Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1116

How Not to Fool Ourselves About Heterogeneity of Treatment Effects

Paul T. von Hippel
University of Texas, Austin

Brendan A. Schuetze
University of Utah

Researchers across many fields have called for greater attention to heterogeneity of treatment effects—shifting focus from the average effect to variation in effects between different treatments, studies, or subgroups. True heterogeneity is important, but many reports of heterogeneity have proved to be false, non-replicable, or exaggerated. In this review, we catalog ways that past researchers fooled themselves about heterogeneity, and recommend ways that we can stop fooling ourselves about heterogeneity in the future.

We make 18 specific recommendations and illustrate them with examples from education research. The most common themes are to (1) seek heterogeneity only when the mechanism offers clear motivation and the data offer adequate power, (2) shy away from seeking "no-but" heterogeneity when there is no main effect, (3) separate the noise of estimation error from the signal of true heterogeneity, (4) shrink variation in estimates toward zero, (5) increase p values and widen confidence intervals when conducting multiple tests, (6) estimate interactions rather than subgroup effects, and (7) check whether findings of heterogeneity are sensitive to changes in model or measurement. We also resolve longstanding debates about centering interactions in linear models and estimating interactions in nonlinear models such as logistic, ordinal, and interval regression. If researchers follow these recommendations, the search for heterogeneity should yield more trustworthy results in the future.

VERSION: January 2025

Suggested citation: von Hippel, Paul, and Brendan A. Schuetze. (2025). How Not to Fool Ourselves About Heterogeneity of Treatment Effects. (EdWorkingPaper: 25 -1116). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/4qtt-0w17

How Not to Fool Ourselves About Heterogeneity of Treatment Effects

Paul T. von Hippel
University of Texas, Austin, USA

Brendan A. Schuetze
University of Utah, USA

Draft of January 8, 2025

Abstract

Researchers across many fields have called for greater attention to *heterogeneity* of treatment effects—shifting focus from the average effect to variation in effects between different treatments, studies, or subgroups. True heterogeneity is important, but many reports of heterogeneity have proved to be false, non-replicable, or exaggerated. In this review, we catalog ways that past researchers fooled themselves about heterogeneity, and recommend steps to stop fooling ourselves about heterogeneity in the future.

We make 18 specific recommendations, which we illustrate with examples from education research. These are the most common themes: (1) seek heterogeneity only when the causal mechanism offers clear motivation and the data offer adequate power; (2) shy away from seeking "no-but" heterogeneity when there is no main effect; (3) separate the noise of estimation error from the signal of true heterogeneity; (4) shrink variation in estimates toward zero; (5) increase *p* values and widen confidence intervals when conducting multiple tests; (6) estimate interactions rather than subgroup effects; and (7) check whether findings of heterogeneity are sensitive to changes in model or measurement. We also resolve two longstanding debates: one about centering interactions in linear models and one about estimating and interpreting interactions in nonlinear models such as logistic, ordinal, and interval regression. Researchers who follow these recommendations will screen out many false, confusing, and non-replicable findings, and claims of heterogeneity that pass the screens should be clearer, more realistic, and more replicable.

Authors' note

ORCID IDs: 0000-0003-4498-4374 (von Hippel), 0000-0002-5210-6785 (Schuetze).
Changes in affiliation: Schuetze began this article as a graduate student at the University of Texas, Austin, USA, and continued as a postdoc at the University of Potsdam, Germany, before moving to the University of Utah as an assistant professor. Parts of this article overlap with Schuetze's dissertation.
Disclosures and acknowledgments: The authors did not seek funding for this research.
Contact: paulvonhippel@utexas.edu, brendan.schuetze@utah.edu

How Not to Fool Ourselves About Heterogeneity of Treatment Effects

In recent years, many scholars have called attention to *heterogeneity*—the idea that treatment effects vary across individuals, populations, contexts, and studies. While most research still focuses on estimating *main effects*, or how well treatments work on average, research on heterogeneity asks when, where, and for whom treatment works best or worst.

