter? *Self and Identity, 10,* 263–277. https://doi.org/10.1080/15298868.2010.485358

Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman mathematics attitudes scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal for Research in Mathematics Education, 7*(5), 324–326. https://doi.org/10.2307/748467

Fischer, R., & Boer, D. (2011). What is more important for national well-being: Money or autonomy? A meta-analysis of well-being, burnout, and anxiety across 63 societies. *Journal of Personality and Social Psychology*, *101*(1), 164–184. https://doi.org/10.1037/a0023663

Fischer, R., & Chalmers, A. (2008). Is optimism universal? A meta-analytical investigation of optimism levels across 22 nations. *Personality and Individual Differences*, *45*(5), 378–382. https://doi.org/10.1016/j.paid.2008.05.008

Flore, P. C., Mulder, J., & Wicherts, J. M. (in press). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: a registered report. *Comprehensive Results in Social Psychology*. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53,* 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Forgasz, H. J., Leder, G. C., & Kloosterman, P. (2004). New perspectives on the gender stereotyping of mathematics. *Mathematical Thinking and Learning*, *6*(4), 389-420. https://doi.org/10.1207/s15327833mtl0604_2

Friese, M., Frankenbach, J., Job, V., & Loschelder, D. D. (2017). Does self-control training improve self-control? A meta-analysis. *Perspectives on Psychological Science*, *12*(6), 1077–1099. https://doi.org/10.1177/1745691617697076

Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development, 85,* 250–263. https://doi.org/10.1111/cdev.1212

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106–137. https://doi.org/10.3102/0091732x19830634

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, *10*(10), 1645–1654. https://doi.org/10.1111/2041-210x.13268

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS ONE, 10*(9), e0138237. https://doi.org/10.1371/journal.pone.0138237

Hargreaves, M., Homer, M., & Swinnerton, B. (2008). A comparison of performance and attitudes in mathematics amongst the 'gifted'. Are boys better at mathematics or do they just think they are? *Assessment in Education: Principles, Policy & Practice, 15,* 19–38. https://doi.org/10.1080/09695940701876037

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 203–217. https://doi.org/10.1037/1082-989X.6.3.203

Heyman, G. D., & Legare, C. H. (2004). Children's beliefs about gender differences in the academic and social domains. *Sex Roles, 50,* 227–239. https://doi.org/10.1023/B:SERS.0000015554.12336.30

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of social-personality psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd Ed., pp. 675-707). London: Cambridge University Press.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*(1), 51–69. https://doi.org/10.1177/0013164404264850

Kurtz-Costes, B., Copping, K. E., Rowley, S. J., & Kinlaw, C. R. (2014). Gender and age differences in awareness and endorsement of gender stereotypes about academic abilities. *European Journal of Psychology of Education*, *29*(4), 603–618. https://doi.org/10.1007/s10212-014-0216-7

Lesaffre, E., Rizopoulos, D., & Tsonaka, R. (2006). The logistic transform for bounded outcome scores. *Biostatistics*, *8*(1), 72–85. https://doi.org/10.1093/biostatistics/kxj034

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*(4), 641–654. https://doi.org/10.1002/sim.698

Martinot, D., & Désert, M. (2007). Awareness of a gender stereotype, personal beliefs and self-perceptions regarding math ability: When boys do not surpass girls. *Social Psychology of Education*, *10*(4), 455–471. https://doi.org/10.1007/s11218-007-9028-9

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences, 18,* 37–45. https://doi.org/10.1016/j.tics.2013.10.011

Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of five decades of U.S. Draw-A-Scientist studies. *Child Development*, *89*, 1943–1955. https://doi.org/10.1111/cdev.13039

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., … Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1). https://doi.org/10.1186/2046-4053-4-1

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell, F. E., Jr. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, *59*(10), 1092–1101. https://doi.org/10.1016/j.jclinepi.2006.01.009

Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*(1). https://doi.org/10.1186/1471-2288-9-2

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2014). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, *6*(2), 143–152. https://doi.org/10.1111/2041-210x.12309

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation & the Health Professions*, *24*(3), 277–307. https://doi.org/10.1177/01632780122034920

Pigott, T. D. (2012). Missing data in meta-analysis: Strategies and approaches. In *Advances in meta-analysis* (pp. 79–107). Boston, MA: Springer. https://doi.org/10.1007/978-1-4614-2278-5_7

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432. https://doi.org/10.3102/0013189x13507104

Plante, I., Théorêt, M., & Favreau, O. E. (2009). Student gender stereotypes: Contrasting the perceived maleness and femaleness of mathematics and language. *Educational Psychology, 29*(4), 385–405. https://doi.org/10.1080/01443410902971500

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes. *Review of Educational Researc*h, *86*(1), 207–236. https://doi.org/10.3102/0034654315582067

Polanin, J. R., & Terzian, M. (2019). A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. *Journal of Clinical Epidemiology, 106,* 60–69. https://doi.org/10.1016/j.jclinepi.2018.10.006

Polanin, J. R., & Williams, R. T. (2016). Overcoming obstacles in obtaining individual participant data for meta-analysis. *Research Synthesis Methods, 7*(3), 333–341. https://doi.org/10.1002/jrsm.1208

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses. *Psychological Science*, *27*(7), 1036–1042. https://doi.org/10.1177/0956797616645672

Pustejovsky, J. (2018). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.3.2). Retrieved from https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Quartagno, M., Grund, S., & Carpenter, J. (in press). jomo: A flexible package for two-level joint modelling multiple imputation. *The R Journal*. Retrieved from https://discovery.ucl.ac.uk/id/eprint/10078316/1/RJwrapper.pdf

Rathbone, J., Hoffmann, T., & Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic Reviews, 4*(1), 80. https://doi.org/10.1186/s13643-015-0067-6

Renkewitz, F., & Keiner, M. (2018, December 20). *How to detect publication bias in psychological research? A comparative evaluation of six statistical methods*. https://doi.org/10.31234/osf.io/w94ep

Rowley, S. J., Kurtz-Costes, B., Mistry, R., & Feagans, L. (2007). Social status as a predictor of race and gender stereotypes in late childhood and early adolescence. *Social Development*, *16*, 150–168. https://doi.org/10.1111/j.1467-9507.2007.00376.x

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J., & Nelson, L. (2018, January 8). *P-curve handles heterogeneity just fine* [Blog post]. Retrieved from http://datacolada.org/67

Simonsohn, U., Simmons, J., & Nelson, L. (2017, June 15). *Why p-curve excludes ps>.05* [Blog post]. Retrieved from http://datacolada.org/61

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, *8*(5), 581–591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). New York, NY: Wiley.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/bf02294384

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, *32*(1), 39–60. https://doi.org/10.3102/1076998606298034

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

What Works Clearinghouse. (2017). *Procedures handbook (Version 4.0)*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Appendix A: Keyword Literature Searches

To enhance reproducibility, this appendix includes the exact search strings for keyword searches. These strings were generated in R to avoid errors and specify the desired searchable fields for each database and group of keyword terms (see 00_keyword_search.R at https://osf.io/a3prx/).

Web of Science Core Collection

The following string generated 1,650 hits on November 8, 2019 when pasted into the "Topic" field that simultaneously searches the title, abstract, author keywords, and Keywords Plus on Web of Science Core Collection accessed at https://www.webofknowledge.com:

("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") AND ("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") AND ("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") AND ("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR "elementary school*" OR "elementary education" OR "middle school*" OR "high school*" OR "highschool*" OR "junior high" OR "primary school*" OR "primary education" OR "secondary school*" OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*")

Scopus

The following string generated 2,333 hits on November 8, 2019 when pasted into the "Advanced" search option on Scopus accessed at https://www.scopus.com/:

(TITLE-ABS("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic

abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR AUTHKEY("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*")) AND (TITLE-ABS("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") OR AUTHKEY("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*")) AND (TITLE-ABS("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR AUTHKEY("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*")) AND (TITLE-ABS-KEY("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR "elementary school*" OR "elementary

education" OR "middle school*" OR "high school*" OR "highschool*" OR "junior high" OR "primary school*" OR "primary education" OR "secondary school*" OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*"))

EBSCOhost Databases

The following string generated these number of hits on November 8, 2019 when pasted into the "Select A Field (optional)" field for these EBSCOhost databases accessed through http://search.ebscohost.com: 1,094 for Academic Search Complete, 603 for Education Full Text, 865 for Education Research Complete, 1,122 for Education Source, 1,395 for ERIC, 444 for SocINDEX with Full Text, and 230 for Social Sciences Full Text.

(TI("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR AB("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR KW("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*")) AND (TI("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") OR AB("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") OR KW("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*")) AND (TI("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR AB("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR

"male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR KW("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*")) AND (("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR "elementary school*" OR "elementary education" OR "middle school*" OR "high school*" OR "highschool*" OR "junior high" OR "primary school*" OR "primary education" OR "secondary school*" OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*"))

PsycINFO

The following string generated 2,115 hits on November 8, 2019 when pasted into the “Select A Field (optional)” field on PsycINFO. Although we searched PsycINFO also using EBSCOhost (http://search.ebscohost.com), the default search fields (when not explicitly specified) does not include the “Age Group” (AG) field; hence, the AG field was explicitly added to the fourth group of terms for the age search terms. In addition, we also added the “Subjects” (SU) field to the first and third group of terms

to take advantage of PsycINFO's controlled vocabulary (e.g., "stereotyp*" would index the subject term "stereotype attitudes" that trained human coders selected as relevant).

(TI("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR AB("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR KW("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*") OR SU("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR "verbal abilit*" OR "verbal achievement" OR "verbal performance" OR "academic domain*" OR "academic abilit*" OR "academic achievement" OR "academic performance" OR "cognitive abilit*" OR "cognitive achievement" OR "cognitive performance" OR "intellectual abilit*" OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*")) AND (TI("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") OR AB("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*") OR KW("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*")) AND (TI("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR AB("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR

"percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR KW("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") OR SU("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie*

about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*")) AND (("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR "elementary school*" OR "elementary education" OR "middle school*" OR "high school*" OR "highschool*" OR "junior high" OR "primary school*" OR "primary education" OR "secondary school*" OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*") OR AG("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR "elementary school*" OR "elementary education" OR "middle school*" OR "high school*" OR "highschool*" OR "junior high" OR "primary school*" OR "primary education" OR "secondary school*" OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*"))

ProQuest Databases

The following string generated these number of hits on November 8, 2019 when pasted into the "Enter search terms…" field for these ProQuest databases accessed through https://search.proquest.com: 473 for GenderWatch, and 1,283 for ProQuest Theses & Dissertations Global. This search syntax used the "anywhere but the full text" (NOFT) field that indexes the title, abstract, authors, publication title, and subjects and indexing terms.

