Why Meta-Analyses of Growth Mindset and Other Interventions Should Follow Best Practices for Examining Heterogeneity

Preprint · November 2022

DOI: 10.13140/RG.2.2.34070.01605

6 authors, including:

Elizabeth Tipton
Northwestern University
72 PUBLICATIONS **6,432** CITATIONS
SEE PROFILE

Christopher J. Bryan
University of Texas at Austin
13 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Jared Scott Murray
University of Texas at Austin
45 PUBLICATIONS **1,433** CITATIONS
SEE PROFILE

Barbara lynn Schneider
Michigan State University
104 PUBLICATIONS **2,943** CITATIONS
SEE PROFILE

Why Meta-Analyses of Growth Mindset and Other Interventions Should Follow Best Practices for Examining Heterogeneity

Elizabeth Tipton, Department of Statistics and Data Science, Northwestern University
Christopher Bryan, Department of Business, Government, and Society, University of Texas at Austin
Jared Murray, Department of Statistics and Data Science, University of Texas at Austin
Mark McDaniel, Department of Psychology, University of St. Louis
Barbara Schneider, Department of Education and Sociology, Michigan State University
David S. Yeager, Department of Psychology, University of Texas at Austin

Revision
November 2022

Note: *This paper includes a re-analysis of data published in Macnamara and Burgoyne (in press). The reanalysis can be found at https://osf.io/mr3yx/.*

This research was supported by the National Science Foundation under award number 1761179 and the National Institutes of Health under award numbers R01HD084772 and P2CHD042849. Research reported in this publication was also supported by an Advanced Research Fellowship from the Jacobs Foundation to D. Yeager. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, the National Institutes of Health, or the Jacobs Foundation.

Abstract

Meta-analysts often ask a yes-or-no question: Is there an intervention effect or not? This traditional, all-or-nothing thinking stands in contrast with current best practice in meta-analysis, which calls for a heterogeneity-attuned approach (i.e., focused on the extent to which effects vary across procedures, participant groups, or contexts). This heterogeneity-attuned approach allows researchers to understand where effects are weaker or stronger and reveals mechanisms. The current article builds on a rare opportunity to compare two recent meta-analyses that examined the same literature (growth mindset interventions) but used different methods and reached different conclusions. One meta-analysis used a traditional approach (Macnamara and Burgoyne, in press), which aggregated effect sizes for each study before combining them and examined moderators one-by-one by splitting the data into small subgroups. The second meta-analysis (Burnette et al., in press) modeled the variation of effects within studies—across subgroups and outcomes—and applied modern, multi-level meta-regression methods. The former concluded that growth mindset effects are biased, but the latter yielded nuanced conclusions consistent with theoretical predictions. We explain why the practices followed by the latter meta-analysis were more in line with best practices for analyzing large and heterogeneous literatures. Further, an exploratory re-analysis of the data showed that applying the modern, heterogeneity-attuned methods from Burnette et al. (in press) to the dataset employed by Macnamara and Burgoyne (in press) confirmed Burnette et al.'s conclusions; namely, that there was a meaningful, significant effect of growth mindset in focal (at-risk) groups. This article concludes that heterogeneity-attuned meta-analysis is important both for advancing theory and for avoiding the boom-or-bust cycle that plagues too much of psychological science.

Keywords: meta-analysis, growth mindset, behavioral science, research integrity, heterogeneity

Why Meta-Analyses of Growth Mindset and Other Interventions Should Follow Best Practices for Examining Heterogeneity

In this article, we discuss two new meta-analyses of growth mindset interventions as an instructive case study that compares more traditional meta-analytic methods with newer, more sophisticated methods (Burnette et al., in press; Macnamara & Burgoyne, in press). In particular, we address the tendency for traditional meta-analyses to focus on the *average effect*, that is, to provide a summary judgment about a phenomenon. In contrast, more modern approaches focus on *heterogeneity of effects*, to build theories of mechanisms and boundary conditions—the who, when, where, and why of intervention effects. These newer approaches seek to capture the interactive and contextual nature of many, if not most, psychological phenomena (Bryan et al., 2019; Gelman, 2015; Jenkins, 1978; Kenny & Judd, 2019; Kitayama, 2017; Linden & Hönekopp, 2021; McDaniel & Butler, 2011; McShane et al., 2019; Miller, 2019; Rahwan et al., 2019; Stanley et al., 2018).

Growth mindset interventions provide a fascinating basis for this commentary for two reasons. First, growth mindset is a well-established phenomenon, having already been subjected to a large, national, confirmatory pre-registered replication with independent verification of the key finding: meaningful effects (relative to cost) for the focal group of lower-achieving students (Zhu et al., 2019). Second, there has been a wide variety of growth mindset intervention studies that include different procedures, populations, and contexts, yielding hundreds of effect sizes—which has produced an intriguing amount of unexplained heterogeneity. The two meta-analyses in question used very different approaches to analyzing that heterogeneity and they reached very different conclusions. The purpose of this commentary is to explain how these analyses reached

such different conclusions as a way of highlighting general lessons for the field of research synthesis.

To preview, Macnamara and Burgoyne's (in press) meta-analysis used a traditional approach that focused primarily—in the abstract and conclusion—on the overall average effect. They found a small effect size, leading them to conclude that the "apparent effects of growth mindset interventions on academic achievement are likely attributable to inadequate study design, reporting flaws, and bias." Their analyses also included estimation of the degree of heterogeneity—though the methods used greatly underestimated it—and they did not discuss or draw conclusions regarding it. Similarly, while they did also include some moderator analyses, these largely resulted in null findings. Meanwhile, Burnette and colleagues' (in press) meta-analysis used a modern, heterogeneity-attuned approach, which focused its analyses—and the reports of results in the abstract and conclusions—around quantifying and understanding the full distribution of effects, as well as the extent to which these could be explained by moderators. Their analyses showed, for example, that there was significant heterogeneity, and their analyses indicated a significant relationship between student risk status and intervention effect size, an effect that has emerged in both past studies and past meta-analyses.

Why did their conclusions differ? One reason might be that their datasets differed, because each review had different inclusion criteria. Although these differences in data may have been important, our article focuses on a second reason: Macnamara and Burgoyne (in press) used different conceptual frameworks, models, and methods than Burnette and colleagues. Our conclusion is justified in part by our own exploratory analysis, in which we applied the multi-level modeling syntax from Burnette and colleagues' meta-analysis to Macnamara and Burgoyne's data. When doing so, we found remarkably similar results to Burnette et al (in press).

For example, when using the same analytic methods (applied to the different data), we found a similar degree of heterogeneity and we found that effect sizes *did* vary as a function of student risk status, which persisted when adjusting for their measures of study quality or publication bias. In this article, we will provide more detail on these analyses and our findings. Overall, our focus is not on detailed comparisons of the findings of the two reviews. Instead, we use these comparisons to illustrate why traditional methods (and software) can make meta-analyses prone to certain oversights and prevent psychology from understanding our phenomena. We aim to show how to take heterogeneity seriously at every stage of the research synthesis process: research questions, effect size coding and modeling, moderator coding, and analysis of bias.

Why Meta-Analysis in the Social Sciences Should Focus on Heterogeneity

As readers of *Psychological Bulletin* know well, meta-analysis can play an important role in the development and testing of theories in psychology. This role is even more pronounced and essential for theories that have been researched extensively and are subject to substantial public interest. By conducting analyses with a focus on reducing bias (e.g., via clearly defined search methods, inclusion criteria, coding, and analysis strategies) these meta-analyses allow the state of the field to be assessed and summarized in ways that do not privilege the earliest studies, the most popular studies, or only the studies that are published in journals.