Enthusiasm for heterogeneity spans many disciplines (e.g., Bolger et al., 2019; McShane et al., 2019; Schudde, 2018). Leading researchers in statistics and social psychology have criticized a focus on main effects as "parochial" and stated that "behavioral science is unlikely to change the world without a heterogeneity revolution" (Bryan et al., 2021, pp.. 982-980). Conferences and thematic issues have urged contributors to look for heterogeneity in education, sociology, economics, psychopathology, and policy research (e.g., Brand et al., 2023; Damme & Mittal, 2023; Reardon & Stuart, 2017). Some medical researchers hope that a better understanding of heterogeneity will usher in an age of "personalized" or "precision medicine," in which each patient receives treatment tailored to their specific needs, instead of an off-the-rack treatment that only works on average (Kosorok & Laber, 2019). Likewise, some psychotherapy scholars have outlined a vision for a "personalized science of human improvement" (Hayes et al., 2022, p. 1), which tailors psychotherapy to an individual's personality, behaviors, goals, and needs.

The importance of true heterogeneity is undeniable. We need look no further than the recent COVID-19 pandemic for examples. Infection with the COVID-19 virus was most dangerous for adults who were older or suffered from certain pre-existing conditions. Knowing which groups were most vulnerable was vital for prioritizing access to vaccines when they first became available in limited supply (Persad et al., 2020). Some vaccines proved to be more effective than others (Self, 2021), and knowing that influenced patients and providers in deciding when to pursue an opportunity for vaccination and when to pass. Remote schooling during the pandemic depressed achievement most for children in high-poverty schools, and that was an important consideration when prioritizing federal aid intended to facilitate academic recovery (Goldhaber et al., 2022).

Yet for seven decades efforts to identify heterogeneity have yielded disappointing results. In a 1957 presidential address to the American Psychological Association (APA), the educational psychologist Lee Cronbach called for a kind of heterogeneity revolution. "Assigning everyone to the treatment with the highest average [response] is rarely the best decision," he wrote. "Ultimately, we should design treatments not to fit the average person, but to fit groups of students with particular aptitude patterns" (pp. 680-682). Cronbach charged psychologists to go forth and find "aptitude-treatment interactions," where aptitude was defined to include "innate differences, motivation, and past experience"—in short, as he later clarified, "any characteristic of the person that affects his response to the treatment" (Cronbach, 1975, p. 116).

By 1975, though, in another APA address, Cronbach reported that the search for "interactions did not turn out as we had anticipated" (p. 119). Few theorized interactions had materialized, and those that had materialized had triggered a replication crisis (Hedges & Schauer, 2018). "I have been thwarted by the inconsistent findings coming from roughly similar inquiries," Cronbach wrote. "Successive studies employing the same treatment variable find

different interactions" (Cronbach, 1975, p. 119). Cronbach conjectured that limiting inquiry to two-way interactions had been naive, speculating that three-way, four-way, and higher-order interactions might be needed to clarify the circumstances and populations where an intervention was or was not effective. Given the number of higher-order interactions that were possible, and the small sample sizes then available to test them, it seemed "unlikely that social scientists will be able to establish [reliable] generalizations" about moderated effects (Cronbach, 1975, abstract). The search for heterogeneity had led into a "hall of mirrors that extends to infinity" (Cronbach, 1975, p. 119).

From the 1970s to the early 1990s, "many authors…lamented the difficulty of detecting reliable moderator effects," and grew "so frustrated by not finding theorized moderator effects that they continue[d] to use and to invent inappropriate statistical procedures" that produced spurious findings of heterogeneity (McClelland & Judd, 1993, p. 377). Truly heterogeneous effects were described in the psychotherapy literature as "infrequent, undependable, and difficult to detect" (Smith & Sechrest, 1991, p. 233).

Even in recent years, as enthusiasm for heterogeneity has revived, some researchers have maintained a skeptical attitude. In organization research, a 2017 review article concluded that "the empirical track record of moderator variable studies is very discouraging," suggesting that scholars need to "mend it or end it" when it comes to searching for heterogeneity (Murphy & Russell, 2017, p. 549). Even when significant, most interactions in psychology are disappointingly small; on a standardized scale, the median interaction is only 0.05, meaning that the moderator usually changes the treatment effect by 0.05 SD or less (Aguinis et al., 2005; Baranger et al., 2022). In medicine, some reporters and scholars believe enthusiasm for precision medicine is premature, as less than 7 percent of cancer patients, for example, have the potential to benefit from available drugs tailored to their personal genome (Marquart et al., 2018; Prasad, 2016; Szabo, 2018). Shares of the consumer genomics company 23AndMe have fallen from over $300 to less than $5 after the company failed to provide value in the form of "personalized wellness plans" tailored to subscribers' genetic profile (Winkler, 2024).