NOFT(("math*" OR "science" OR "technol*" OR "engineering" OR "comput*" OR "STEM" OR "spatial*" OR "mental rotation" OR "quant* abilit*" OR "quant* achievement" OR "quant* performance" OR ("verbal abilities" OR "verbal ability") OR "verbal achievement" OR "verbal performance" OR ("academic domain") OR ("academic abilities" OR "academic ability") OR "academic achievement" OR "academic performance" OR ("cognitive abilities" OR "cognitive ability") OR "cognitive achievement" OR "cognitive performance" OR ("intellectual abilities" OR "intellectual ability") OR "intellectual achievement" OR "intellectual performance" OR "reading" OR "writing" OR "intelligen*")) AND NOFT(("gender*" OR "sex" OR "boy*" OR "girl*" OR "female*" OR "male*" OR "women*" OR "son*" OR "daughter*")) AND NOFT("stereotyp*" OR "gender* perception*" OR "gender* belief*" OR "gender* bias" OR "male domain" OR "female domain" OR "belie* gender*" OR "percei* gender*" OR "percep* gender*" OR "belie* about gender*" OR "percei* about gender*" OR "percep* about gender*" OR "belie* that gender*" OR "percei* that gender*" OR "percep* that gender*" OR "belie* sex" OR "percei* sex" OR "percep* sex" OR "belie* about sex" OR "percei* about sex" OR "percep* about sex" OR "belie* that sex" OR "percei* that sex" OR "percep* that sex" OR "belie* boy*" OR "percei* boy*" OR "percep* boy*" OR "belie* about boy*" OR "percei* about boy*" OR "percep* about boy*" OR "belie* that boy*" OR "percei* that boy*" OR "percep* that boy*" OR "belie* girl*" OR "percei* girl*" OR "percep* girl*" OR "belie* about girl*" OR "percei* about girl*" OR "percep* about girl*" OR "belie* that girl*" OR "percei* that girl*" OR "percep* that girl*" OR "belie* female*" OR "percei* female*" OR "percep* female*" OR "belie* about female*" OR "percei* about female*" OR "percep* about female*" OR "belie* that female*" OR "percei* that female*" OR "percep* that female*" OR "belie* male*" OR "percei* male*" OR "percep* male*" OR "belie* about male*" OR "percei* about male*" OR "percep* about male*" OR "belie* that male*" OR "percei* that male*" OR "percep* that male*" OR "belie* women*" OR "percei* women*" OR "percep* women*" OR "belie* about women*" OR "percei* about women*" OR "percep* about women*" OR "belie* that women*" OR "percei* that women*" OR "percep* that women*" OR "belie* son*" OR "percei* son*" OR "percep* son*" OR "belie* about son*" OR "percei* about son*" OR "percep* about son*" OR "belie* that son*" OR "percei* that son*" OR "percep* that son*" OR "belie* daughter*" OR "percei* daughter*" OR "percep* daughter*" OR "belie* about daughter*" OR "percei* about daughter*" OR "percep* about daughter*" OR "belie* that daughter*" OR "percei* that daughter*" OR "percep* that daughter*") AND NOFT((("child*" OR "adolescen*" OR "boy*" OR "girl*" OR "grade*" OR "preschool*" OR "pre-school*" OR "pre-kindergart*" OR "prekindergart*" OR "kindergart*" OR ("elementary school" OR "elementary schoolchildren" OR "elementary schoolhome" OR "elementary schooling" OR "elementary schoolk" OR "elementary schools" OR "elementary schoolteacher") OR "elementary education" OR ("middle school" OR "middle schooler" OR "middle

schoolers" OR "middle schooling" OR "middle schools") OR ("high school" OR "high schooler" OR "high schoolers" OR "high schoolin" OR "high schooling" OR "high schools" OR "high schoolthe") OR "highschool*" OR "junior high" OR ("primary school" OR "primary schooling" OR "primary schools" OR "primary schoolteacher") OR "primary education" OR ("secondary school" OR "secondary schooling" OR "secondary schools") OR "secondary education" OR "elementary secondary" OR "youth*" OR "teen*" OR "K12*" OR "K-12*" OR "PK12*" OR "PK-12*" OR "school-age*")))

Appendix B: Screening Manual

The abstracts you will screen will have color-coded highlighting for the same keywords that were used in the literature database searches to first find these citations (Table 1).

Table 1. Search Terms for Keyword Searches (Asterisks Denote Wildcard Characters)

Domain	Gender	Stereotype	Age
math*	gender*	stereotyp*	child*
science	sex	gender* perception*	adolescen*
technol*	boy*	gender* belief*	boy*
engineering	girl*	gender* bias	girl*
comput*	female*	male domain	grade*
STEM	male*	female domain	preschool*
spatial*	women*	belie* about [gender term]	pre-school*
mental rotation	men*	belie* that [gender term]	pre-kindergart*
quant* abilit*	son*	belie* [gender term]	prekindergart*
quant* achievement	daughter*	percei* about [gender term]	kindergart*
quant* performance		percei* that [gender term]	elementary school*
verbal abilit*		percei* [gender term]	elementary education
verbal achievement		percep* about [gender term]	middle school*
verbal performance		percep* that [gender term]	high school*
academic domain*		percep* [gender term]	highschool*
academic abilit*			junior high
academic achievement			primary school*
academic performance			primary education
cognitive abilit*			secondary school*
cognitive achievement			secondary education
cognitive performance			elementary secondary
intellectual abilit*			youth*
intellectual achievement			teen*
intellectual performance			K12*
reading			K-12*
writing			PK12*
intelligen*			PK-12*
			school-age*

GENDER DIFFERENTIAL ITEM FUNCTIONING IN SLOVAK VERSION OF INTELLIGENCE STRUCTURE TEST 2000-REVISED

Journal: Studia Psychologica (Database: Web of Science)

Authors: Kohut Michal; Halama Peter; Dockal Vladimir; Zitny Peter

The study focused on the gender differential item functioning in Slovak version of the Intelligence Structure Test 2000 – Revised (Amthauer et al., 2011). The sample included 744 middle and high school students with mean age of 16.94 years. The non-parametric method SIBTEST for identification of items with differential functioning was used in order to detect uniform and non-uniform DIF. The analysis showed that the I-S-T 2000 R includes several items with DIF favoring either males or females, but in most subtests, with no or small effect on differences between genders. Substantial but nonsignificant effect of DIF items on subtest score was found for Verbal Analogy, which contained six items with DIF all favoring females. These items included verbal content related to areas more common for females such as diet or food. The results suggest that specific content of verbal intelligence items can be a potential source of gender bias.

keywords: intelligence testing; differential item functioning; gender; I-S-T 2000 R

Link to Full Text

Search Google Scholar

Search Google

Literature Database Page

Stage 1: Abstract Screening Questions

Answer each question with *yes*, *unsure*, or *no*. If the answer is *yes* or *unsure* to all questions, then approve the study in Abstrackr, which passes it to full-text review. If the answer is a definite *no* to one or more questions, then exclude. The below is a suggested mental order, but you may exclude from further review if you notice any answer is a definite no, even out of order. When you are unsure, include the study; the worst possible error at this stage is a false exclusion.

As shown in the previous screenshot, some studies will have links to the full text (searching in Google Scholar also often shows a link to a full text PDF). If you are unsure about any of these questions from the abstract, you may quickly skim the full text (especially its methods section) to see, but do not do this if you cannot immediately find the full text. In general, you should base your decision on the abstract, but having it readily available could help check your assumptions about interpreting the abstract (e.g., you think a particular term is likely irrelevant but you are not 100% sure). Please also keep efficiency in mind. Any study that passes this initial stage will be re-reviewed more carefully in full-text review (Stage 2). If you find yourself spending more than ~1-2 minutes looking over the full text, then you should likely include in Abstrackr and move on.

1. **Age: Did the study include a sample of children or adolescents (18 years or younger)?**
    a. Include any grade level before college (PreK-12).
    b. Include all-female, all-male, or mixed-sex samples.
    c. Exclude samples of only college students, teachers, parents, or adults.
    d. Exclude analyses of only cultural artifacts (e.g., children's textbooks).

2. **Gender Stereotypes: Did the abstract mention gender stereotypes were studied?**
    a. Include beliefs about gender differences, gendered beliefs, perceptions of boys as good at math, etc. ("stereotype" does not need be explicitly mentioned).
    b. Exclude stereotypes about other groups (e.g., racial but not gender stereotypes).
    c. Exclude studies about actual gender differences (e.g., gender gaps in math achievement) but not beliefs or perceptions about those differences.
    d. Exclude studies only about self-perceptions of abilities (e.g., "How good do you think you are at math?") but not about perceptions of gender differences.

3. **Stereotype Domain: Were the stereotypes about academic or cognitive domains?**
    a. Include gender stereotypes about general or specific academic/cognitive domains at this initial screening stage (we want to later flag the domain-general ones).
    b. Include if the abstract mentions a stereotype threat study because the study may have also measured ability stereotypes (but you may exclude if a skim of the full text clearly indicates that this is not the case).
    c. Include if the abstract mentions cultural fit stereotypes (e.g., math-gender Implicit Association Test), but you may also exclude here with a skim of the methods section.
    d. Exclude gender stereotypes about other domains (e.g., being nice, agentic).

4. **Sample Size: Did the sample include at least 10 children or adolescents in total?**
    a. Include if the numbers of boys and girls sum to 10 or more.

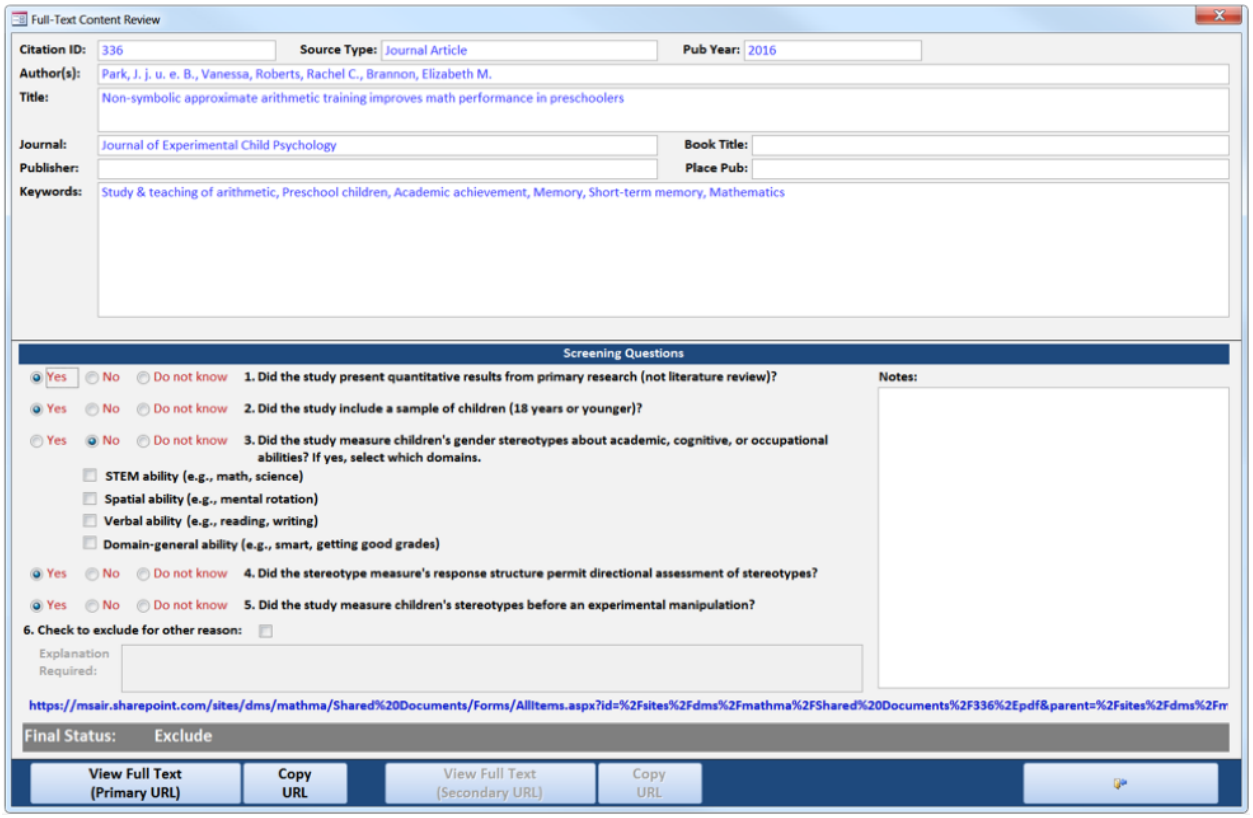
5. **Primary Research: Did the article report findings from primary research?**
    a. Exclude systematic reviews, meta-analyses, or other literature reviews.