Psychology is a large field and its diversity of theories combined with a thorough search for mechanisms can often result in dozens of independent studies, each measuring multiple outcomes under different conditions. The resulting meta-analyses that synthesize these findings tend to be large because the interventions studied are not fixed or marketed programs but instead are theories that have been operationalized in multiple ways. This practice contrasts with meta-

analyses in medicine which—because they focus on very specific treatments (e.g., pharmaceuticals)—tend to be quite small (75% include fewer than 6 studies; Turner et al., 2012). The best practice methods for small and large meta-analyses differ. Their small size and strict inclusion criteria lead medical meta-analyses to focus their results on 'the' effect of a treatment (that is, the average effect)[1]. Larger meta-analyses, resulting from broader inclusion criteria—as are typical in the social sciences—however, allow for (and call for) this heterogeneity to be tested and explored. That is, by nature of the "heterogeneity in" (the procedures, populations, and contexts included), there is often profound "heterogeneity out" (variation in effect sizes). Thus, there is a *distribution* of treatment effects, which can be characterized by *both* its mean and a measure of its spread. For these reasons, with a research literature of this scope, by design the purpose of a meta-analysis is to understand and explain this variation.

Thus in the social sciences (and especially in psychology) the purpose of a meta-analysis—from beginning to end—should be to understand and elucidate not only the average effect but also variation in effects and the extent to which moderators can explain this variation. Researchers may hope for a distribution of effect sizes that only includes positive values and that vary tightly around a meaningful mean, but in practice, it is not this simple. Most effective interventions have beneficial effects in some conditions (e.g., certain procedures, populations, or contexts) but not in others. The point of social science meta-analysis is therefore to understand to what extent effects vary and, furthermore, to what extent this variation in effects can be explained and understood using moderators rooted in the theory of the intervention (as well as other things).

[1] The small sample sizes make it difficult to estimate well and test hypotheses about the degree of heterogeneity.

Unfortunately, this focus on understanding and explaining variation in effect sizes is not that common. Indeed, in a review of meta-analyses published in *Psychological Bulletin* in 2016, Tipton, Pustejovsky, and Ahmadi (2019b) showed that most articles did not implement key heterogeneity analysis methods (e.g., reporting all effect sizes rather than averaging; testing more than one moderator at a time in a meta-regression; Tipton et al. 2019b). Similarly, in a review of 150 meta-analyses in psychology, Linden and Hönekopp (2021) found that heterogeneity was quantified in only 29% of the studies. When reported in the meta-analyses, however, the heterogeneity was meaningful, indicating that effects range from trivial or even negative to strikingly large (Linden & Hönekopp, 2021). In such cases, the mean alone is not informative.

A recent exchange illustrates this trend rather clearly. Mertens and colleagues (2022) meta-analyzed the "nudge" literature (which involves low-cost interventions that change the 'choice architecture' [e.g., the way that information is presented] as a means for encouraging positive behavior) and focused their conclusions on a substantial average effect of $d = 0.45$. In a commentary, Maier and colleagues (2022) adjusted the same data for the possibility of publication bias (leading to $d = 0.00$ to 0.08) and concluded that "no evidence for the effectiveness of nudges remains." Both articles, however, were rooted what may be regarded as dichotomous, all-or-nothing thinking, neglecting the high degree of unexplained heterogeneity in effect sizes. In a comment by some of the present authors (Szaszi et al., 2022), we pointed out that the initial article by Merten et al. (2022) found that 95% of effects ranged from -0.92 to +1.08 — a tremendous amount of heterogeneity. Thus, the unexplained heterogeneity, not the average, was the heart of the story. As we explain below, the meta-analysis of growth mindset effects recently conducted by Macnamara and Burgoyne (in press) follows more of a traditional

"all or nothing" pattern, whereas the meta-analytic approach used by Burnette and colleagues (in press) allowed the authors to probe more deeply.

We are writing this article to call attention to the ways in which this traditional approach (and, as we explain, commonly used software) can limit one's ability to conduct a meta-analysis that interrogates the richness of heterogeneous effects. We expect that this article will be of interest to both those interested in growth mindset—those wanting to know more about the findings in this literature—and those who simply seek to gain a better understanding of meta-analytic best practices in psychology.

Our article proceeds as follows. First, we provide background and context on growth mindset and on the two meta-analyses. Next, we comment on four categories of best practices for heterogeneity-attuned meta-analyses, based upon consensus in the community of statisticians, methodologists, and experts in the field. These categories are *research questions, modeling variation in effect sizes, moderators,* and *adjusting for bias.* We compare Macnamara and Burgoyne's meta-analysis to Burnette and colleagues' as an illustration of why it is important to follow best practices. In the process, we report a simple exploratory analysis of Macnamara and Burgoyne's (in press) data using Burnette and colleagues' (in press) methods, which conform to best practices.[2] Note that our purpose in doing so is only to illustrate that following best-practices recommendations can, under some conditions, lead to very different conclusions from the exact

[2] For these analyses, we used a correlated, hierarchical effects meta-analysis that included all of the effect sizes available within each study found in Macnamara and Burgoyne's meta-analysis. Since the correlation between effect sizes in the same sample was unknown, we used robust variance estimation to guard against misspecification (see Pustejovsky & Tipton, 2022). When Macnamara and Burgoyne (in press) had calculated multiple effect sizes but excluded them from analyses, we included all calculated effect sizes (four studies). For six studies, Macnamara and Burgoyne included only one effect size, while Burnette et al. included several; in these cases, we borrowed the effect size information from Burnette et al. Further information—including the data and R code—can be found at https://osf.io/mr3yx/.

same dataset (although we do not claim to deliver a final verdict on parameter estimates). Finally, we conclude with a discussion of how to implement these best practices more broadly.

Background on Growth Mindset

Growth mindset interventions teach students the belief that people's abilities can be developed—for example, through effort, use of effective strategies, and appropriate help-seeking–and how to apply this belief in their classes (Hecht et al., 2021; Yeager et al., 2019; Yeager & Dweck, 2020). The interventions are often brief (<50 minutes) and low-cost (as low as 20 cents per child). Several randomized trials—including the pre-registered, U.S. nationally-representative, double-blind, experiment cited earlier (of which several of the authors of this article were affiliated), with data collected and analyzed by third parties—have found that a growth mindset intervention can improve the grades of lower-achieving, academically-at-risk, or low-income students (Outes-Leon et al., 2020; Yeager et al., 2019). As noted, an independent team of policy analysts from MDRC re-processed and analyzed all data from the mindset intervention, without interference from mindset researchers, and found the same results as the published article (see Zhu et al., 2019). This independent analysis addressed whether mindset was a false-positive result that only appeared when financially interested researchers were at the helm.

The effects of growth mindset may seem small relative to laboratory experimental effects in psychology, but they are meaningful relative to established benchmarks in the field of education (Kraft, 2020). A year of learning in 9th grade math tends to increase test scores by an average of 0.22 *SD* (Hill et al., 2008), and having a high-quality math teacher for a year is associated with an increase of 0.16 *SD* (Chetty et al., 2014). A year of intensive one-on-one

tutoring for high schoolers is 0.08 *SD* (in intent-to-treat analyses, Guryan et al., 2021). Thus, Kraft (2020) concluded that "effects of 0.15 or even 0.10 *SD* should be considered large and impressive" (p. 248) if the intervention is scalable, rigorously evaluated, and shown to improve consequential, objective outcomes (e.g., grades). Given common effect sizes in education, it is impressive that brief online growth mindset interventions have tended to improve the grades of lower-achieving high school students by 0.11 *SD* (Yeager et al., 2019; see also Paunesku et al., 2015; Yeager, Romero, et al., 2016), even though they do not teach any academic content (e.g., math or reading). Furthermore, heterogeneity analyses, guided by pre-registered analysis plans and conservative Bayesian modeling, have found school contexts where the effects approach 0.15 or even 0.20 *SD*, specifically, those contexts with mindset-supportive cultures (Yeager et al., 2019, 2022; see also Hecht et al., 2021).