The resurgence of interest in heterogeneity stems in part from concerns about the replication crisis. A substantial fraction of main effects fail to replicate when tested in new contexts and populations (Boulay et al., 2018; Ioannidis, 2005; Open Science Collaboration, 2015). Some failures to replicate main effects may result from inattention to *hidden moderators* which might explain why a treatment worked in one setting but not in another (Bryan et al., 2021, p. 981). For example, some social psychology findings replicate better among college students, where they were discovered, than in broader populations (Yeager et al., 2019), suggesting that effects are moderated by participants' age and education. Similarly, the facial feedback effect—the hypothesis that smiling, for example, makes the smiler happier—may replicate only when the experimenter is present to see the smiler's face, and not when the experimenter is absent (Phaf & Rotteveel, 2023). Despite these encouraging examples, a review of 68 meta-analyses of replication studies in psychology found limited evidence of heterogeneity across time, situations, and persons, leading the authors to describe their findings as "an argument against so-called 'hidden moderators'" (Olsson-Collentine et al., 2020, p. 936).

In fact, despite the hope that moderators might solve the replication crisis, moderated effects have proved to be less replicable than main effects. Across four large-scale efforts to replicate published results in psychology, economics, and social science, only one in five moderated effects (interactions) replicated successfully, vs. half of main effects (Altmejd et al., 2019; Open Science Collaboration, 2015). In a review of clinical trials, out of 117 subgroup

effects claimed in study abstracts, only 5 had been subjected to replication attempts, and none of those attempts successfully replicated the subgroup effect (Wallach et al., 2017). In short, while attention to heterogeneity may sometimes clear up the mystery of a non-replicable effect, to date it seems that the pursuit of moderators—at least as it has typically been conducted—may have made the replication crisis worse instead of better.

None of this means that heterogeneity is nonexistent, or impossible to detect. But it does suggest that modern researchers should be more careful than their predecessors when looking for heterogeneous effects.

In this article, we will diagnose major reasons for past disappointments and make 18 concrete recommendations that should reduce disappointments in the future. We will explain how so many past scholars have fooled themselves about heterogeneity, and recommend steps to ensure that we don't get fooled again.

Overview

Our plan for the paper is as follows. We will start by clarifying terminology in a section titled "What we talk about when we talk about heterogeneity." Among other things, our terminology will highlight the key distinction between heterogeneity that we can *explain* through treatment-moderator interactions, and heterogeneity that remains *unexplained* and simply represents variation in effects from one study, treatment, time, or place to another.

We then group 18 recommendations into three broad sections: "How not to fool ourselves about *unexplained* heterogeneity," "How not to fool ourselves about *explained* heterogeneity," and "How not to let models and measures fool us about heterogeneity." Our discussion of heterogeneity is not exhaustive, but we spotlight the most common pitfalls encountered in the search for heterogeneity, and we recommend ways to avoid them.

What we talk about when we talk about heterogeneity

Before we get started, we should explain what we mean by heterogeneity, since the word means different things in different settings.

Explained vs. unexplained heterogeneity

The most important distinction for our purposes is the distinction between *explained* and *unexplained* heterogeneity. These terms are meant to evoke the distinction between explained and unexplained variance in regression, where explained variance is associated with specific explanatory variables, and unexplained variance is not.

Explained heterogeneity

Many experimental and observational studies try to *explain* why effects vary by identifying specific *moderator* variables that predict larger or smaller effects. We call this *explained heterogeneity*; it is also called *moderation*. Attempts to explain heterogeneity often estimate *interactions* between the treatment and one or more moderator variables.