Common Reasons for Exclusion at Stage 1

- Studies that may be relevant but use solely qualitative approaches
- Reviews of previous literature and studies
- Studies analyzing children's self-perceptions (e.g., confidence)
- Studies that look at adults' (often parents' and/or teachers') perceptions
- Content analyses of cultural artifacts (e.g., science textbooks)

Stage 2: Full Text Screening Questions

During full-text screening, you will answer the following questions in Microsoft Access. As shown below, question 3 could be broken into at least 4 smaller subcomponents (but for efficiency, we leave it as one question in Access).



Do I need to answer all the above questions?

- If the answer is "no" to any of Questions 1-3, then you may simply click "no" for that question, without needing to answer all the other questions.
    - For instance, if the study did not include children (#2), then do not spend time trying to determine if the stereotype measure meets our requirements (#3 or #4).
- However, if the answer is "yes" to Questions 1-3 (i.e., study presented quantitative results on children's ability stereotypes), then please fill out *all* questions.

1. **Did the study present quantitative results from primary research (not literature review)?**
    a. Include quantitative studies (e.g., 1–5 stereotype scale).
    b. Include qualitative studies that also present quantitative summaries of results (e.g., number of children indicating boys are better based on coding structured interviews or open-ended survey questions).
    c. Exclude qualitative studies without quantitative summaries (e.g., qualitative thematic analysis of interview data).
    d. Exclude systematic reviews, meta-analyses, other literature reviews, narrative essays, opinion articles not reporting primary research, and so on.

2. **Did the study include a sample of children (18 years or younger)?**
    a. Include any grade level before college (PreK-12).
    b. Include all-female, all-male, or mixed-sex samples.
    c. Exclude samples of only college students, teachers, parents, or adults.
    d. Exclude analyses of only cultural artifacts (e.g., children's textbooks).

3a. Did the study measure *gender stereotypes*?
    a. Include beliefs about gender differences, gendered beliefs, perceptions of boys as good at math, etc. ("stereotype" does not need be explicitly mentioned).
    b. Exclude stereotypes about other groups (e.g., racial but not gender stereotypes).
    c. Exclude studies about actual gender differences (e.g., gender gaps in math achievement) but not beliefs or perceptions about those differences.
    d. Exclude stereotype threat studies that manipulated stereotype salience but did not measure children's stereotypes.
    e. Exclude studies only about self-perceptions of abilities (e.g., "How good do you think you are at math?") but not perceptions of gender differences among others.
        i. Note that the results might be described in ambiguous ways (e.g., "girls perceived their abilities in stereotypical ways" could imply measuring either girls' self-perceptions of their own individual abilities or gender stereotypes).

3b. Did the study measure *children's* gender stereotypes?
    a. Include children's beliefs about gender differences among child targets (e.g., "Are girls or boys better?") or adult targets (e.g., "Are men or women better?")
    b. Include children's perceptions of others' stereotypes, including perceptions of the stereotypes held by adults (e.g., "Do adults think that boys or girls are better?")
    c. Exclude studies measuring teachers' or parents' stereotypes but not children's (e.g., investigating how parents' stereotypes relates to children's math achievement).

See Table 2 in the following appendix for different types of eligible and ineligible perceptions.

3c. Did the study measure children's gender stereotypes *about abilities*?
    a. Include stereotypes about ability, including innate talent or ability more broadly defined as performance (in an academic, cognitive, or occupational domain).
    b. Include indirect or "implicit" measures *if* they directly address ability beliefs.
    c. Exclude measures about cultural fit such as…
        i. Indirect or "implicit" associations not about ability (e.g., math-gender IAT).
        ii. Liking mathematics more.
        iii. Gender role attitudes ("Is science more appropriate for girls or boys?").
        iv. Gender representation in a field ("Are engineers typically men or women?").

See Tables 3 and 4 in the following appendix for example measures and distinctions between ability and cultural fit stereotypes. If a stereotype scale has a mix of eligible and ineligible items (e.g., ability items mixed with cultural fit items), still include the study; we will later author query for results from just the eligible items.

3d. Did the study measure children's gender stereotypes about *STEM or verbal* abilities?

a. Include spatial abilities such as mental rotation or navigation (they will later be considered "STEM" abilities, but at this stage, we want to flag them).
b. Include occupational abilities about STEM jobs (e.g., "Who are better physicists?")
c. Include academic or cognitive abilities in STEM or verbal domains (e.g., "Who is better at math? Who is better at reading?").
d. Exclude stereotypes about domain-general academic performance (e.g., "Who gets better grades in school?") or cognitive performance (e.g., "Who is more intelligent?").
    i. But still mark the "domain-general" checkbox, as we may want to later return to these studies, so we need to flag them.
    ii. Only select this checkbox if the stereotype was about domain-general *ability* (e.g., getting good grades, being smart) rather than other domain-general traits (e.g., being disruptive in classrooms).

4. Did the stereotype measure's response structure permit unambiguous directional assessment of stereotypes?

a. Include measures that allow respondents to express beliefs of male versus female superiority in a symmetric way. Examples include, but are not necessarily limited to:
    i. Direct comparison of female and male targets ("Are girls or boys better?").
    ii. Separate ratings of female and male targets ("How good are girls/boys?").
    iii. Likert measures based on agreement to male-biased items ("boys are better") subtracted by agreement to analogously worded female-biased items ("girls are better"). The linguistic structure should exactly match.
    iv. Indirect measures that present male and female targets (choose the best student at science among pictures of boys and girls).
b. Exclude Likert measures based on agreement to statements about gender equality (e.g., "girls are as good as boys").
c. Exclude Likert measures based on agreement to directional statements ("boys are better than girls") that lack analogously worded statements in the opposite direction ("girls are better than boys").
d. Exclude response structures based on cross-domain comparisons for one sex (e.g., "Are girls better in math or reading?").
    i. However, still complete all the other eligibility questions because we want to know if the study satisfies all inclusion criteria but this one.

5. Did the study measure children's stereotypes before an experimental manipulation?

a. Include non-experimental studies.
b. Include stereotype threat experiments if children's stereotypes were measured before the manipulation (one example is Muzzatti & Agnoli, 2007, Study 1).
c. Exclude if stereotypes were measured after the experimental manipulation (one example is Galdi et al., 2014).
    i. However, still complete all the other eligibility questions because we want to know if the study satisfies all inclusion criteria but this one.

Stage 3: Data Availability Screening

Before coding an eligible study, you should first check if at least one effect size can be extracted from it, based on the following questions. If not, David will send an author query.

1. Were means reported separately for when ability stereotypes were first measured?
    a. Yes
    b. No, an AQ would be needed for information to calculate effect sizes
2. Were stereotype means reported separately for ability-related items?
    a. Yes
    b. No, an AQ would be needed for information to calculate effect sizes
3. Were any STEM outcomes (motivational or performance) also measured?
    a. Yes
    b. No
4. If “yes” to Q3, were bivariate correlations with STEM outcomes reported separately for boys and girls?
    a. STEM stereotypes: (a) yes, (b) no, (c) N/A
    b. Spatial stereotypes: (a) yes, (b) no, (c) N/A
    c. Verbal stereotypes: (a) yes, (b) no, (c) N/A
    d. Domain-general stereotypes: (a) yes, (b) no, (c) N/A

Supplemental Tables

Table 2. Example Types of Eligible (Boxed in Red) and Ineligible Perceptions

Category	Examples
Personal stereotype endorsement (include)	• As for you, how well do you think girls do in mathematics? • As for you, how well do you think boys do in mathematics?
Perceptions of others' stereotypes (include)	• How well do adults think girls do in mathematics? • How well do adults think boys do in mathematics?
Self-perceptions of ability (exclude)	• As for you, how well do you think you do in mathematics?

Note. The last row would be an eligible outcome for Meta-Analysis 2, but the study must also measure ability stereotypes to be eligible (in either meta-analysis). Hence, at this screening stage, you may essentially ignore measures such as confidence and interests.

Table 3. Example Eligible Gender Stereotype Measures

Category	Description	Examples
BETTER	Lacks explicit language about "abilities" or "skills"	• Are girls or boys better at mathematics? • Who does well in science, women or men? • How good are boys/girls at reading?
SCHOOL	Explicitly concerns school performance	• Do girls or boys earn higher grades in science? • Do girls or boys do better in mathematics classes?
ABILITY/ TALENT	Explicitly uses language suggesting innate traits	• Do women or men have more natural aptitude for physics? • Rate the reading ability of girls and boys. • Are girls or boys more talented in geometry? • Women/men are more gifted in math than men/women.
SKILLS	Uses language about (potentially learned) skills	• Who has better writing skills, girls or boys? • Are girls or boys better skilled in science?
CAN	Uses language about future potential	• Who can be the best student in math, boys or girls? • Who could become an engineer, women or men?
INDIRECT	Indirect measure lacking an explicit question or evaluative prompt	• An experimenter reads a story about a student who wins state math contests, solves math problems that teachers cannot solve, and earns straight As in math classes. The experimenter then asks the child participant to repeat the story, noting if the child used masculine or feminine pronouns (e.g., "he" or "she"). • An experimenter reads a similar story, asks the child participant to draw a picture of the described student, and records whether the drawing depicted a male or female character based on cues such as hair style and clothing.

Table 4. Example Distinctions Between Ability and Cultural Fit Stereotypes

	Ability (Include)	Cultural Fit (Exclude)
Direct measure	• Who do you think is better at math: boys or girls? • How much do you associate being good at science with men or women? • Who has the abilities to become engineers: men or women?	• Who do you think likes math more: boys or girls? • How much do you associate science with men or women? • Who are engineers: men or women?
Indirect measure	• Draw a student good at science. • Modified Affect Misattribution Procedure task with response keys for "good at math" or "bad at math." • Children select the student talented at math from pictures of women and men (but their gender is not explicitly mentioned).	• Draw a scientist. • Math-gender Implicit Association Test. • Children select the scientist from pictures of women and men (but their gender is not explicitly mentioned).