This prior literature indicates that growth mindset *can* work under some conditions (Broda et al., 2018; Yeager et al., 2022), and *does not* work under others (Ganimian, 2020). Mindset researchers have now gone to considerable lengths to make clear that they do not believe growth mindset works everywhere, under all conditions; nor do they claim to explain a large share of the variance in grades or test scores (see Yeager & Dweck, 2020). Instead, the key claim from mindset theory in the literature is a claim of cost-effective and meaningful benefits for students who need it, provided they are in supportive contexts that allow them to put a growth mindset into practice (Hecht et al., 2021; Yeager & Dweck, 2020).

Meta-Analysis Best Practices

This article is not the first to focus on best practices in research synthesis and meta-analysis. For example, the *Handbook of Research Synthesis and Meta-Analysis* (Cooper et al.,

2019) provides chapters on best practices in all parts of systematic reviewing, from search procedures to effect size coding to analysis. Another good resource is a recent Editorial in *Psychological Bulletin* (Johnson, 2021) that provides a general overview of best practices in psychology. These best practices overlap in many regards with those in education research, which also produces large meta-analyses (Pigott & Polanin, 2020). Regarding statistical models and methods, Tipton, Pustejovsky & Ahmadi (2019a) provide a review of 40 years' worth of meta-regression methodological developments, distilling consensus points for the field. Finally, the PRISMA guidelines—standards for reporting—map onto best practices in psychology and the broader social sciences (Moher et al., 2015).

For the sake of brevity, we do not address the full range of best practices, instead focusing on four that are especially important in large meta-analyses of heterogeneous literatures. These include a variety of ways through which researchers conducting meta-analyses should address heterogeneity in psychology, including:

- *Research questions:* The purpose of a review should be to understand the distribution of effect sizes and the extent to which this variation in effects can be explained by existing theories and moderators.
- *Effect sizes:* A review should characterize not only the mean effect, but the distribution of effects, including the degree of heterogeneity. Analyses should include all relevant within-study effect sizes, and statistical models should appropriately account for the dependence structure of the data (e.g., using multilevel modeling).
- *Moderators:* Moderators should be planned with a focus on testing theory-driven hypotheses and tested simultaneously in meta-regression.

- *Adjusting for bias*: Reviews should address and adjust for sources of potential bias, including confounders and publication bias, and test them side-by-side with moderators in the same meta-regression model.

In the remainder of this section, we address each of these guidelines in order.

Research Questions

The first best practice starts with the research questions asked by a meta-analysis. Burnette et al. (in press) stated their research questions in a way that was aligned with both best practices and mindset theory. They stated that "heterogeneity in effects is expected for growth mindset interventions," and that "meta-analyses in fields with clear heterogeneity in outcomes should not try to deliver" a "simple verdict" on effects. Therefore, their research questions focused on "clarifying *for whom* these interventions work best." These questions led to analyses that yielded real insights that clarified the substantial heterogeneity in the field. The authors concluded that "effects are stronger to the degree that the analyses and/or interventions were targeted to focal groups" and highlight the role of implementation fidelity and context as well. Altogether, these findings from their research questions suggested that growth mindset is a promising intervention *in the right contexts* and for *the right students*.

In contrast, Macnamara and Burgoyne pre-registered research hypotheses (osf.io/ga9jk) that were less nuanced:

> "*Hypothesis 1a*: Benefits of growth mindset interventions on academic achievement are due to efficacious interventions.
>
> *Hypothesis 1b*: Perceived benefits of growth mindset interventions on academic achievement are largely due to poor design, analytical, reporting, or other practices."

These two hypotheses were meant to be at odds with one another, with the meta-analysis focused on proving *either* Hypothesis 1a or 1b to be true. Either growth mindset interventions work overall, or all effects in the literature are due to things like bias and low-quality methods. The dueling Hypotheses 1a versus 1b, as posed by the authors seek an answer in terms of a single number summary—the mean effect size, also referred to as 'the' effect or the 'summary' effect.

The reason that Macnamara and Burgoyne's (in press) formulation is problematic can be illustrated in Figure 1. It shows that the composition of a sample with respect to any given moderator necessarily changes the meaning of the *average* of a truly heterogeneous effect. Suppose a study intentionally over-samples from participant populations or contexts that are expected to have null effects (as was done in large growth mindset studies; Yeager et al., 2019), so that the study has the necessary power to detect interaction effects. In that case, the average in the whole sample would be lower because it combined a meaningful effect in one subgroup with predicted a null effect in another group that was included in the study specifically to provide a contrast with the group in which an effect was anticipated (Tipton, Yeager, et al., 2019). In many ways, a study that oversamples from sub-groups that theory predicts will show null effects is a more informative study because it is designed to reveal important group differences rather than simply to document an average effect that glosses over meaningful moderation. Yet, a meta-analysis that ignores this meaningful heterogeneity and focuses only on the sample-wide average can give the misleading impression that the phenomenon is weak or overclaimed (See Figure 1, Panel C).

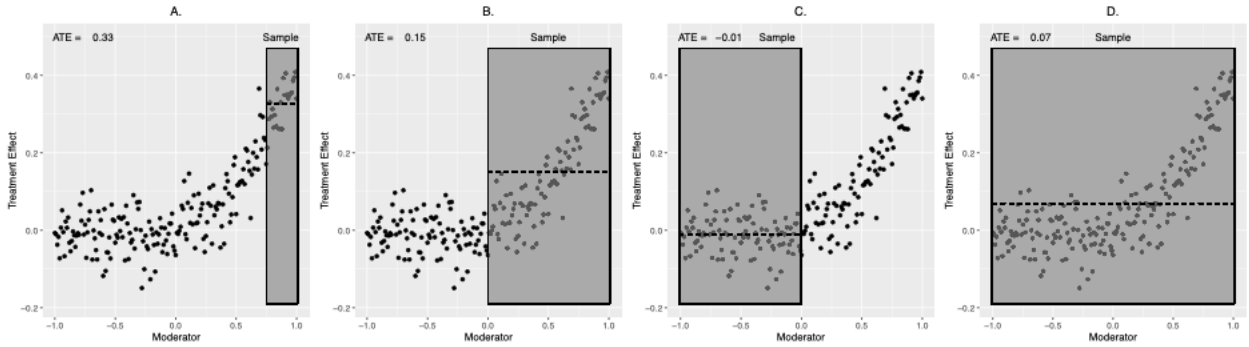

Figure 1. Hypothetical data showing the effects of a promising but heterogeneous intervention, such as growth mindset, by moderator and sampling type: (A) Initial promising effect in a hand-picked sample. (B) Replication study conducted in larger sample expected to show the effect, such as low-achieving students. (C) Replication study conducted in sample expected *not* to show the effect, such as high-achieving students. (D) Nationally-representative sample that includes all students regardless of prior achievement. Note: Dots are the expected effect sizes at each level of a moderator. Gray windows are the part of the sample included in a given study's sample or meta-analysis's sample. The dashed line is the average treatment effect (ATE) in a given sample. This figure is meant to illustrate the problems with a narrow focus on the ATE when there is true moderation. We may over-interpret a study's effect size if it is conducted in a small slice of the population (panel A). It also shows the problems with conducting a large sample study in a group of students who are not expected to benefit (panel C) and presuming that, because of the large sample size, the average effect in the sample is more informative with respect to the population average. Finally, the figure illustrates the limitations of averaging together heterogeneous effects (e.g., Panels B and C), as in panel D, and presuming that the effect is homogeneous. It would be far better to have all of the dots in the figures and model them using modern meta-regression analysis, as we explain below. (Figure reproduced with permission from (Bryan et al., 2021).