We explained heterogeneity in our introduction when we observed that the first COVID-19 vaccines had larger benefits for the elderly, and that pandemic school closures did greater harm to children from low-income families. In other words, age explained or moderated the effects of the vaccine, and income explained or moderated the effects of school closures.

Unexplained heterogeneity

Other studies estimate how much effects vary without trying to explain that variation with moderator variables. We call this *unexplained heterogeneity*.

One research design that highlights unexplained heterogeneity is *meta-analysis*, which combines the results of several studies to estimate an average treatment effect and the variance across studies that is due to heterogeneity—which in a simple meta-analysis means unexplained heterogeneity. In a more complex meta-analysis, researchers may try to explain heterogeneity by comparing studies with different characteristics. Moderator or subgroup analysis estimates the relationship between effect size and a single study characteristic, while meta-regression can model the relationship between effect sizes and multiple study characteristics simultaneously (Viswesvaran & Sanchez, 1998).

Another design that highlights unexplained heterogeneity is a *mega-study*, which assigns many different treatments, ideally at random, to different participants in a study population (Milkman, Gromet, et al., 2021). The goal of a mega-study is typically to identify and select the most effective treatments, without necessarily trying to explain why some treatments were more effective than others. For example, one mega-study tried to identify which of many text messages did the most to increase participants' vaccination rates (Milkman, Patel, et al., 2021), while another tried to identify which of many hints did the most to help students to solve math problems (Haim et al., 2022).

A popular type of mega-study is a *value-added study*, which highlights unexplained heterogeneity among teachers—trying to identify which of many teachers had the largest effects on student test scores, without necessarily trying to explain why. Because students are not assigned to teachers at random, value-added studies commonly control for prior test scores and other covariates (Staiger & Rockoff, 2010).

The bulk of heterogeneity is often unexplained

In many studies, the bulk of heterogeneity remains unexplained. After every available moderator has been considered, it often remains largely unclear why effects vary from one study, setting, or individual to another. For example, although age explained some heterogeneity in the benefits of the COVID vaccines, there were elderly individuals who died despite being vaccinated, or experienced minimal symptoms despite being unvaccinated. Likewise, although family income explained some of the heterogeneity in the learning effects of pandemic school closures, there were many high-income families whose children learned little during the pandemic. And although teachers clearly vary in effectiveness, only a small fraction of the variance can be explained by observed characteristics such as experience (Staiger & Rockoff, 2010).

Persistently unexplained heterogeneity is a fundamental challenge to any heterogeneity revolution. Although it is surely true that many treatments vary in their effects, it is often difficult to explain effect variation in terms of observed moderators—as we will see. And without knowing what variables moderate an effect, it is difficult to know which treatments will work best for which recipients.

Table 1. Types of Heterogeneity

Unexplained Heterogeneity	**vs.**	**Explained Heterogeneity**
Quantifies how much effects vary in total.		Links varying effects to characteristics of respondents, settings, or treatments.
Heterogeneity Within Studies	**vs.**	**Heterogeneity Between Studies**
Within a single study, effects may vary across respondents, contexts, treatments, or time-points.		Effects vary between different studies reviewed in a meta-analysis.
Heterogeneous Treatments	**vs.**	**Heterogeneous Responses**
Different participants receive different treatments.		Responses vary even though everyone receives exactly the same treatment..
Yes-and Heterogeneity	**vs.**	**No-but Heterogeneity**
There is a significant main effect, and the effect is larger or smaller for some subgroups.		There is no significant main effect, but there may be effects in a subgroup.
Heterogeneity Between Groups	**vs**.	**Heterogeneity Between Individuals**
Effects are different for some groups.		Effects are different for each individual.

Note. This table is not exhaustive but illustrates the breadth of approaches to describing heterogeneity in the social sciences.

Other distinctions in the heterogeneity literature

While our methodological recommendations are oriented primarily around explained vs. unexplained heterogeneity, Table 1 lists several other distinctions that help to clarify thinking about heterogeneity.

Heterogeneity within studies vs. heterogeneity between studies

One distinction is that heterogeneity can exist within or between studies. *Within-study heterogeneity* refers to a single study where different subgroups experience different effects (detected by interactions) or receive different treatments (as in a mega-study). *Between-study heterogeneity* is the province of meta-analysis, which compares effects across different studies of the same or similar treatments.