Appendix C: Codebook

T_01_Study		
Variable Name	**Description**	**Format**
StudyID	Unique study ID	Numeric
Authors	Authors	Text
PubYear	Year of publication	Numeric
CollectionYear	Year of data collection	Numeric
PubType	Publication type 1 = Book 2 = Book Section 3 = Conference Paper 4 = Journal Article 5 = Report 6 = Thesis 7 = Other 8 = Unpublished data	Numeric
PubTypeOth	Publication type - Other, specify	Text
APA	Full citation	Text
OverallNotes	Overall notes	Text
StaffID	Staff ID 1 = David Miller 2 = Jillian Lauer 3 = Abigail Jeffreys 4 = Robert Schwarzhaupt	Numeric

T_02_Sample

Variable Name	Description	Format
SampleID	Unique sample ID	Numeric
SampleName	Name of sample	Text
AgeProvided	Type of age variable coded Mean age Grade level Midpoint of age or grade level range	Dropdown
SampleAge	Age in years (two decimal places)	Numeric
Total_N	Sample size for all genders	Numeric
Female_N	Sample size female	Numeric
Male_N	Sample size male	Numeric
EconDisadv_Pct	Percent sample considered economically disadvantaged	Numeric
EconDisadv_Description	Description of sample's socioeconomic class, if provided by authors 1 = Low 2 = Middle 3 = High 4 = Mixed SES 99 = Not reported (exclusive option)	Numeric
EconDisadv_Meas	Economic disadvantage measure 1 = Free/Reduced Price Lunch Program 2 = Family Income 3 = Parental Education Level 4 = Other 99 = Not reported (exclusive option)	Numeric
EconDisadv_MeasOther	Economic disadvantage measure – Other, specify	Text
Pct_White	Percent sample White, non-Hispanic	Numeric
Pct_Black_AA	Percent sample Black or African American	Numeric
Pct_Hispanic	Percent sample Hispanic/Latinx	Numeric
Pct_A_Ind_Alask	Percent sample American Indian or Alaska Native	Numeric
Pct_Asian	Percent sample Asian	Numeric
Pct_Nat_Haw_PI	Percent sample Native Hawaiian or Pacific Islander	Numeric
Pct_ChildrenofColor	Percent of sample described as children of color and/or racial-ethnic minority children, if authors reported an aggregated %	Numeric
Sample_Locality	Sample study locality (Select All That Apply) 1 = Urban 2 = Suburban 3 = Rural 99 = Not reported (exclusive option)	Boolean
Sample_US_Region	Sample study geographic region, if in the United States (Select All That Apply) 1 = West 2 = Midwest 3 = Southwest 4 = Northeast 5 = Southeast 6 = U.S. Territories (Puerto Rico, U.S. Virgin Islands, Guam, Northern Mariana Islands, American Samoa) 7 = Other 99 = Not reported (exclusive option)	Boolean
Sample_US_Region_Other	Other region of U.S., if Other specified	Text
Sample_Country	Sample country of testing's alpha 3 code	Text

Sample_SchoolType	**Sample school type, if applicable (Select All That Apply)** **1 = Public** **2 = Private, religious** **3 = Private, non-religious** **99 = Not reported (exclusive option)**	**Boolean**
Sample_SchoolComp	**Sample school composition, if applicable** **1 = Co-educational** **2 = Single-sex** **99 = Not reported (exclusive option)**	**Boolean**

T_03_StereotypeMeasure

Variable Name	Description	Format
SampleID	Unique sample ID	Numeric
StereoMeasureID	Unique stereotype measure ID	Numeric
StereoMeasureName	Name of stereotype measure	Text
StereoMeasureDesc	Brief description of stereotype measure and/or citation for task	Text
StereoDomain	Domain described in stereotype measure (Select all that apply) Biology / life sciences Chemistry Computer science / technology Engineering Math Physics Science (General) Spatial ability Verbal ability Other	Boolean
StereoDomain_Other	Other academic discipline, if Other specified	Text
StereoMeasureType	Type of stereotype measure (Select only one) Direct / overt Indirect / covert	Dropdown
StereoTargetAge	Age of targets in stereotype measure (Select only one) Children (under 12) Adults (18-64) Other Unspecified or mixed Unknown	Dropdown
StereoTargetAge_Other	Other target age group, if Other specified	Text
StereoExceptionality	Stereotype content (Select only one) Average differences Differences in exceptional performers Both Unspecified	Dropdown
StereoAbilityPct	Percentage of stereotype scale items that are ability related	Numeric
StereoReli	Reliability statistic (two decimals)	Numeric
StereoReliType	Reliability metric (Select only one) Cronbach's alpha Other	Dropdown
StereoReliType_Other	Reliability metric, if Other specified	Text
StereoScaleType	Type of stereotype scale (Select only one) Forced-choice Likert Visual analog Other	Dropdown
StereoScaleType_Other	Other scale type, if Other specified	Text
StereoScaleNumOptions	Number of response options on stereotype scale	Text
StereoScaleItems	Text from items on the stereotype measure, if provided	Text
StereoScaleItemsNo	Number of items on the stereotype measure	Numeric
StereoScore_Midpoint	The midpoint of the scale/the gender-neutral response	Numeric
StereoScore_Max	Maximum score on stereotype scale	Numeric
StereoScore_Mean	Sample raw mean score on the stereotype scale	Numeric

StereoScore_SD	**Sample standard deviation on the stereotype scale**	**Numeric**
StereoScore_PageNum	**Page number on which the sample mean was located**	**Text**
StereoScore_Directionality	**Gender stereotyped as having greater ability** **Girls / women** **Boys / men** **Equal**	**Dropdown**
ES_Type	**Type of effect size, if reported** **Cohen's *d*** **Hedges' *g*** ***t* test statisitc** ***p*-value** **Other**	**Dropdown**
ES_TypeOther	**Type of effect size, if Other specified**	**Numeric**
ES_Val	**Reported effect size estimate**	**Numeric**
ES_PageNum	**Page number on which the effect size was located**	**Text**

T_04_OutcomeMeasure

Variable Name	Description	Format
SampleID	Unique sample ID	Numeric
StereoMeasureID	Unique stereotype measure ID	Numeric
OutcomeMeasureID	Name of stereotype measure	Text
OutcomeDomain	Domain described in outcome measure (Select all that apply) Biology / life sciences Chemistry Computer science / technology Engineering Math Physics Science (General) Spatial ability Verbal ability Other	Boolean
OutcomeDomain_Other	Other academic discipline, if Other specified	Text
OutcomeType	Type of outcome measure (Select only one) Motivational Attitudinal Performance Other	Dropdown
OutcomeType_Other	Other outcome measure description, if Other specified	Text
OutcomeType_AuthorDesc	Description of constructs assessed by outcome measure, as provided by authors	Text
ES_SampleSize	Sample size used when computing effect size	Numeric
ES_Reportedr	Reported effect size estimate (correlation coefficient), to 2 decimals	Numeric
ES_RegCoef	Regression coefficient, if reported and a single-predictor model	Numeric
ES_RegSE	Regression standard error, if reported and a single-predictor model	Numeric
ES_OtherType	Effect size type (e.g., Spearman's *r*) if other is available	Text
ES_OtherVal	Effect size value (e.g., 0.20) if other is available	Numeric

Appendix D: Selective Reporting Bias Analyses

This appendix details how we will conduct and interpret our three chosen approaches for diagnosing and adjusting for selective reporting bias: (a) selection modeling, (b) meta-regression to assess small-study effects, and (c) comparison of published versus unpublished studies. As justified below, if these three approaches yield diverging conclusions, we will place the greatest weight on selection models because of their superior performance in research conditions that are likely relevant to our meta-analysis such as moderate to large between-study heterogeneity (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2019).

For readers wishing to verify if our reported analyses matched our preregistered plan, they can skip directly to the section entitled *Analytic Plan* if they care less about the detailed considerations and data simulations used in forming that plan.

Hypotheses About Publication Bias for This Literature

Following Carter et al.'s (2019) recommendation, we first consider how selective reporting may operate in the specific literature that we will synthesize, before detailing our analytic plan to statistically assess it.

Potential Publication Bias Mechanisms: Some social psychological and developmental researchers assume that stereotypes about female inferiority in mathematics are widespread, at least among adult participants. For instance, in their seminal article, Spencer, Steele, and Quinn (1999) argued that, "widely known stereotypes in this society impute to women less ability in mathematics and related domains…women bear the extra burden of having a stereotype that alleges a sex-based inability" (p. 6).

Studies that fail to find pro-male STEM ability stereotypes could therefore be viewed with suspicion and additional scrutiny, potentially leading to selective reporting due to internal pressures (e.g., study authors distrusting their own results or deciding they would be too difficult to publish) or external pressures (e.g., from journal editors and peer reviewers). As a specific example, before this meta-analytic project began, one researcher told this project's PI about the researcher's unpublished findings on STEM ability stereotypes, noting that, "our girls do not seem to be endorsing gender stereotypes. In fact, even our older girls (up to around 14) do not seem to endorse gender stereotypes…I had been worried about getting pushback when making the claim that our young girls do not seem to endorse gender stereotypes." The comment about "pushback" indicates perceived external pressures that could, in the aggregate, lead to the underreporting of findings that do not show pro-male STEM ability stereotypes (e.g., by making study authors less likely to submit their unexpected results for publication).

Evidence of Publication Bias: Some empirical evidence suggests that publication bias may affect related developmental literatures. For instance, in a meta-analysis of 47 stereotype threat experiments about girls' STEM test performance, Flore and Wicherts (2015) found suggestive evidence of publication bias based on (a) significant funnel plot asymmetry per Egger's regression and (b) excessive significance (i.e., more statistically significant effects than expected based on average statistical power). Analyses also indicated sensitivity to adjustments for potential publication bias; the mean stereotype threat effect was reduced from $g = 0.22$ to $g = 0.07$ after applying trim and fill procedures. The authors emphasized the need for a large preregistered experiment to provide a less biased effect estimate. Filling this need, a subsequent preregistered experiment with a large sample of Dutch high school students ($n = 2064$) found no evidence of an overall stereotype threat effect on girls' mathematics performance; furthermore, key theoretical moderators such as domain identification and test difficulty did not significantly moderate experimental effects (Flore, Mulder, & Wicherts, 2019). Although other interpretations are possible (e.g., funnel plot asymmetry tests have notable limitations as discussed later), these findings are consistent with publication bias affecting the developmental literature on stereotype threat in STEM domains.

More optimistically, however, other considerations temper these concerns about publication bias. For instance, several peer-reviewed journal articles have reported findings failing to show STEM ability stereotypes favoring boys and men (e.g., Martinot & Désert, 2007; Plante, Théorêt, & Favreau, 2009).

Heyman and Legare (2004) even highlighted in their study's abstract that, "perceived gender differences were minimal for math, and those that were seen were consistent with same-sex biases" (p. 227). Other journal articles have even reported findings of perceived female superiority in math and science (e.g., Rowley, Kurtz-Costes, Mistry, & Feagans, 2007; Evans, Copping, Rowley, & Kurtz-Costes, 2011; Kurtz-Costes, Copping, Rowley, & Kinlaw, 2014). Hence, at least some studies failing to find pro-male stereotypes have clearly been published. Lastly, in a meta-analysis on U.S. children's associations of science with men, no evidence of overall publication bias was found, based on either (a) comparison of published versus unpublished studies or (b) small-study effects (Miller, Nolla, Eagly, & Uttal, 2018).

The available evidence therefore suggests both concern and optimism regarding potential publication bias in the developmental literature on gender stereotypes about STEM abilities. Either way, these considerations show that the role of selective reporting should be carefully considered for this project.

Varying Patterns of Publication Bias: The discussion so far has focused on overall publication bias that could distort estimates of aggregate mean effects. However, varying patterns of publication bias could also bias our planned moderator analyses (Coburn & Vevea, 2015). For instance, publication bias could be more extreme with studies of older than younger children. If a study fails to find pro-male stereotypes with younger children (e.g., pre-school children), the researcher may conclude that the children simply have not "learned" societal stereotypes yet. In contrast, null results or results of pro-female stereotypes may be more controversial for samples of older children and adolescents (e.g., high schoolers) who are closer to adults in age. Hence, studies with older children might plausibly be upwardly biased to finding stereotypes favoring boys and men, whereas studies with younger children may be less biased. In this hypothetical scenario, strengthening of pro-male stereotypes with age could reflect differing patterns of publication bias rather than true developmental change.