In sum, to make a contribution to the field's understanding of growth mindset effects, Macnamara and Burgoyne needed to ask research questions aligned with the goals of a heterogeneity-attuned meta-analysis (McShane & Böckenholt, 2018; Tipton, Pustejovsky, et al., 2019b), as Burnette and colleagues' (in press) meta-analysis did.

Characterize the Distribution of Effect Sizes, Including Variation

Meta-analyses should prominently quantify the heterogeneity in effect sizes. In literatures that are known to be heterogeneous—because of variation in the populations, interventions, outcomes and so on—one of the primary results should be the quantification of the heterogeneity of effects. This goal can be found in both Macnamara and Burgoyne's and Burnette et al.'s meta-analyses, which used broad inclusion criteria, having included different

growth mindset interventions, age groups (children, adolescents, and adults), and outcomes (test scores, grades, and more; see Tables 2, 4, and 5 Macnamara and Burgoyne, in press).

When there is heterogeneity, a random effects model is appropriate. This model assumes that there is a distribution of effect sizes, which can be characterized by both the mean effect size (μ) and a measure of the variation in effects (τ^2). If we can assume, as is common, that the true effects of the intervention are normally distributed, then 95% of the true effects fall into the interval $\mu \pm 1.96\tau$ (called a "95% prediction interval"). Both Macnamara and Burgoyne (in press) and Burnette et al. (in press) modeled their data using random effects meta-analysis, consistent with convention. But the authors differed in whether they prominently displayed the heterogeneity of effects.

Burnette et al. (in press) report an overall mean effect of 0.09 *SD* and an effect among targeted (at-risk) students of 0.16 *SD,* consistent with noteworthy effects according to conventions in education, as noted earlier (Kraft, 2020). Further, they prominently report a 95% prediction interval of effects, ranging from -0.08 *SD* to 0.35 *SD,* in the abstract and in the text. Importantly, these intervals are 95% *prediction* intervals, not *confidence* intervals. That is, these intervals are not about how precisely the mean treatment effect is estimated, but instead about the variation in true effects across studies.

Macnamara and Burgoyne used a different, more traditional approach. In the abstract and the article, they report an average effect of 0.05 *SD* overall, and smaller effects (0.03 *SD* and 0.02 *SD*) in subgroup analyses (which they call Meta-analysis 2 and 3). In a footnote, the authors reported the information needed to calculate their 95% prediction intervals (e.g., $0.05 \pm 1.96\sqrt{0.005} = 0.05 \pm 0.14 = (-0.09,\ 0.19)$). although, they did not calculate these themselves and they do not report the prediction interval in the abstract. Macnamara and

Burgoyne's intervals, however, indicate meaningful heterogeneity across a range of studies, subgroups, outcomes, and treatment versions. Further, as we show next, this range is truncated and masks larger positive effects because of how the authors analyzed studies that contributed multiple effect sizes.

Meta-analysis should include all the relevant within-study variation in effect sizes. Note that the variation and the means of the two meta-analyses differed, with Macnamara and Burgoyne's (in press) data indicating a tighter prediction interval. A primary reason for this difference is that these authors used traditional methods and software that excluded variation in effects *within* studies, while Burnette and colleagues (in press) included and modeled the within-study variation using a multilevel model with modern software. Specifically, Macnamara and Burgoyne used software that is commonly used for smaller meta-analyses—Comprehensive Meta Analyses (CMA) version 2 (Borenstein et al., 2006)—and that can only handle one effect size per study. The software forced the authors either to choose only one effect per study (e.g., the study's overall mean effect, or in a sensitivity analysis, only one subgroup effect) or allow the software to average the different effects from a study into one mean effect size. In contrast, Burnette and colleagues (in press) used a freely available software package in R, *metafor* (Viechtbauer & Viechtbauer, 2015), which allows for the inclusion of all of the relevant effects via statistical models that properly account for their dependence (see Pustejovsky & Tipton, 2022 for an overview of how to implement such analyses). This package allowed them to estimate two sources of heterogeneity—within (ω^2) and between study (τ^2) variation in effect sizes—which together account for the wider prediction intervals with larger positive effects. In order to guard against misspecification of this model, they also calculated standard errors and hypothesis tests

using robust standard errors (Hedges et al., 2010; Tipton, 2015; Tipton & Pustejovsky, 2015) implemented in the *clubSandwich* (Pustejovsky, 2017) package in R.

Not only does within-study variation give a more comprehensive view of the variation in effects, but it is also better for moderation analyses because it allows for other study features to be controlled. That is, it is a stronger comparison between effect sizes for low versus high-risk adolescents if the intervention and study team are held constant, and only the risk status varies within studies, rather than comparing two different studies with populations with different risk levels but using different materials, investigators, and so on. This approach is one of the benefits of including all effect sizes in the meta-analysis: It allows for better estimates of the effects of moderators.

An example to illustrate this point comes from another literature focused on anti-bullying programs (Yeager et al., 2015). Ttofi and Farrington (2011) meta-analyzed the literature on bullying using the traditional approach (one effect size per study) and reported that studies with older participants had larger effects than studies with young participants. However, this between-study moderation was contradicted by the within-study pattern in the individual studies, which in almost every case showed weaker effects when the same intervention was given to older students relative to younger students. Yeager et al. (2015) meta-analyzed the separate effect sizes for the studies and modeled them using multilevel meta-regression and found more predictable (on the basis of theory) patterns of moderation: In general, anti-bullying interventions are ineffective or even iatrogenic for older adolescents. This example shows why meta-analysis moderation tests that simplify complex within-study moderation by modeling only a single effect size per study can mask or even reverse true moderation results.

Like Ttofi and Farrington's (2011) meta-analysis of bullying interventions, the traditional approach used by Macnamara and Burgoyne (in press) can lead to illusory moderation results, because it averages over within-study variation that is potentially both theoretically informative and statistically meaningful. Consider the *National Study of Learning Mindsets* (NSLM), which was a double-blind RCT conducted in a nationally representative sample of 9th graders. It had an extensive pre-registered classical statistical analysis, a complementary conservative Bayesian analysis, and independent data collection and processing. The NSLM was designed in advance (see pre-registration: osf.io/tn6g4) to find beneficial effects on academic achievement in only *half* of the sample: those with low prior achievement. The primary reason is that 9th graders who already have straight "A" grades (as about 40% of 9th graders do) cannot get higher grades even if they have a growth mindset. Indeed, the NSLM found—as hypothesized in the pre-analysis plan and consistent with three previous replications (Paunesku et al., 2015; Yeager, Romero et al., 2016; Yeager, Walton et al., 2016), one of which was also pre-registered—effects for the subgroup of low-achieving students, and no statistically discernible effects for the subgroup of high-achieving students. Thus, by design, the NSLM included both subgroups in which the effect of growth mindset was hypothesized to be noteworthy and those in which it was hypothesized to be small or absent. Yet, analysts using CMA v2 would be forced to ignore this within-study heterogeneity.[3] Averaging a null effect with a meaningful and significant effect cuts the estimated effect size in half (see Panel D in Figure 1). The result is that a pre-registered, multiply

[3] Macnamara and Burgoyne (in press) claim that in the NSLM study "6,222 participants were missing from the published version" (p. 50). That is, they claim that the study did not report an important subgroup effect (high achiever) and did not report an interaction effect (comparing low to high-achieving subgroups). In fact, these subgroup and interaction effect analyses appeared in the Extended data Table 1, rows 8 to 12 and the data were not missing.

replicated effect size comes out looking trivial and non-significant, much like the erroneous conclusions from the Ttofi and Farrington (2011) meta-analysis.

