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

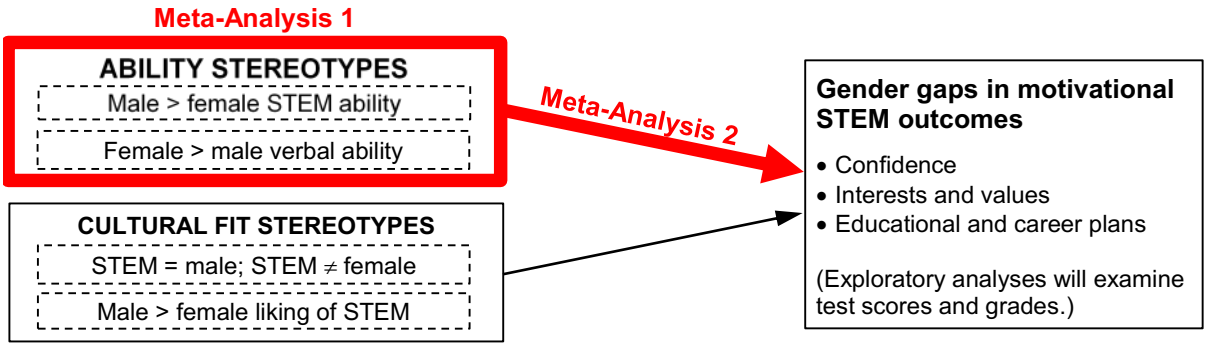
Ryan T. Williams
Courtney Tanenbaum
American Institutes for Research

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)



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 1) 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, 2009; 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 1).

Table 1. 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 1). 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 1, which is a log transformation of the raw POMP score that can range from $-\infty$ 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.

References

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

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

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.

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

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

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

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

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

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

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

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