Encouragingly, Miller et al.'s (2018) meta-analysis found no indication that the evidence for publication bias (i.e., published-unpublished differences and funnel plot asymmetry) varied as a function of the average age of the study sample (supplemental results available upon request; see https://osf.io/3awvj/ for the underlying data). Nevertheless, these hypotheses illustrate how publication bias could influence both the overall *mean* effect and *pattern of variation* across studies (i.e., moderator analyses). Both sets of possible distortions should be considered when evaluating our focal confirmatory analyses. The possibility of biased moderator analyses has largely gone unaddressed in the methodological literature (though see Coburn & Vevea, 2015), but we will return to this point in later sections.

Goals for the Planned Publication Bias Analyses

Several meta-analytic methods now exist that aim to adjust the meta-analytic estimates for publication bias (e.g., Citkowicz & Vevea, 2017; Duval & Tweedie, 2000; Stanley & Doucouliagos, 2014; Vevea & Hedges, 1995) rather than only test for the presence of it (e.g., Egger, Smith, Schneider, & Minder, 1997). However, given the inherent statistical challenges in adjusting for bias, meta-analysis methodologists have often recommended to use bias-correcting methods more as sensitivity analyses and less as definitive estimates of true "corrected" effects (e.g., McShane, Böckenholt, & Hansen, 2016). For instance, when assessing publication bias, Coburn and Vevea (2015) advocated for the approach of triangulation of "using multiple methods of assessment and reporting the range of results, rather than relying on one method and one point estimate" (p. 328).

Carter et al.'s (2019) recent simulation studies and tutorial aimed to provide explicit guidance for conducting such sensitivity analyses. They cautioned that a simplistic triangulation approach could lead to the application of inferior publication bias methods that perform poorly in too many realistic scenarios. Because these methods can often yield diverging conclusions, Carter et al. recommended that, "the set of methods employed in a sensitivity analysis should include only those that can be expected to perform reasonably well. Put differently, if a method is known to perform poorly under the conditions that apply to a meta-analysis at hand, it should not be included in a sensitivity analysis, or it should at least be treated with skepticism and given less weight than other methods when the results are evaluated" (p. 137).

As one specific example, we consider popular "trim and fill" approaches to be inappropriate for our meta-analysis because they perform poorly under conditions of moderate to large between-study heterogeneity (see Duval & Tweedie, 2000 for a detailed description of the method). Based on their simulation results, Moreno et al. (2009) concluded that, "with respect to the popular Trim & Fill method, we find it hard to recommend over the regression-based alternatives due to its potentially misleading adjustments and poor coverage probabilities, especially when between-study variance is present" (p. 12). Other simulation studies further reinforce this conclusion (e.g., Carter et al., 2019; Terrin, Schmid, Lau, & Olkin, 2003). For our study, this limitation is critical because our scoping review found that the magnitude and direction of STEM ability stereotypes varies widely across studies (e.g., some finding stereotypes favoring boys and men, but others finding the exact opposite). Hence, we require methods that perform reasonably well with heterogeneity, eliminating some candidate methods (e.g., trim and fill) from further consideration.

Carter et al. (2019) formalized the concept of a *methods performance check* in which meta-analysts first evaluate which bias-correcting methods can be expected, based on simulated data, to perform reasonably well in conditions that are plausible for the specific research environment at hand. The authors provided an interactive web application (http://www.shinyapps.org/apps/metaExplorer) and a tutorial for applying this approach, using a meta-analysis from social psychology as a specific example. Carter et al. recommended conducting this methods performance check prior to data analysis to avoid "assumption hacking" (e.g., after viewing the results, researchers may more heavily weight methods that yield conclusions favoring their prior hypotheses). Hence, analysts should "preregister a method performance check prior to data collection and define which methods will be given the greatest weight if different methods provide conflicting results" (Carter et al., 2019, p. 140). We follow this advice in the next section.

Hence, though exploratory, our planned selection bias analyses will still be constrained and guided by simulation studies on what methods can be expected a priori to perform reasonably well in conditions plausible for the research environment at hand (i.e., developmental literature on gender stereotypes about STEM abilities). The methods described in the next section can also provide formal significance tests of publication bias (e.g., significant evidence of funnel plot asymmetry). Though we will examine these publication bias tests, we will interpret them cautiously because they can be underpowered, even for meta-analyses with many studies (e.g., 50 or more; Citkowicz & Vevea, 2017; Macaskill, Walter, & Irwig, 2001; Renkewitz & Keiner, 2018; though see Pustejovsky & Rodgers, 2019 for selection models). Hence, our sensitivity analyses will focus more on adjusting for, rather than detecting, publication bias.

Methods Performance Check

This section details how we used Carter et al.'s (2019) interactive website to identify methods that might perform reasonably well for our meta-analysis (http://www.shinyapps.org/apps/metaExplorer). We did not conduct any new simulations ourselves but instead graphed the specific results from that website that are likely the most relevant to specific our project. In this following section, we group the bias correction methods into two categories of those based on the *p*-value (e.g., selection models, *p*-curve analyses) and those based on small-study effects (e.g., trim and fill, PET). At the end, we also consider another approach based on comparing published and unpublished studies.

Simulation Conditions: Carter et al.'s (2019) simulations were based on individual studies with a two-group experimental design summarized by a Cohen's *d* effect size metric. Simulated sample sizes were based on an empirical distribution of per-group sample sizes in studies published between 1995 to 2006 in four psychology journals; the median per-group sample size was 23 (25% quantile: 14, 75% quantile: 50). Although these simulated conditions do not exactly match our meta-analysis (e.g., the effect size metrics are different), Carter et al.'s simulations should be sufficient to at least generate reasonable hypotheses about methods performance for our meta-analysis. Readers should see Carter et al.'s original paper for an in-depth description of these simulations; here we instead focus on how we applied them.

As noted earlier, given our scoping review, we required methods that perform well in cases of moderate to large between-study heterogeneity, so we choose the largest simulated heterogeneity value available

on the metaExplorer website (τ = 0.4). We choose the number of studies in the meta-analysis to be k = 100 because we expect to code approximately 80 to 120 studies for Meta-Analysis 1 (reducing this value to k = 60 increased the variability of mean effect sizes did not substantially change the results in the following section about relative methods performance).

Carter et al. (2019) simulated three conditions of publication bias: (a) no publication bias, (b) moderate publication bias, and (c) strong publication bias. These scenarios assumed that the probability of study publication depended solely on the p value for between-group differences (e.g., significant results in the predicted direction were always "published" with 100% probability). The below figure shows the simulated probability of publication as a function of the one-tailed p value (ps < .025 indicate significant results based on a two-tailed test). We examined methods performance under both medium and strong publication bias, but the conclusions about relative methods performance were largely the same, so the following section only present results for performance under strong publication bias.

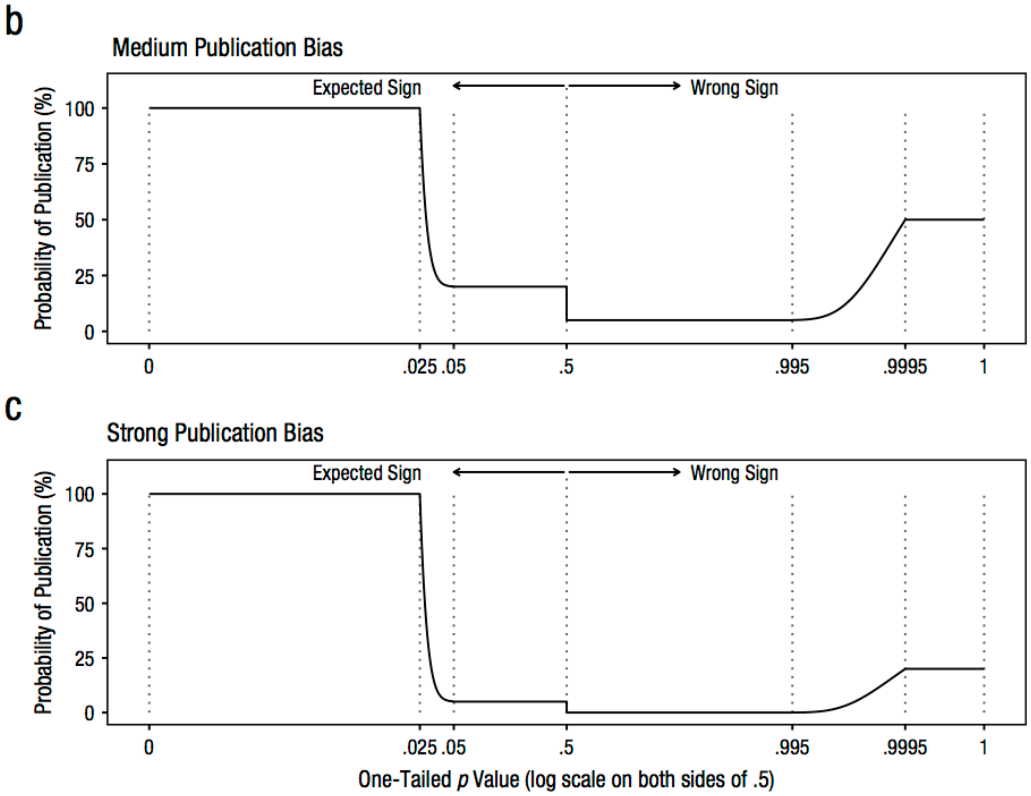


Fig. 2. Implementation of publication bias in the simulation. The graphs show the probability of publication as a function of the one-tailed p value of the simulated results in (a) the no-publication-bias condition, (b) the medium-publication-bias condition, and (c) the strong-publication-bias condition. The x-axes have a logarithmic scale on both sides of $p_{one\text{-}tailed}$ = .5 to increase the visibility of the function at the high and low ends of the scale. This figure is available at https://osf.io/f6esc/, under a CC-BY4.0 license.

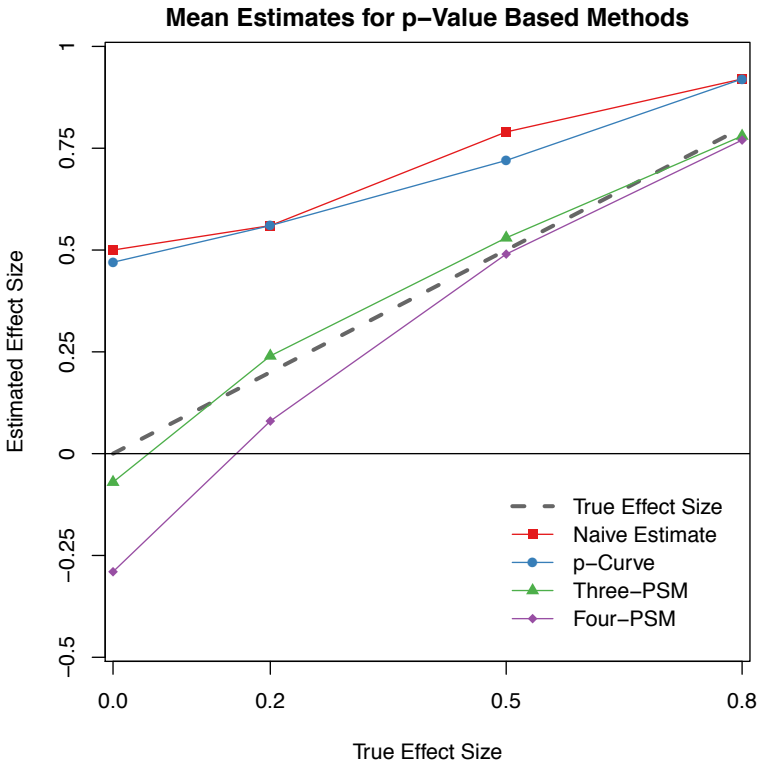
p-Value Based Methods: Carter et al. (2019) examined the performance of three-parameter selection models that estimated (a) the mean effect size, (b) between-study heterogeneity, and (c) the relative likelihood that nonsignificant results are published as compared to directionally consistent, significant results (i.e., p < .05 in the expected direction). The last parameter is modeled by a step function with one cut point at the one-tailed p-value of .025 (corresponding to .05 for a conventional two-tailed test).