Another example is the very large (>54,000 students, 799 schools), independent, randomized trial conducted in Peru by the World Bank (Outes-Leon et al., 2020). This study used a very low-dose approach: simply mailing schools a packet of paper-and-pencil growth mindset interventions and posters with supportive messaging to be displayed in classrooms (in a treatment group), or not (in a control group). The study, which cost less than 0.20 cents per student, yielded a range of effect sizes for different groups. Again, what Macnamara and Burgoyne (in press) included differed from what Burnette and colleagues (in press) included. For the high-poverty schools (at-risk) schools (>50% receiving government assistance), where student achievement was lower (See Outes-Leon et al., 2020, Table 4), Outes-Leon et al. (2020) report 5 effect sizes that range from 0.23 *SD* to 0.35 *SD*. For low-poverty and high-achieving schools, the 5 effect sizes were essentially zero (-0.02 to 0.02 *SD*). The traditional approach used by Macnamara and Burgoyne, does not include the ten varying effect sizes from this large study, but instead calls for only a single average effect for this country's very different populations: d = 0.02 (note that this is not an effect size that is reported in Outes-Leon et al.'s, 2020 working paper). The traditional approach, therefore, gives the impression that the intervention had no meaningful effect and was not moderated, when in fact it did have an important effect with the tens of thousands of students in precisely the group that mindset theory would predict.

We see similar issues for many other studies included in the Macnamara and Burgoyne (in press) meta-analysis. Several studies found evidence of moderation (Broda et al., 2018; Fink et al., 2018; Paunesku et al., 2015; Rienzo et al., 2015; Yeager, Romero, et al., 2016) but the traditional approach excluded the relevant interaction effects. As well, using the traditional

approach and CMA v2 meant that a 2×2 laboratory experiment in which one cell was supposed to make the mindset effect appear and another cell was supposed to make it disappear, was reported as just an average of the two cells (Wilson, 2009).

In general, the use of CMA v2 with a heterogeneous literature causes simplifications that can distort the overall conclusions and prevent analysts from accurately testing key moderators. To illustrate, we used the *metafor* package in R to explore what would happen if a multi-level meta-analysis method was applied to Macnamara and Burgoyne's (in press) data. This analysis included all the coded effect sizes, not study-level averages.[4] With Macnamara and Burgoyne's (in press) included studies, but with all 122 effect sizes, we found a mean growth mindset effect of 0.09 *SD, p* < .001 and an estimate of the standard deviation of true effects ($\sqrt{\tau^2 + \omega^2}$) that was twice as large ($\sqrt{\tau^2 + \omega^2}$ = 0.16 vs. $\sqrt{\tau^2 + \omega^2}$ = 0.07). Putting these together produced a 95% prediction interval of -0.22 to 0.40, very similar to Burnette and colleagues' estimate (-0.08 to 0.35 *SD*). Furthermore, this analysis indicated a mean effect among at-risk groups of 0.15 *SD*, *p* < .001, and a significant moderation of low versus high risk level of *B* = -0.08, *p* <.05, again consistent with findings from Burnette and colleagues. Overall, when analyzed using the same models—that appropriately accounted for all the effect sizes—Macnamara and Burgoyne's (in press) and Burnette and colleagues' (in press) data lead to strikingly similar results (Recall that we do not claim that this method is the only way to analyze the data and we recommend that readers re-analyze the data themselves; see syntax at https://osf.io/mr3yx/)).

[4] Here are the assumptions of our re-analysis. When Macnamara and Burgoyne reported a subgroup effect (which they analyzed separately), we simply included that subgroup effect nested within a study. When they reported only a mean but Burnette et al. (in press) reported the subgroups, or when there was a verifiable error in reporting (in one case), we took the estimates from Burnette et al. (in press). All decisions are explained in the posted spreadsheet at https://osf.io/mr3yx/.

Meta-Analyses Should Appropriately Adjust for Confounders, Including Study Quality and Publication Bias

Operationalizing study quality. Best practice for meta-analysis requires using validated measures of study quality. If ad hoc methods are used instead then analyses can be prone to two kinds of errors. First, the measure of study quality can be difficult to interpret because the measure does not capture the study quality construct. Second, measures of study quality that stray from validated frameworks can be prone to errors because there are no established methods to guide coders or to assist reviewers in evaluating the coding decisions. Macnamara and Burgoyne's (in press) meta-analysis illustrates both concerns.

There are many established frameworks for study quality (see the Equator Network: https://www.equator-network.org/), with the CONSORT-SPI reporting standards perhaps most relevant in psychology and education. These guidelines—developed and validated through a Delphi consensus process based on a large community of researchers—include guidelines related to how randomization should be reported (including type) and how methods and data should be reported. Other guidelines have also been proposed. For example, a series of comprehensive lists of threats to four validity types (internal, external, statistical conclusion, and construct) can be found in Shadish, Cook, and Campbell (2002). Chacon-Moscoso et al. (2016) provide a comprehensive review of study quality checklists in psychology and propose and validate a new, briefer one (of 12 items). And in education research, the What Works Clearinghouse Standards Handbook (NCES) provides standards for reporting, effect size computation, and indicators of study quality. These standards are based upon a consensus process including a panel of statisticians and methodologists in the field (one of us is involved in this work). In summary, there are several vetted and established criteria for evaluating study quality.

Instead of using or adapting one of these existing measures, Macnamara and Burgoyne's (in press) meta-analysis developed their own. Of the 10 standards coded by Macnamara and Burgoyne, five are not mentioned in the three major consensus-based standards for reporting and study quality. For one, Macnamara and Burgoyne's measure requires that *a priori* power analyses are reported, yet this requirement is not found in other guidelines. The CONSORT reporting guidelines do suggest inclusion of information regarding how a study sample size was determined (not a formal power analysis), but these guidelines are not uniformly implemented in journals. Another example is their requirement that high quality studies not only use random assignment, but also that this random assignment must be at the student level. In general, the guideline regarding the use of random assignment can be found in many existing standards for quality. However, none of these existing standards require random assignment to be at the student level. Instead, they require that the analysis conducted must be consistent with the study design and level of assignment. In school-based studies in education, in fact, group randomization is far more common than student-level randomization. Thus, experts in meta-analysis might not classify Outes-Leon et al. (2020) as "low quality" because it used school random assignment. Adapting to the context might instead be a sign of higher quality research.

Macnamara and Burgoyne's (in press) measure also includes a unique definition of manipulation checks. If a study manipulated growth mindset (versus control), and then measured growth or fixed mindsets at post-test, and showed that mindsets were different between the conditions, that study was coded as *not* having a manipulation check. A study was only coded as having a successful manipulation check if the study *measured mindset at baseline* and post-test, calculated a difference score, or controlled for baseline mindset, and then reported a test of the difference in change or post-test scores. This is not a standard definition of a manipulation check.

It resulted in several studies that did in fact have a post-test manipulation check (e.g., Aronson et al., 2002; Wilson, 2009) as being coded as though they did not have a manipulation check. Again, this ad hoc definition makes it hard to interpret the results.