Between-study heterogeneity is often exploratory because it is hard to know whether the results of different studies differ because of observed moderators or unobserved confounders. For example, it might appear that effects obtained in different studies differ because of theoretically interesting variation in how the intervention was implemented or how responsive different subgroups were—but the differences might actually be due to methodological issues such as publication bias, how the outcome was measured, or how the data were analyzed (Holzmeister et al., 2024).

Within-study heterogeneity, by contrast, is easier to take at face value since researchers conducting a single study typically try to apply the same treatment, measure the same outcome, and make the same analytic decisions in different subgroups. On the other hand, estimates of within-study heterogeneity may not generalize beyond the study where they were observed unless participants (and stimuli) were sampled from a large and well-defined population (Tipton & Olsen, 2022; Yarkoni, 2022).

Heterogeneous treatments vs. heterogeneous responses

Another important distinction is between heterogeneous *treatments* and heterogeneous *responses*. When different units respond differently to treatment, it could be because the *treatments* they receive are not actually the same, or it could be because they are not equally sensitive and would respond differently even if treated identically.

For example, in a mental health intervention, some clients may receive more effective therapists than others, while other clients may respond more strongly to therapy. Or, in an intervention that reduces class size, some classes may shrink from 30 to 15 students, while others shrink from 25 to 20—but even at the same class size, some children may benefit more than others. The issue of heterogeneous treatments is important in the literature on *intervention fidelity*, which examines how well clinicians, educators, or others tasked with delivering an intervention actually carry out the treatment plan in practice. Slippage between the treatment plan and the treatment as actually delivered might explain some differences in efficacy across schools, clinics, or other settings (Gearing et al., 2011).

Yes-and vs. no-but heterogeneity: Heterogeneity with and without main effects

Another consideration is whether the main effect of the treatment is itself large enough to be practically important. When the main effect is large, then we call variation in treatment effects *yes-and heterogeneity*: Yes, the treatment usually works, and it works better for some groups than for others. The first COVID-19 vaccines were a clear example of yes-and heterogeneity. Nearly everyone benefited from vaccination, *and* the highest risk groups stood to benefit even more (Persad et al., 2020).

If the main effect is near zero, though, what we are left with is a search for *no-but* heterogeneity. No, the treatment usually doesn't work, but perhaps we can find a subgroup that benefits nonetheless. Lithium salts are an excellent example. Despite irresponsible health claims made in the first half of the twentieth century, lithium salts have *no* benefits for 99 percent of humanity, and in fact can be toxic at doses not far above the therapeutic dose. *But* among the 1 percent of the population that suffers from bipolar disorder, about two-thirds of sufferers benefit from lithium, and in some cases the benefit can be transformative or even life-saving (Brown, 2019).

There is some evidence that yes-and heterogeneity is more common than no-but heterogeneity—that is, that generally we should expect more heterogeneity when there is a large main effect. For example, in meta-analyses of direct replication studies in psychology, there is a strong correlation (0.66 to 0.91) between the average effect size and measures of unexplained heterogeneity (Olsson-Collentine et al., 2020).

Although examples of no-but heterogeneity certainly exist, their interpretation is not always as attractive as it is for lithium. If the average effect of treatment is near zero, at least one of the following statements must be true of any heterogeneity:

1. Any subgroup that benefits is small; and/or
2. The benefit experienced by any subgroup is small; and/or
3. The benefits to one subgroup are offset by harms to another.

Lithium checks box 1 because it benefits only 1 percent of the population, and checks box 3 because it can be toxic. Yet lithium remains an important treatment because it does not check box 2: that is, the benefit of lithium is not small. The benefit is substantial to a subgroup that can be readily identified by a diagnosis of bipolar disorder.

Other examples of no-but heterogeneity may be less attractive if they check all three boxes, or if the subgroup that benefits is difficult to identify in a replicable fashion—that is, if one subgroup appears to benefit in one study, but a similar subgroup does not benefit in another. We will return later to the problem of slippery subgroups.