The three-parameter model clearly misrepresents the true simulated selection process because the model assumes a constant probability of publication for all one-tailed ps > 0.025; in contrast, the true simulated publication probability varies in that range, as shown above. Hence, if this model still performs

well, the results would provide some evidence of robustness to model misspecification. To address this point, Carter et al.'s (2019) website also includes four-parameter models with one more cut point at p = .50, allowing for the possibility that directionally consistent findings may be favored in the publication process, even if they are not significantly significant.

The more recently proposed *p*-curve and *p*-uniform methods (e.g., Simonsohn, Nelson, & Simmons, 2014) are closely related in principle to selection models, as McShane et al. (2016) discussed, even though these methods were developed separately from each other. The *p*-curve and *p*-uniform methods are based on finding an effect size parameter that best recreates the observed distribution of statistically significant *p* values; only true effects are expected to generate right-skewed *p*-curves, containing more low (.01s) than high (.04s) *p* values (Simonsohn et al., 2014). However, for reasons detailed below, we consider the underlying goals of *p*-curve analysis to be fundamentally misaligned with ours. It would require us to exclusively focus on studies with directionally consistent, significant results (e.g., pro-male stereotypes), forcing us to exclude any other studies with different results (e.g., pro-female stereotypes).

Performance of *p-Value Based Methods:* The below figure shows the mean estimates for the *p*-value based methods as a function of the true average effect size (these simulations assume strong publication bias, large heterogeneity of τ = 0.4, and k = 100 studies). Perfectly unbiased estimates would fall exactly along the dotted diagonal line; any deviation from it indicates bias from the true population averages. We made the below graph, but the plotted values came directly from Carter et al.'s (2019) website.



Note. These simulations assumed strong publication bias, large heterogeneity (τ = 0.4), k = 100 studies, and no questionable research practices (see later section). The naïve estimates are based on simple random-effects models. The *p*-uniform method is omitted because it had nearly identical results to *p*-curve. PSM = parameter selection model. The plotted values came directly from Carter et al.'s (2019) website (http://www.shinyapps.org/apps/metaExplorer).

As shown, in this simulated case of strong publication bias, the three-parameter selection model has the best overall performance, falling the closest to the diagonal (i.e., the true estimates). When the true average effect size is zero, the naïve random-effects estimate is upwardly biased as expected (d = 0.50),

whereas the three-parameter model only slightly overcorrects with a small downward bias (d = -0.07). These results show overall mean bias, but similar conclusions were found when examining a measure of estimation accuracy (root mean square error) that accounts for both mean bias and estimation variance.

These results are notable given that the three-parameter model misrepresents the true simulated selection process, as noted earlier. Similar favorable results were found in the case of moderate publication bias (results not shown), which has selection probabilities that vary even more widely in the range of $0.025 < p < 1.0$, showing some robustness to model misspecification. Somewhat surprisingly, the four-parameter model performs less well by overcorrecting with small true effect sizes, counter to the naïve wisdom that more selection model parameters are always better, even with many studies (k = 100).

In contrast, *p*-curve analysis severely overestimates the population average when the true effect size is zero (estimated d = 0.47), barely offering any improvement over the naïve random-effects estimate (d = 0.50). This overestimation comes from the presence of heterogeneity (i.e., the mean estimate is no longer biased when heterogeneity is set to zero). Furthermore, when heterogeneity is large, *p*-curve analysis is severely biased *even in the absence of publication bias* (d = 0.47 when the true average effect is zero and no selection is present; results not shown in the graph). Based on similar simulated results, McShane et al. (2016) recommended to avoid using *p*-curve (or *p*-uniform) analysis when between-study heterogeneity is expected or observed, which characterizes most of psychological research.

The original developers of *p*-curve analysis have responded to these critiques (Simonsohn, Simmons, & Nelson, 2018; see http://datacolada.org/67), arguing that the systematic overestimation of population effect sizes is desired because the method only aims to recreate the true average effect size for studies with an observed $p < .05$ in the expected direction, not for all studies. Any studies with $p > .05$ (or results in the opposite direction) are discarded from analysis, as the method developers have also noted (Simonsohn, Simmons, & Nelson, 2017; see http://datacolada.org/61). Hence, when the true underlying effects vary across studies (i.e., heterogeneity is present), *p*-curve analysis will preferentially select studies with larger true effects because those will be more likely to find $p < .05$ in the expected direction. This selection at the analysis stage therefore explains the overestimation of population averages.

Although the focus on studies with directionally consistent, significant results might be appropriate for some research goals, we agree with Carter et al.'s (2019) point that, "meta-analysts generally aim to recover the average of the distribution of all true effects related to the phenomenon of interest" (p. 122). Furthermore, one of our central goals is to understand variability across studies such as characterizing why some studies find stereotypes strongly favoring boys and men, whereas other studies find the exact opposite. In contrast, *p*-curve analysis would force us to only consider significant findings of pro-male stereotypes. In this regard, we view the underlying goals of *p*-curve analysis to be fundamentally misaligned with our research goals. In contrast, selection models are better aligned and show satisfactory performance in simulated conditions plausible for the research environment at hand.

Methods Based on Small-Study Effects: Many conventional methods for examining publication bias (e.g., trim and fill, funnel plots) rest on the assumption that small studies with small observed effects often are not published due to lack of statistical significance (see Borenstein et al., 2009, Chapter 30). Small studies must instead observe large effects to find $p < .05$. In contrast, large studies with small observed effects can still obtain statistical significance due to larger sample sizes. Consequently, in the presence of publication bias, a funnel plot of effect sizes plotted against their standard errors will be asymmetric, due to the selective censoring of small studies with small nonsignificant effects.

Regression approaches have existed for decades aiming to *detect* these small-study effects. More recent methodological developments have aimed to also *adjust* for small-study effects using meta-regression (e.g., Moreno et al. 2009; Stanley & Doucouliagos, 2014). For instance, with Stanley and Doucouliagos' precision-effect test (PET), effect sizes are predicted by their standard error in a weighted least squares regression ($ES = \beta_0 + \beta_1 SE$ + error) with fixed-effects inverse variance weights. The intercept in this

regression model (β_0) is the adjusted mean estimate when *SE* = 0, extrapolating to studies with infinite sample size, presumably correcting for the bias from the selective reporting of small studies.

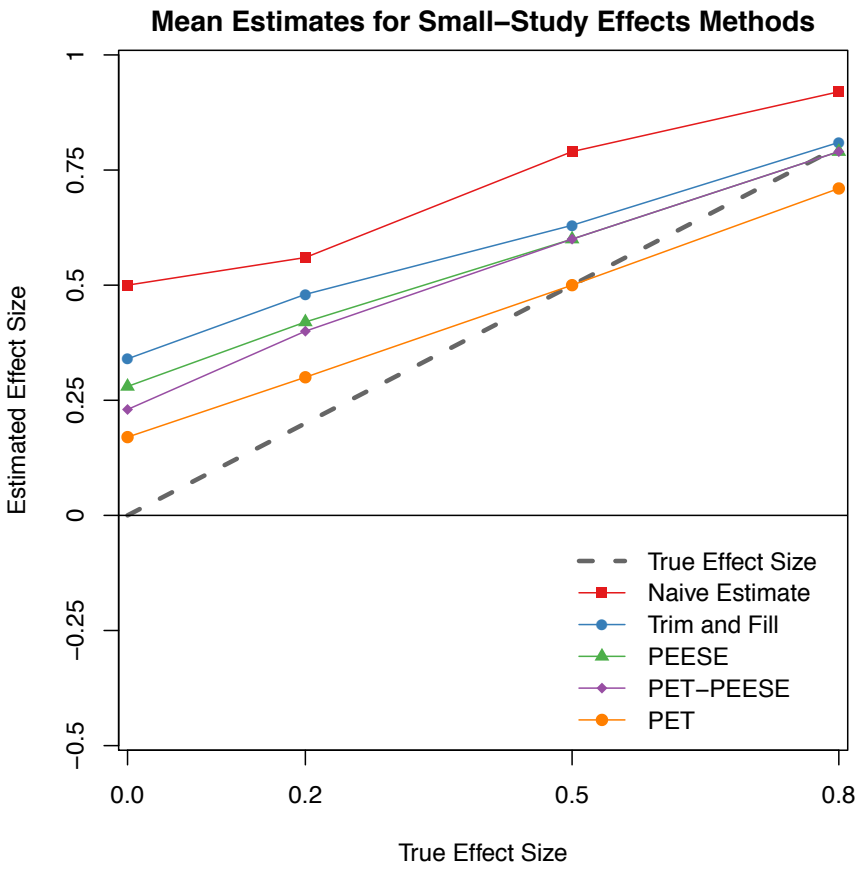
The significance test for the predictor coefficient (β_1) in PET can be shown to be equivalent to the commonly used Egger's regression test of funnel plot asymmetry (Egger et al., 1997). Hence, because they rest on the same assumptions about small-study effects, these newer regression-based adjustment methods share many similarities with earlier methods based on funnel plots. As such, they also share many of the same limitations. Notably, the presence of small-study effects may not necessarily indicate publication bias if smaller studies systematically differ from larger studies in the phenomena they study, as methodologists have repeatedly noted (e.g., Sterne & Egger, 2005, Terrin et al., 2003). For instance, a recent meta-analysis on teacher professional development programs found evidence that intervention effects were larger for studies with smaller samples of teachers (Garrett et al., 2019). However, the study authors cautioned that, "one potential explanation is that studies with more teachers represent scale-up studies and reflect the difficulties of scaling teacher professional development" (p. 129-130).

Hence, small-study effects may not necessarily provide direct evidence of publication bias. This limitation is important to keep in mind because Carter et al.'s (2019) simulations did not examine it; the simulations assumed no confound between sample size and other study characteristics (though see the simulations from Terrin et al., 2003, which provide some evidence that selection models are robust to such confounds, though trim and fill is not). Stanley and Doucouliagos (2014) proposed other regression-based adjustments such as the precision-effect estimate with standard error (PEESE) that uses the effect size variance (SE^2) as the predictor[1], but this alternate functional form obviously shares the same limitations.

However, one key advantage of these regression-based approaches is their flexibility to control for study moderators, account for dependent effect size structures, and include interaction terms (e.g., controlling for study moderators can help mitigate against ambiguous interpretations of small-study effects). We elaborate on this point in the following sections after this current section.

Performance of Small-Study Effects Adjustment Methods: The below graph shows mean estimates for the examined methods based on small-study effects (e.g., PET, trim and fill).

[1]The name precision-effect estimate *with standard error* may be confusing because PEESE uses the effect size variance, not standard error, as the predictor. However, this naming reflects that the predictor used in PET (i.e., standard error) is multiplied by the standard error (i.e., SE^2 or the variance).