Perhaps most curious is Macnamara and Burgoyne's (in press) measure's definition of financial conflicts of interest (FCOI). The CONSORT guidelines also mention FCOI, so a reasonable approach to coding FCOI in a meta-analysis would be to examine any FCOIs reported in each individual publication. Instead of the CONSORT definition Macnamara and Burgoyne defined FCOI as any *subsequent* financial success that an author experienced outside of academia following the publication of an article evaluating the effects of growth mindset interventions (e.g., paid talks, private consulting, proceeds from a popular press book). For example, a study co-authored by Dr. Mark McDaniel, a cognitive scientist who has not been involved with mindset research apart from one study, is listed as having an FCOI because McDaniel subsequently described mindset research for a few pages in a popular book he wrote about another topic. Macnamara and Burgoyne (in press) also state that Dr. Kasey Orvidas' role as a middle author on an article in 2018 could have biased the manuscript because in 2020 she formed a consulting company. The definition in the Macnamara and Burgoyne (in press) article means that an event that will happen in the future (e.g., consulting or giving a paid talk) is being interpreted as a causal explanation for an event that happened in the past (the publication of a mindset article). Again, this non-standard definition makes the findings difficult to interpret.

These examples show why it is important to have standardized, validated measures of study quality. They can lead to an overall mistaken impression about a literature. For example, Macnamara and Burgoyne claimed that very few studies met their criteria for 6 out of 10 best practices but fixing a few simple errors on their part—such as errors in coding pre-registrations,

manipulation checks, financial stakes, or power analyses—shows a different picture: the high-quality studies are the large, pre-registered, multi-site, team-science replications (see https://osf.io/mr3yx/).

One limitation of Burnette et al.'s (in press) meta-analysis is that it used a mix of standardized and non-standardized study quality criteria. Their study quality measures come in two categories: open science-related practices and RCT-related practices (e.g., internal validity). In favor of standardization, the authors score studies as higher in quality when they followed published open-science reporting standards (e.g., CONSORT), and several RCT-related study quality measures also appear in the WWC standards (e.g., baseline equivalence). But other quality measures are anachronistic, such as the requirement for pre-registration when studies' data were collected in the late 1990s or early 2000s, more than a decade before pre-registration was discussed in psychology (e.g., Aronson et al., 2002; Blackwell et al., 2007). Although the study quality standards in the Burnette et al. (in press) meta-analysis could be improved, they avoided claims and analyses that could have compromised their overall conclusions. They avoided all-or-nothing hypotheses about the effects being solely due to low-quality research (cf. Macnamara and Burgoyne's Hypothesis 1B) and they controlled for study quality continuously in multiple meta-regression analyses (discussed next).

Adjusting for study quality in analyses. How should analysts account for study quality? Here too the traditional approach that relies on the CMA v2 software can lead scholars away from best practices. Tipton et al. (2019a, 2019b) show that the consensus in the meta-analysis field is that methodological characteristics (including study quality) should be included as control variables in moderator analyses, consistent with best practices for the analysis of quasi-experimental data. The point is that a relationship observed between a focal variable (here a

moderator) and outcome (here an effect size) might be confounded with other features of the study design. Readers interested in this approach might turn to a tutorial by Tanner-Smith et al. (2017).

CMA v2 (unlike CMA v3 or the *metafor* package in R) only allows for moderator analyses with a single moderator at a time, making it impossible to include study quality as a confounder in analyses of other moderators. Therefore, analysts using the traditional approach must make choices about cut points for "high" versus "low" quality and divide the meta-analysis into smaller and smaller groups of effect sizes, as Macnamara and Burgoyne (in press) did. This dichotomization causes two problems. First, dichotomizing continuous measures is generally not a best practice in psychology and has not been for over 20 years (MacCallum et al., 2002) because analysts must make (at times, arbitrary) choices about the dividing line. Dichotomization can lead to false positive results or false negative results, depending on where analysts choose to put cases at or near the dividing line. Furthermore, dichotomization implies that all studies on one side of a dividing line are essentially the same. Thus, a study with 50% best practices is "low quality" just like a study with only 10% best practices, but a study with just one additional best practice—perhaps including a baseline measure of a manipulation check—is "high quality." This kind of decision can be difficult to defend. The second problem is that testing smaller and smaller subgroups reduces statistical power, and therefore makes subgroup average effects sensitive to small numbers of outliers, and masks moderator tests.

For an example of why this idea matters, consider Macnamara and Burgoyne's (in press) finding that in higher-quality studies, the mean effect of growth mindset interventions on outcomes was non-significant. To conduct this analysis, they first had to determine which studies were above a cut-point of their study quality measure, which required them to establish

this cut-point. But this cut-point choice was somewhat arbitrary, resulting in an underpowered test of the mean effect.[5] Note that this choice is not necessarily an indictment of the authors' work so much as it is a sign that dichotomization combined with subgroup analysis in general is not best practice because the results can hinge on a few arbitrary decisions. These problems with the traditional approach to modeling study quality are why experts recommend modeling the full, continuous measure of study quality in a meta-regression, even when the underlying quality measure follows validated and widely agreed-upon standards (as noted above).

Adjusting for publication bias. Findings that are not statistically significant have historically been harder to publish than those that are (Rosenthal, 1979). The traditional way to test for this publication bias is to examine if there is a relationship between the size of a study and the size of the effect (i.e., the standard error vs. the effect size). The logic of such analyses is that if larger studies have smaller effects (and small studies have larger effects), then this relationship provides circumstantial evidence that non-significant effects are "in the file drawer" due to publication bias. Macnamara and Burgoyne (in press) largely follow this traditional approach and conclude that the findings in the mindset literature are prone to publication bias. However, this traditional approach can lead to erroneous conclusions of publication bias when effect sizes are heterogeneous.

[5] The Macnamara and Burgoyne (in press) results by study quality were an artifact of one study being coded erroneously and then included in the high-quality group, and two other studies that were erroneously excluded. Specifically, a small study by Brougham and Kashubeck-West (2018) was incorrectly coded as having a power analysis when it did not (the authors said was that they aimed for "statistical power," without any power calculations), putting it in the "high quality" category. Several large studies with positive effects were incorrectly coded as not having a pre-registration, keeping them out of the "high quality" category. When we corrected these errors in Macnamara and Burgoyne's data and re-conducted the high-quality studies subgroup analysis using the *metafor* package, we found a significant overall effect of 0.05 *SD*, $p = .004$, and a significant effect for at-risk groups of 0.12 *SD*, $p = .01$—essentially identical to the results of the Burnette et al. (in press) meta-analysi and the NSLM study (Yeager et al., 2019). See https://osf.io/mr3yx/.

What is the traditional approach? The first statistical tool developed in the literature was a scatterplot of effect sizes (x-axis) versus standard errors (y-axis), called a *funnel plot.* Meta-analysts inspect the plot and look for "asymmetry," which is defined as more small studies (i.e., large-standard-error-studies) with large treatment effects, but fewer corresponding small studies with small treatment effects, as would be expected if effects vary randomly around a mean. Going beyond a visual test, authors can report Egger's test (Egger et al., 1997), which is a hypothesis test of the correlation between standard errors and effect sizes. Another approach is Trim-and-Fill (Duval & Tweedie, 2000), which imputes missing values until the funnel plot is symmetric, to estimate what the effect size might have been if publication bias had not been present. More recently, PET-PEESE analyses (Stanley & Doucouliagos, 2014) have been proposed to adjust effect sizes, relying on similar assumptions. These are the methods used by Macnamara and Burgoyne's (in press) meta-analysis, like many other articles in the literature.

The issue with all these traditional approaches to publication bias analysis, as suggested above, is that they can lead to false conclusions when they are applied to literatures that are truly heterogeneous. One prominent example comes from Open Science Collaboration (Open Science Collaboration, 2015) which replicated 100 studies from different subfields in psychology using new data collection. Because an independent team of scholars replicated the studies, and all effect sizes were published regardless of their significance, they knew that there was no publication bias. And yet an Egger's test of publication bias was significant ($z = 3.47$, $p < .001$), yielding an (impossible) finding of publication bias. This incorrect conclusion resulted because the effects were truly heterogeneous ($I^2 = 90\%$), and sources of heterogeneity of the sizes of the true underlying effects happened to be correlated with sample sizes[6].