Heterogeneity between individuals vs. heterogeneity between groups

Lastly, some studies estimate heterogeneity between individuals rather than heterogeneity between groups. Many statistical methods, such as interactions, estimate heterogeneity between defined groups, but within groups there may still be individual heterogeneity in response to treatment (McManus et al., 2023). Simply knowing that individual heterogeneity exists has limited value unless the heterogeneity can be explained, but the existence of individual heterogeneity is a sign that moderators are worth looking for.The search for individual heterogeneity often requires longitudinal data in which the same individuals are observed under both treatment and control (Golino et al., 2022; Hamaker, 2012; Molenaar, 2004; Molenaar & Campbell, 2009). Such data are often analyzed with multilevel models that distinguish variation at the level of groups, individuals, and measurement error (Singer & Willett, 2003; von Hippel et al., 2018).

We will not have much to say about individual heterogeneity except that it is still possible for researchers to fool themselves. For example, a recent article claiming individual heterogeneity in response to antidepressants was retracted and replaced with a reanalysis that found no clear evidence of individual heterogeneity (Maslej et al., 2020, 2021; Öngür & Bauchner, 2020).

What is a treatment?

Before discussing methods, we should clarify what we mean by a treatment. A treatment or intervention is any variable that has the potential to be manipulated (Holland, 1986) and might affect some outcome variable related to health, educational success, etc. We can estimate the causal effect of a treatment without bias if the treatment is not confounded with other variables that might affect the outcome. Random assignment ensures that treatments are not confounded, and so can some non-randomized designs—such as instrumental variables, regression discontinuity, and difference-in-differences—provided their assumptions are met (Angrist & Pischke, 2014). Other research designs, such as propensity score matching, only control for observed variables and cannot ensure that estimated treatment effects are free from confounding.

Our point in this article is that, even when a study design permits us to estimate treatment effects without bias, it is still possible to fool ourselves into seeing spurious heterogeneity between the effects observed in different groups.

How not to fool ourselves about unexplained heterogeneity

In this section, we will make five recommendations that can help us to avoid fooling ourselves about heterogeneity:

1. Decompose the variance of estimates into heterogeneity and estimation error, or signal and noise.
2. Graph the null distribution on your plot of estimated effects.

3. Correct for multiple inferences when testing hypotheses or estimating confidence intervals.
4. Shrink the estimated effects toward the mean effect using a simple empirical Bayes procedure.
5. Use large samples—not just overall but for each estimate.

We will motivate these recommendations with an example of unexplained heterogeneity, but the same recommendations apply to explained heterogeneity as well.

Motivating example

Figure 1a is a "caterpillar plot" that compares 92 Texas teacher preparation programs with respect to their teachers' effects on student reading scores (von Hippel, Bellows, et al., 2016). The point estimates are shown in ascending order, each with a 95 percent confidence interval. Test scores were standardized to a mean of 0 and an SD of 1, so an average program would have an effect of 0, and an effect of 0.1 would mean that a program's teachers raised student test scores by 0.1 SD. In the 2010s, 21 states used plots like this in an effort to identify the best and worst teacher preparation programs, with the goal of expanding the best programs and fixing or closing the worst (von Hippel & Bellows, 2018b).