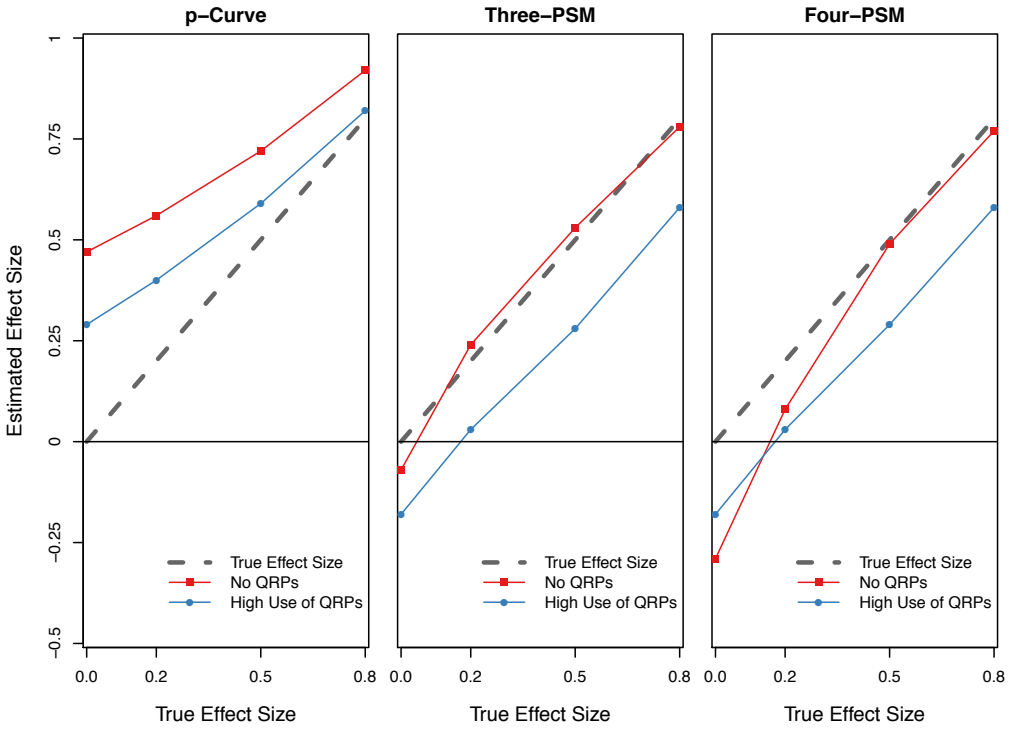
Note. These simulations assumed strong publication bias, large heterogeneity (τ = 0.4), k = 100 studies, and no questionable research practices (see later section). The naïve estimates are based on simple random-effects models. The plotted values came directly from Carter et al.'s (2019) website (http://www.shinyapps.org/apps/metaExplorer).

Consistent with our earlier discussion, the popular "trim and fill" method performs poorly in this case of large heterogeneity, yielding an average upward bias of d = 0.34 when the true effect is zero, performing the worst for this class of adjustment methods. In contrast, PET performs the best for small true effects, though it is still upwardly biased for that scenario (estimated d = 0.17 when the true effect is zero). This mean bias for PET reverses for larger simulated effects, underestimating the true d = 0.80 effect as 0.71, whereas PEESE performs better in this scenario of true large effects. As Stanley and Doucouliagos (2014) explained in detail, this behavior is expected, leading the authors to propose a conditional estimator (PET-PEESE) that uses PEESE if the mean estimate is significant, but uses PET otherwise.

However, we view the conservative bias of PET for large true effects to be tolerable because an effect estimate of 0.80 and 0.71 would both lead to the same substantive conclusion (i.e., the effect is large). Except in this case of large true effects, PET performs the best in terms of mean bias correction among this class of methods (examining the root mean square error, a measure of estimation accuracy, yields similar conclusions largely due to PET's reduced bias with small true effects).

Influence of Questionable Research Practices (QRPs): In addition to studying publication bias (i.e., likelihood of publication based on the p value), Carter et al. (2019) varied the extent of questionable research practices (QRPs) or "p-hacking" that seek to find significant results based on flexible analytic decisions (operationalized here as optional removal of outliers, optional selection between two dependent variables, optional use of moderators, and optional stopping). Again, we direct readers to the original paper for a full description. To examine the robustness of our conclusions, we examined both the simulated use of no QRPs versus a high QRP environment.

Adding the presence of QRPs to the previously mentioned scenario (strong publication bias, large heterogeneity, k = 100 studies) did not substantially change simulated results for the methods based on small-study effects (e.g., PET, PEESE). However, for the *p*-value based methods, adding QRPs pulled the mean estimates down by approximately 0.1 to 0.2 SDs in most cases, as shown below.



Note. These simulations assumed strong publication bias, large heterogeneity (τ = 0.4), and k = 100 studies. The naïve estimate is based on a simple random-effects model. PSM = parameter selection model. The plotted values came directly from Carter et al.'s (2019) website (http://www.shinyapps.org/apps/metaExplorer).

Consequently, the three-parameter selection model is now downwardly biased, losing its previous advantage over the four-parameter model; both are now equally conservative. The upward bias in *p*-curve estimates is also now slightly improved because of the downward pull of adding QRPs.

Although this behavior may seem counterintuitive, it is well understood in the context of *p*-curve analysis. As noted earlier, that method relies on the assumption that only true effects are expected to generate right-skewed *p*-curves, containing more low (.01s) than high (.04s) *p* values (Simonsohn et al., 2014). However, QRPs or "*p*-hacking" will distort the observed distribution of statistically significant *p* values. As Simonsohn et al. (2014) explained, "because *p*-hacking leads researchers to quit conducting analyses upon obtaining a statistically significant finding, *p*-hacking is disproportionately likely to introduce 'large' significant *p* values into the observed distribution (i.e., *p* values just below .05). As a result, *p*-hacking reduces the right skew of *p*-curve. Because smaller effect sizes are associated with less right-skewed *p*-curves, *p*-hacking causes *p*-curve to underestimate effect sizes" (p. 670). A related principle may likely apply to selection models because *p*-curve analysis shares some similarities with selection models, especially Hedges' (1984) original formulation. As McShane et al. (2016) explained, *p*-curve and *p*-uniform approaches "can be viewed as alternative implementations of the original Hedges (1984) selection method approach that employ different estimation strategies" (p. 731).

The key take-away from these considerations is that selection models may overcorrect (i.e., be negatively biased) in the presence of QRPs, which we will keep in mind when interpreting our results. However, Carter et al. (2019) noted that future simulation studies may show different results with different

implementations of QRPs, cautioning that their study is “best considered a sensitivity analysis that explored the effect of a range of three plausible QRP environments” (p. 120).

Conclusions and Method Selections: From these results, we conclude that three-parameter selection models generally performed the best for the simulated conditions studied, though they slightly overcorrected in the presence of QRPs (i.e., researcher behaviors that seek to find significant results based on analytic decisions such as optional removal of outliers). Without QRPs, three-parameter models surprisingly performed better than four-parameter models that had an additional $p = .50$ cut point, even though the number of studies ($k = 100$) was presumably large enough to estimate additional parameters.

However, little methodological work has directly compared which selection model specifications perform the best in different conditions, so these simulated results are tentative. We therefore will examine the sensitivity of our results to a set of reasonable selection model specifications. In addition to the $p = .025$ and $p = .50$ cut points already mentioned, we will explore adding the $p = .05$ (one-tailed) cut point, which is a meaningful boundary because two-tailed p values in the range between 0.05 and 0.10 are often interpreted as “marginally significant” or “approaching significance” (Pritschet, Powell, & Horne, 2016). To avoid these manual decisions about what p value cut points to choose, Citkowicz and Vevea (2017) recently proposed a beta density weight function that uses two parameters to specify a wide range of possible nonlinear selection models. We will use this beta density model as a further sensitivity analysis.

Another contending method is PET, which uses the standard error as a predictor in regression models to adjust for small-study effects. It was somewhat upwardly biased for true small effects, though less biased than other methods based on small-study effects (e.g., PEESE, trim and fill). This method may also be appropriate for our meta-analysis because our scoping review found studies with generally large, but also variable, sample sizes with a median $n = 241$ (25% quantile: 146, 75% quantile: 383). This point is important because these methods generally require large and variable study sample sizes to perform well because they involve extrapolations to $SE = 0$ (i.e., infinite sample size). In contrast, the sample sizes were much smaller for Carter et al.’s (2019) simulations with a median $n = 23$ (25% quantile: 14, 75% quantile: 50), which can degrade the performance of these meta-regression methods (Stanley, 2017). Hence, we may have found more favorable performance for PET if the distribution of simulated study sample sizes more closely matched those for the specific literature we will meta-analyze.

We will still examine the PEESE estimator as a robustness check but will place less weight on it, especially if the mean estimate is small. As with any other method based on small-study effects, caution is needed when interpreting results because small-study effects can reflect study confounds. However, controlling for other moderators in regression models can help mitigate this possibility.

Adding Published-Unpublished Comparisons: Because published studies often find larger effects than unpublished studies (Polanin, Tanner-Smith, & Hennessy, 2016), we will also examine publication status as a moderator. Like methods based on small-study effects, these analyses should be viewed cautiously because published-unpublished differences could reflect other systematic confounds such as differences in study quality. However, controlling for other moderators can help address this possibility.

As an “adjustment” method, we will consider if our results for the full meta-analytic dataset match those for estimates restricted to the obtained unpublished literature (e.g., dissertations, conference presentations, emails to us). However, this approach should also be viewed cautiously because “unpublished studies” are still presumably subject to some selective reporting pressures. Primary researchers must still decide to present the results in some form for us to meta-analyze them. Furthermore, if published and unpublished studies differ, then estimates restricted to unpublished studies might not be representative of the full population of studies (including published and unpublished ones).

Nevertheless, examining publication status as a moderator can still provide a useful data point for considering the role and magnitude of possible selective reporting biases. Like the other regression-based methods (e.g., PET), we will include publication status (1 = published; 0 = unpublished) as a

moderator in meta-regression models. The regression intercept will serve as the "adjusted" estimate (i.e., estimate for unpublished literature), and the significance test for the moderator will serve as a publication bias test. In contrast to simply excluding published studies, this analytic approach will leverage the published studies to help stabilize the estimation of the between-study heterogeneity parameter.

In the following sections, we will use the term "meta-regression approaches" to refer both to those based on small-study effects (e.g., PET, PEESE) and publication status.

Addressing Effect Size Dependencies

Limitations of Current Methods: One major limitation of all current publication bias methods is that they were developed and studied assuming the statistical independence of effect sizes, not accounting for dependent effect sizes nested within studies. For instance, all of Carter et al.'s (2019) simulations assumed only one effect size from each study. As discussed in a recent meta-analysis, "there is not yet an accepted method in the field to correct for publication bias when using robust variance estimation models" (Bediou, Adams, Mayer, Tipton, Green, & Bavelier, 2018, p. 88). In the absence of explicit guidance from rigorous methodological literature, we created our plan for handling dependent effects by making plausible assumptions about selective reporting mechanisms and consulting our project's meta-analysis methodological advisors.