[6] For example, between-subjects studies of personality variables in surveys used larger sample sizes and had smaller effects than within-subjects studies of cognitive psychology variables tested in laboratory settings with small sample

The same kind of finding can occur within a heterogeneous literature on a single phenomenon, such as growth mindset. For example, Macnamara and Burgoyne (in press) included in their meta-analysis a small laboratory experiment that assessed test performance at immediate post-test (Wilson, 2009), an early-stage intervention that involved personal attention from trainers over several weeks (Blackwell et al., 2007), and the large study in Peru (over 50,000 students) that examined test scores months later (Outes-Leon et al., 2020), among others. Just as in the Open Science Collaboration (2015) meta-study, sample size and effect size can be correlated for reasons other than publication bias. Larger studies might have less control over experimental procedures, they might examine outcomes that are less under the control of the researcher (such as standardized test scores), or their effects might wear off over time. Further, as in the NSLM and shown in Figure 1 (compare panels A and D), later-conducted studies might intentionally include populations where large effects are unlikely, to test for moderation, which would increase sample size but decrease effect size. In general, because the traditional method for assessing publication bias used by Macnamara and Burgoyne (in press) cannot distinguish file drawer effects from other factors, such as fadeout or heterogeneity in outcome variables, then it is of limited use for assessing bias in heterogeneous literatures.

What should scholars use instead? Detecting potential publication bias is an area of statistical methodology in meta-analysis that is developing rapidly. A review and benchmarking study by McShane and colleagues (2016) suggests that at that time the Vevea and Hedges selection model approach was stronger than the traditional methods noted above (Hedges & Vevea, 2005; Vevea & Hedges, 1995; Vevea & Woods, 2005). An advantage of the selection

sizes, which resulted in a correlation between effect size and sample size (Open Science Collaboration, 2015). This example illustrates why Egger's test will often be confounded by omitted variable bias: sample size is simply an observed variable, not a manipulated one, and so it can be a signal of many other factors besides putting null studies in the file drawer.

model approach is that users can examine different selection rules (p-value thresholds) and can implement these methods not only with the mean effect size but also with moderator analyses.[7] These are the methods used in Burnette et al.'s (in press) meta-analysis, but not Macnamara and Burgoyne's (in press) meta-analysis. As a result, Burnette and colleagues do not find evidence of publication bias.

Meta-analyses Should Seek to Explain Heterogeneity Using Moderation Analyses

Explanatory analyses that suggest causality need to account for confounders with multivariate meta-regression. Taking a step back, it is important to consider why we conduct moderator analyses in the first place. Usually, it is because we want to understand *mechanisms*. We want to know why the effects are appearing as they are, and in the case of interventions, how to design future interventions that produce positive effects appear more reliably. This attention to mechanism moves analysts from a descriptive focus (where one-variable models are sufficient) to an explanatory focus. And just as in original research, where psychologists should not make strong claims from a *t-test* with a single explanatory variable, meta-analysts should not make strong claims based on solitary dichotomized moderators. Instead, it is important to model the multiple competing moderators continuously in a meta-regression to account for potential confounding.

In regard to multiple moderators as well, the two meta-analyses used different tactics. The traditional approach, used by Macnamara and Burgoyne (in press), relied on CMA v2, which, as noted, only allows for a single moderator at a time and cannot model moderators continuously. As of 2016, this single-moderator approach was regrettably common in

[7] These methods can be implemented in R using the *weightr* package (Coburn & Vevea, 2015).

Psychological Bulletin and in the field in general (Tipton, Pustejovsky, & Ahmadi, 2019b). Burnett et al. (in press), in contrast, used the *metafor* and *clubSandwich* packages and included the competing moderators in the same model. When we re-analyzed Macnamara and Burgoyne's (in press) data using this approach and modeling the multiple moderators continuously, we found results that aligned with Burnette et al.'s (in press) findings: significant moderation by student at-risk group, and no significant moderation by FCOI status (see https://osf.io/mr3yx/). Thus, a heterogeneity-attuned analysis yielded a finding that is inconsistent with the Macnamara and Burgoyne (in press) conclusion that growth mindset effects for targeted groups are solely due to researcher bias or can only be obtained by the originators of mindset interventions.

Why was it important for the growth mindset meta-analyses to model moderators simultaneously? Because the moderators could be correlated. For example, authors who developed mindset interventions or measures and who worked on prominent studies might both understand the concept clearly (allowing them to develop high-quality interventions that are targeted to sensible populations) and might later be asked to talk about their results to public audiences (putting them in Macnamara and Burgoyne's FCOI category). As in an original study, a multiple meta-regression analysis can help to control for potential confounding.

Absence of evidence is not evidence of absence. Finally, we remind readers that in moderator analysis, just as in original research, the absence of evidence is not evidence for absence. Hypothesis tests focus on evaluating whether there is evidence to show that effect sizes are moderated by a given variable. These tests cannot be used to prove that the effect does *not* vary in general. Instead, evidence of variation in effect sizes comes from estimates and tests of the heterogeneity parameter themselves (e.g., the 95% prediction interval). It is entirely possible—and often likely—that all tests of moderators are not statistically significant and yet

there is considerable heterogeneity in effects. It could be that the moderators tested are not the right moderators either on a conceptual basis or because they were measured with too much error. It could also be that the effects of individual moderators are small—leading to underpowered tests (see Hedges & Pigott, 2004). The traditional approach employed by Macnamara and Burgoyne (in press) leads scholars to faulty conclusions when interpreting the results of their moderator analyses. For example, these authors conduct a variety of hypothesis tests and find that none are supported (ps > .05). They then conclude that the effect of mindset interventions is *not* moderated and thus interpret their small average effect size as if it were a constant effect—one that does not vary. And yet, as noted previously, there is substantial heterogeneity in effects. It is just not explained by their model. Burnette et al. (in press) do not make this mistake, perhaps because their approach allowed them to detect theoretically meaningful moderators that explained the heterogeneity. As noted, they found average mindset effects that were impressive compared to published standards in the field (Kraft, 2020), especially among focal (at-risk) groups of participants

Conclusion

Where does this commentary leave us with respect to the effect of interventions designed to promote a growth mindset? Large studies using extensive pre-registration and independent confirmatory analyses have already shown that growth mindset interventions can work (e.g., Zhu et al., 2019) and can produce meaningful and sizeable effects in terms of effect sizes for real-world educational outcomes that are determined and unfold over time. Although, as Burnette et al.'s (in press) review shows, growth mindset is not a magic bullet (see also Yeager & Walton, 2011). As hypothesized, it works better for some students than others, including students who

face academic struggles, as opposed to students who are already high performing. But this hypothesis is exactly the mark of a good theory, one that does not promise to solve everything, but sufficiently predicts for whom and under what conditions something works (Bryan, Tipton, & Yeager, 2021).

Moving beyond the specific case of growth mindset, the primary purpose of this article was to encourage the field of psychology to embrace modern meta-analytic techniques, ones that allow us to analyze heterogeneity. Instead of just focusing on mean effects and treating heterogeneity as a nuisance, we argue that given the variation in theories and interventions tested in the social sciences, the purpose of a meta-analysis should be to quantify and explain this heterogeneity, to help readers make sense of often conflicting findings, and to test hypotheses regarding the contexts and conditions under which theories hold and when they do not. These are complex questions and require complex analyses. There is not a single "effect" of interest.