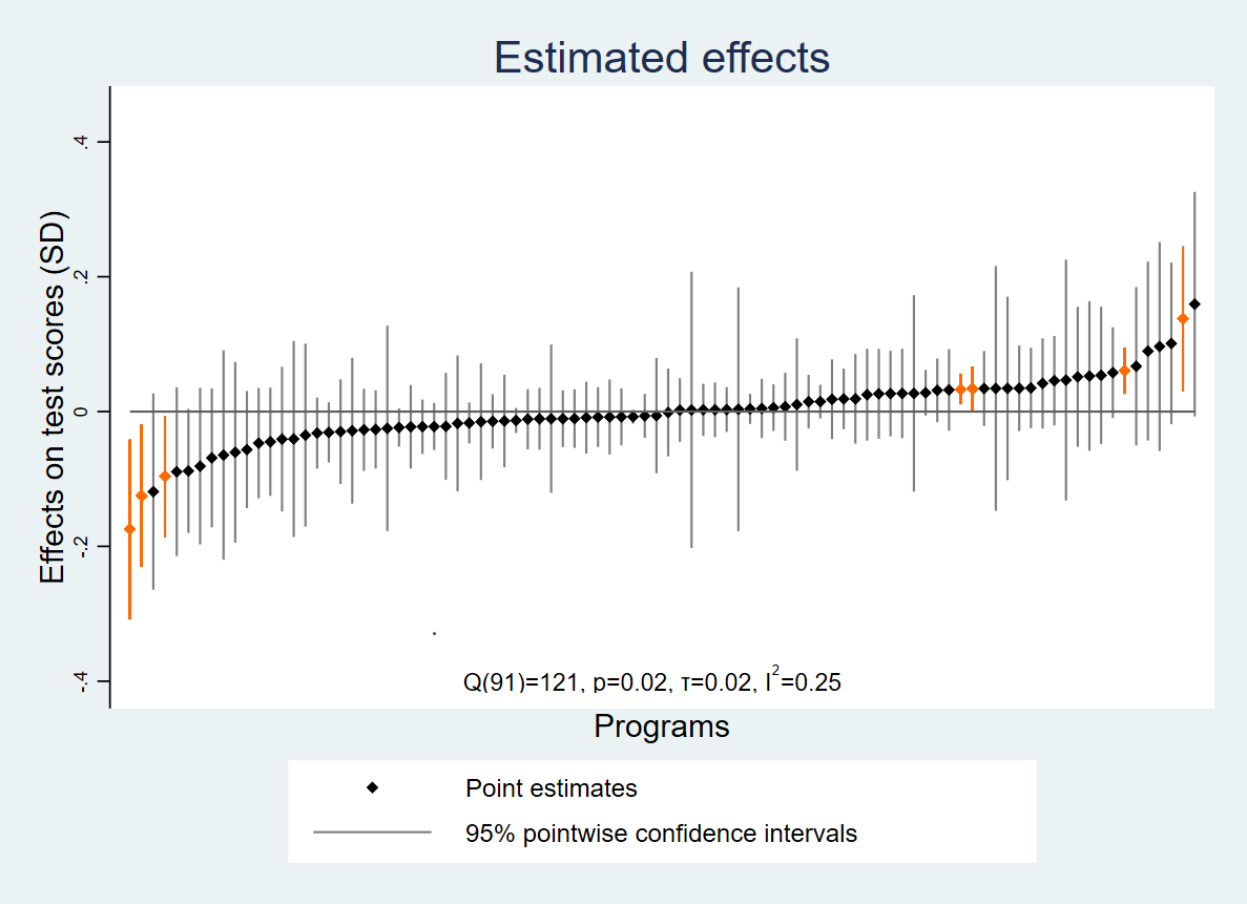

Figure 1a. Estimated effects on test scores of teachers from 92 teacher preparation programs. Effects of the orange programs differ significantly from zero.

Although this particular caterpillar plot compares 92 program effects, a similar plot could be used to compare 92 different treatments in a mega-study, 92 different subgroup-by-treatment interactions in an experiment, or 92 study results in a meta-analysis. (Meta-analyses often use a "forest plot," which does not order the estimates by effect size, but caterpillar plots are occasionally used as well.)

At first glance, it appears that there is substantial heterogeneity between programs. Compared to teachers from an average program, it appears that teachers from the "best" and "worst" programs can add or subtract nearly 0.2 SD from student test scores. Four programs appear to be significantly better than average (p<.05), and three appear to be significantly worse (p<.05). Programs that differ significantly from the average, highlighted in orange, have confidence intervals that do not cross zero.

Yet most of the differences in Figure 1a do not reflect true heterogeneity. The vast majority of programs do not really differ in their effects, and any true program differences are much smaller than they appear.

There are a few ways to avoid fooling yourself about unexplained heterogeneity.