Plausible Selective Reporting Mechanisms: The most common way to address dependent effect sizes in selection models has been to first aggregate dependent effect sizes and then conduct standard analyses on the aggregated article-level estimates (Coles, Larsen, & Lench, 2019). Although this approach technically addresses the problem of dependent effects, it raises deeper conceptual issues regarding the assumed selective reporting mechanisms. Aggregating effects assumes that publication decisions are based on the statistical significance of an aggregate estimate, not individual estimates, for each study. Although theoretically possible, this assumption seems problematic given empirical evidence on the nature of outcome reporting bias and other selective reporting within articles (e.g., Pigott, Valentine, Polanin, Williams, & Canada, 2013). For instance, in a survey of 2,155 academic psychologists, 65% said they did not report all dependent measures in studies, and 48% said they selectively reported studies that "worked" in an article (John, Loewenstein, & Prelec, 2012). Furthermore, authors could disaggregate their data until a statistically significant effect is found for at least one subgroup, as suggested by research on "*p*-hacking" (e.g., Simmons, Nelson, & Simonsohn, 2011).

For instance, in our project's specific context, a researcher could collect data for fourth- and eighth-graders; find that stereotypes correlate with STEM attitudes among the eighth-graders ($p < .05$) but not fourth-graders ($p = .48$); and then conclude the results reflect developmental change, even if the aggregate correlation is nonsignificant. The researcher could then pursue publication by arguing that the results align with theoretical developmental perspectives, emphasizing the significant results for the older students (but could still report results for the younger students). Publication could therefore be driven by finding "$p < .05$" for at least one outcome, subgroup, or study within an article, even if an aggregate article-level estimate is nonsignificant, not reported, or not even computed by the authors. Hence, aggregating dependent effects could lead to misspecification of even article-level selection processes.

This model misspecification could cause articles with at least one statistically significant result to be represented as a nonsignificant study (depending on the magnitude of the other effects of course). The presence of these "nonsignificant studies" may result in an overly optimistic estimate of the survival of nonsignificant results, meaning that analyses might underestimate the true extent of publication bias.

Selection Modeling: Handling dependent effects in selection models is an active area of methodological research. Hence, if a new validated method is developed in time to address this issue, we will use it in our selection model analyses. However, in the absence of a new method, we will include all effect sizes (without first aggregating) in standard selection models in the *weightr* R package (Coburn & Vevea,

2017), focusing our interpretation on how much the mean estimates vary across models, consistent with a recent meta-analysis on teacher practice (Garrett, Citkowicz, & Williams, 2019).

Although this "back-up" approach essentially ignores dependencies, this limitation will primarily affect the standard errors and *p* values, rather than mean estimates. For example, in the "metafor + clubSandwich" implementation of robust variance estimation (RVE) in our main analyses, RVE is exclusively used to adjust the standard errors and degrees of freedom for significance tests, not the parameter estimates themselves. Hence, interpreting mean estimates from this approach is defensible, even if the *p* values and likelihood ratio tests cannot be trusted. (Though some caution is still warranted because ignoring dependencies could affect parameter estimates via the relative weighting of effect sizes.)

As a comparison, we will also apply the aggregated effect size approach, given its common use, but we will place less weight on it when interpreting results for the reasons noted above. We will use the *MAd* R package (Del Re & Hoyt, 2014) to aggregate effect sizes, assuming a correlation of $r = .50$ among clusters of dependent effect sizes within samples, following Coles et al.'s (2019) analytic approach.

Meta-Regression Approaches: In contrast to selection models, meta-regression approaches such as PET and PEESE can easily accommodate RVE when assessing small-study effects. As with any other moderator, the standard error (PET) or variance (PEESE) can simply be included in a standard meta-regression model implementing RVE. At least three recent meta-analyses have applied this approach (Bediou et al., 2018; Coles et al., 2019; Friese, Frankenbach, Job, & Loschelder, 2017). The same point obviously applies to examining publication status (1 = published, 0 = unpublished) as a moderator. We will therefore adopt this approach because it explicitly addresses dependent effects and matches the estimation models for our main analyses.

However, one major drawback is that the statistical properties of this RVE implementation have not yet been extensively studied. We therefore will also apply the more common implementation of PET and PEESE: weighted least squares regression using aggregated article-level effects. The weights will be fixed-effects inverse variance weights ($1 / SE_i^2$), like Egger's regression, consistent with the implementation that was originally proposed and used in simulation studies (Carter et al., 2019; Stanley & Doucouliagos, 2014). However, it is not clear this approach will be superior to the RVE implementation because it relies on aggregating effects to the article level (also see Moreno et al., 2009 for the drawbacks of the fixed-effects weighting for this implementation). Hence, we will use this more common implementation as a comparison approach, but we will place less weight on it when interpreting results.

Adjusting Moderator Analyses

Limitations of Current Methods: The methods described thus far have exclusively focused on the role of overall publication bias in biasing mean estimates. However, as discussed in the *Hypotheses About Publication Bias* section, patterns of publication bias could vary based on study characteristics, potentially biasing the moderator analyses (see Coburn & Vevea, 2015 for discussion). We will therefore consider how selective reporting might bias our moderator analyses of children's age. We focus here on children's age because (a) this moderator is central to our theoretical goals and (b) publication bias could plausibly depend on how significant results are interpreted in the context of the sample's average age.

None of the statistical models detailed earlier allow for this possibility. For instance, although standard selection models technically allow for the inclusion of moderators, they only do so through the effect size model component, not the selection model component (i.e., the weight function; Coburn & Vevea, 2015). In other words, standard selection models assume that publication bias is solely driven by the *p* value, but not by how that *p* value is interpreted in the context of other study characteristics. The models can control for other study confounds that may appear to be overall publication bias, but they do not necessarily provide less biased estimates of moderator effects.

Selection Modeling: Addressing varying patterns of publication bias is complicated for selection models. However, Coburn's (2018) dissertation recently proposed an extension of the step function model that can accommodate moderators of publication bias. In its simplest form, Coburn's lambda model uses one additional parameter to model the relative likelihood of survival of nonsignificant results for one level of a moderator (e.g., older samples) versus another (e.g., younger samples). Her simulation results showed promising performance for this method in accounting for moderators of publication bias, including in simulations with moderate to large between-study heterogeneity (see Chapter 4 in the dissertation).

However, it is not immediately clear how the lambda model would extend to continuous moderators. Furthermore, the model is not currently available in the *weightr* package, but Coburn (2018) noted that an update is forthcoming (p. 131). Hence, we plan to use the lambda model if it becomes available by the time we conduct our analyses (or if we can otherwise obtain the needed R code to implement it). Otherwise, we plan to use the approach that Coburn and Vevea (2015) described of splitting our dataset by a categorical moderator (e.g., median split of the average age); conducting separate selection models for each subgroup of effect sizes; and examining if the pattern of mean estimates across moderator levels is consistent across unadjusted versus adjusted estimates.

Meta-Regression Approaches: Meta-regression approaches can more easily accommodate varying patterns of publication bias. Following Miller et al.'s (2018) example, we will modify the adjustment models by including interaction terms with children's age. For instance, in its simplest form, the fixed-effects part of the regression equation for the modified PET method can be represented as follows:

$$ES = \beta_0 + \beta_1 SE + \beta_2 Age + \beta_3 (SE * Age)$$

where *ES* is the effect size, *SE* is the effect size standard error, and *Age* is the sample's average age. The regression coefficient for β_2 is the adjusted age effect because it is the estimate when *SE* = 0, extrapolating to studies with infinite sample size. The interaction term (the β_3 coefficient) indicates if small-study effects (i.e., funnel plot asymmetry) vary by the average age of the sample; if significant, the term would provide suggestive evidence of varying patterns of publication bias. This approach also obviously extends to PEESE (replace *SE* with SE^2) and published-unpublished difference methods (replace *SE* with publication status dummy code). However, one major limitation is that the statistical properties of these interaction-based adjustments have not yet been studied to our knowledge.

Analytic Plan

This section combines the considerations from the previous sections to specify a concrete analysis plan for examining how selective reporting bias might impact our central results.

Adjustments to Mean Estimates: We will examine adjustments to overall mean estimates as a starting point for considering publication bias. We will prepare a table (or graph) like the one below that will show adjusted mean estimates across various model specifications. When interpreting results, we will give the greatest weight to selection models that include all dependent effects, for reasons detailed earlier. The multiplicity of possible bias adjustments emphasizes that these results should be viewed as sensitivity analyses, not as definitive estimates of "true" corrected effects.

Mean Estimates Across Adjustments for Selective Reporting

	All Effect Sizes*		Aggregated Effect Sizes	
Adjustment Model	Simple Adjustment	Add Moderators	Simple Adjustment	Add Moderators
Unadjusted				
None	?	?	?	?
Selection Models*				
p = .025	?*	?*	?	?
p = .025, .05	?*	?*	?	?
p = .025, .50	?*	?*	?	?
p = .025, .05, .50	?*	?*	?	?
Beta density	?*	?*	?	?
Small-Study Effects				
PET	?	?	?	?
PEESE	?	?	?	?
Publication Status				
Unpublished	?	?	?	?

Note. The *p* value cut points for the selection models correspond to one-tailed tests. A cut point of *p* = .025 denotes finding directionally consistent, significant results with a two-tailed test; *p* = .50 specifies the boundary for directional differences (e.g., pro-male vs. pro-female stereotypes). Publication status will be dummy coded (1 = published; 0 = unpublished), meaning that the regression intercept can be interpreted as the estimate for unpublished literature.

The "add moderators" column will control for differences in our confirmatory moderators in addition to dummy codes for the stereotype scale type. These moderators will be grand mean-centered so that the intercept can be interpreted as an overall mean estimate. Including moderators is important to control for confounds that might falsely appear as publication bias (e.g., larger studies may focus on smaller effects that are harder to detect in smaller studies).

*These adjustments will be given the greatest weight when interpreting results.

Adjustments to Moderator Analyses: The methods to adjust moderator analyses for selective reporting are far less developed than for adjusting mean estimates. We will therefore apply a similar set of procedures for adjusting moderator analyses of children's age but with some modifications. We will only consider simple adjustments, given the potential instability that could result from adjusting other moderators for selective reporting concurrently (e.g., for PET, including multiple interaction terms between the standard error and other moderators). Also, although the lambda model can provide adjusted estimates of moderator effects for step function selection models (Coburn, 2018), the beta density weight function model currently cannot (Citkowicz & Vevea, 2017). Assuming we can apply the lambda model to the continuous moderator of children's age, we will prepare a table like the one below:

Moderator Effects for Children's Age Across Adjustments for Selective Reporting

Adjustment Model	All Effect Sizes*	Aggregated Effects
Unadjusted		
None	?	?
Selection Models*		
p = .025	?*	?
p = .025, .05	?*	?
p = .025, .50	?*	?
p = .025, .05, .50	?*	?
Small-Study Effects		
PET	?	?
PEESE	?	?
Publication Status		
Unpublished	?	?

*These adjustments will be given the greatest weight when interpreting results.

Selective Reporting Bias Tests: All methods we plan to use also provide formal significance tests for the evidence of publication bias (including if it significantly varies as a function of the average age of the sample). These tests are based on likelihood ratio tests for selection models (Coburn, 2018) and individual parameter tests (e.g., regression coefficients for publication status) for the other methods. We will prepare a supplemental table (or graph) summarizing *p* values from these significance tests, but we will interpret them cautiously because they can be underpowered and likely may not be valid with dependent effects. Furthermore, these tests can only provide *suggestive* evidence of publication bias (e.g., small-study effects or unpublished-published differences could reflect other study confounds, though adding other moderators to the models will help rule out this possibility