We are not the first to alert the field to this goal or focus. Donald Rubin (1992), one of the founders of modern causal inference methods, wrote that a meta-analysis should not be "tied to the conceptualization of average effects, weighted or otherwise, in a population of studies." Thirty years later, these recommendations are echoed in the work of many other statisticians (Berlin, 1995; Gelman, 2014; McShane & Böckenholt, 2020; Rothman et al., 2008; Thompson, 1994; Tipton, Pustejovsky, et al., 2019a). These methods and approaches continue to be regrettably uncommon in meta-analyses in psychology, even in *Psychological Bulletin,* but Burnette et al. (in press) offer a promising example that follows many best-practices recommendations.

We are concerned that if the heterogeneity-naive approach persists unabated, there will be many more false conclusions about large areas of research–whether they are over-statements

of effect sizes or premature rejections of promising but heterogeneous effects. Both could lead to wasted scientific resources and could harm the credibility of the field[8]. Therefore, we hope that by spelling out these issues, we can encourage more meta-analysts to implement best practices more routinely, and ultimately produce a more robust and theoretically informative evidence base.

[8] Those conducting analyses and reviews may wonder what they should do next. We suggest an online resource library that may be of interest and use to the field: https://www.meta-analysis-learning-information-center.com/.

References

Berlin, J. A. (1995). Invited commentary: Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American Journal of Epidemiology*, *142*(4), 383–387.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2006). Comprehensive meta analysis (Version 2.2. 030). *Englewood, NJ: Biostat*.

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness*, *11*(3), 317–338. https://doi.org/10.1080/19345747.2018.1429037

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bryan, C. J., Yeager, D. S., & O'Brien, J. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1910951116

Chacón-Moscoso, S., Sanduvete-Chaves, S., & Sánchez-Martín, M. (2016). The development of a checklist to enhance methodological quality in intervention programs. *Frontiers in Psychology*, *7*, 1811.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

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

Fink, A., Cahill, M. J., McDaniel, M. A., Hoffman, A., & Frey, R. F. (2018). Improving general chemistry performance through a growth mindset intervention: Selective effects on underrepresented minorities. *Chemistry Education Research and Practice*, *19*(3), 783–806.

Ganimian, A. J. (2020). Growth-mindset interventions at scale: Experimental evidence from Argentina. *Educational Evaluation and Policy Analysis*, *42*(3), 417–438. https://doi.org/10.3102/0162373720938041

Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research a Bayesian perspective. *Journal of Management*, 0149206314525208. https://doi.org/10.1177/0149206314525208

Gelman, A. (2015). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management*, *41*(2), 632–643. https://doi.org/10.1177/0149206314525208

Guryan, J., Ludwig, J., Bhatt, M., Cook, P., Davis, J. M. V., Dodge, K., Farkas, G., Fryer, R., Mayer, S., Pollack, H., & Steinberg, L. (2021). *Not Too Late: Improving Academic*

Outcomes Among Adolescents (No. w28531; p. w28531). National Bureau of Economic Research. https://doi.org/10.3386/w28531

Hecht, C. A., Yeager, D. S., Dweck, C. S., & Murphy, M. C. (2021). Beliefs, affordances, and adolescent development: Lessons from a decade of growth mindset interventions. In *Advances in Child Development and Behavior* (Vol. 61, pp. 169–197). JAI. https://doi.org/10.1016/bs.acdb.2021.04.004

Hedges, L. V., & Pigott, T. D. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods*, *9*(4), 426–445. https://doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments*, 145–174.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Jenkins, J. J. (1978). Four points to remember: A tetrahedral model of memory experiments. *Levels of Processing in Human Memory*, 429–446.

Johnson, B. T. (2021). *Toward a more transparent, rigorous, and generative psychology.*

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Kitayama, S. (2017). Journal of Personality and Social Psychology: Attitudes and social cognition. *Journal of Personality and Social Psychology*, *112*(3), 357–360. https://doi.org/10.1037/pspa0000077

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science. *Perspectives on Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. pdh. https://doi.org/10.1037/1082-989X.7.1.19

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e2200300119.

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork*, 175–198.

Mcnamara, B. N., & Burgoyne, A. P. (in press). Do Growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices. *Psychological Bulletin*.

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1), 255–271.

McShane, B. B., & Böckenholt, U. (2020). Enriching Meta-Analytic Models of Summary Data: A Thought Experiment and Case Study. *Advances in Methods and Practices in Psychological Science*, *3*(1), 81–93. https://doi.org/10.1177/2515245919884304

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-Scale Replication Projects in Contemporary Psychological Research. *The American Statistician*, *73*(sup1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, *119*(1), e2107346118.

Miller, D. I. (2019). When Do Growth Mindset Interventions Work? *Trends in Cognitive Sciences*, *23*(11), 910–912. https://doi.org/10.1016/j.tics.2019.08.005

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1–9.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Outes-Leon, I., Sánchez, A., & Vakis, R. (2020). *The Power of Believing You can Get Smarter: The Impact of a Growth-Mindset Intervention on Academic Achievement in Peru*. The World Bank. https://doi.org/10.1596/1813-9450-9141

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784–793. https://doi.org/10.1177/0956797615571017

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46.

Pustejovsky, J. (2017). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. R package version 0.2. 3. *R Found. Stat. Comput., Vienna*.

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438.

Rahwan, Z., Yoeli, E., & Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature*, *575*(7782), 345–349. https://doi.org/10.1038/s41586-019-1741-y

Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). Changing mindsets: Evaluation report and executive summary. *Education Endowment Foundation*.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (Vol. 3). Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, *17*(4), 363–374.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Houghton Mifflin.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018a). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018b). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), e2200732119.

Thompson, S. G. (1994). Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *Bmj*, *309*(6965), 1351–1355.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375.

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.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Tipton, E., Yeager, D. S., Schneider, B., & Iachan, R. (2019). Designing probability samples to identify sources of treatment effect heterogeneity. In P. J. Lavrakas (Ed.), *Experimental*

Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment. Wiley.

Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., & Higgins, J. P. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology*, *41*(3), 818–827.

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.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

Viechtbauer, W., & Viechtbauer, M. W. (2015). Package 'metafor.' *The Comprehensive R Archive Network. Package 'Metafor'. Http://Cran. r-Project. Org/Web/Packages/Metafor/Metafor. Pdf*.

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., Muller, C., Murray, J., Mhatre, P., Kersting, N., Hulleman, C., Kudym, M., Murphy, M., Duckworth, A. L., Walton, G. M., & Dweck, C. S. (2022). Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychological Science*, *33*(1), 18–32. https://doi.org/10.1177/09567976211028984

Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, *75*(9), 1269–1284. https://doi.org/10.1037/amp0000794

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., …

Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H. Y., O'Brien, J., Flint, K., Roberts, A., Trott, J., Greene, D., Walton, G. M., & Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, *108*(3), 374–391. https://doi.org/10.1037/edu0000098

Yeager, D. S., & Tipton, E. (2022, November 7). Re-analysis of growth mindset meta-analysis using meta-regression methods. Retrieved from osf.io/mr3yx

Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., Kamentz, D., Ritter, G., Duckworth, A. L., Urstein, R., Gomez, E. M., Markus, H. R., Cohen, G. L., & Dweck, C. S. (2016). Teaching a lay theory before college narrows achievement gaps at scale. *Proceedings of the National Academy of Sciences*, *113*(24), E3341–E3348. https://doi.org/10.1073/pnas.1524360113

Zhu, P., Garcia, I., & Alonzo, E. (2019). An Independent Evaluation of the Growth Mindset Intervention. Issue Focus. *MDRC*.