Recommendation 1: Decompose the variance

One way to avoid fooling yourself is to decompose the variance of the estimates. Each program's true effect β is estimated with random error e, so that the estimated effect $\hat{\beta}$ consists of the true effect plus the error: $\hat{\beta} = \beta + e$. As a result, the *total variance V* of the estimated effects is larger than the variance of the true program effects, known as the *heterogeneity variance* τ^2. The excess variance is due to estimation error and is known as the *error variance* σ^2 . Here are three equivalent ways to write the formula for total variance:

$$Var(\hat{\beta}) = Var(\beta) + Var(e)$$
$$V = \tau^2 + \sigma^2$$
$$Total\ variance = Heterogeneity\ variance\ +\ Error\ variance$$

This formula is well-known in meta-analysis (Hedges & Olkin, 2014), but its implications are underappreciated in other settings. Whenever multiple effects are estimated with error, the estimated effects will vary more than the true effects (von Hippel & Bellows, 2018a, 2018b; von Hippel et al., 2016). This is true when we compare study results in a meta-analysis, but it is also a concern when we estimate multiple treatment effects in a mega-study, or multiple interactions in an experiment.

If the estimation error is substantial, as it is here, then the *heterogeneity variance*—the differences among the *true* effects—can be much smaller than the total variance—the differences among the *estimated* effects.

There are several statistics that shed light on the true extent of heterogeneity in a group of noisy estimates. All these statistics are commonly reported by meta-analysis software, but scholars looking for heterogeneity in other contexts are often unaware of them.

- Cochran's (1954) *Q* tests the null hypothesis of *homogeneity*—the null hypothesis that true effects do not vary at all. We reject the null hypothesis—and conclude that some heterogeneity is present—if if *Q* exceeds the critical value of a chi-square distribution with *df*=*k*-1 degrees of freedom, where *k* is the number of estimates,
- The I^2 statistic estimates what fraction of the variance in estimated study effects is due to true heterogeneity rather than estimation error. It is calculated as $I^2 = 1 - df/Q$

(Higgins & Thompson, 2002) but can be biased when the number of estimates is less than 10 (von Hippel, 2015).

- The heterogeneity variance τ^2 or heterogeneity SD τ represents how much effects vary from one study to another. There are several formulas for estimating τ^2 (Rukhin, 2013). The most transparent is just to subtract the error variance, estimated by the mean square of the standard errors, from the total variance of the estimates $Var(\hat{\beta})$.

Figure 1a reports statistics for the heterogeneity between our 92 teacher preparation programs:

- Cochran's (1954) *Q* is 121 which, with 91 degrees of freedom, allows us to reject the null hypothesis of homogeneity (p=0.02). So there probably is some heterogeneity among programs.
- But the I^2 statistic is just 0.25, suggesting that only one-quarter of the variance in Figure 1a is due to heterogeneity—true differences between programs. Three-quarters is due to estimation error.
- Finally, the heterogeneity SD τ is estimated to be just $\tau = 0.02$ student-level standard deviations. So contrary to the visual impression made by the graph, there are almost surely no programs with true effects as large as 0.2 student-level standard deviations. A program with an effect that large would be 10τ above average in the distribution of program effects—an almost impossibly rare event.[1]

In short, while there probably is some heterogeneity among programs, the heterogeneity variance is much smaller than you might guess from Figure 1a.

Recommendation 2: Plot the null distribution

Figure 1b shows visually how much estimation error contributes to the estimates. It compares the caterpillar plot to a curve representing the *null distribution*, which shows what the distribution of point estimates would look like under the null hypothesis that all 92 program effects were equal and only estimation error were present. This curve represents a mixture of 92 normal distributions, each representing the distribution of estimation error for one program estimate (von Hippel & Bellows, 2018a, 2018b; von Hippel et al., 2016).

The null distribution fits so well that we should emphasize it was not actually calculated from the estimated effects. It was calculated from the standard errors under the assumption that the true effects were all the same—i.e., that there was no heterogeneity. The fact that the null distribution is barely distinguishable from the empirical distribution of estimates confirms that little heterogeneity is present, and the bulk of the variance is due to estimation error.

The shape of the null distribution exposes a common misinterpretation of caterpillar plots. Looking at caterpillar plots like this one, investigators commonly imagine that the vast majority of programs are practically indistinguishable, but there are a few exceptional programs with larger positive or negative effects. It is easy to see where that idea comes from, since the middle of the plot is nearly flat while the tails flare. However, the null distribution has flaring tails as well, implying that flaring tails are not in themselves evidence of large outlying effects, but should be expected even when no heterogeneity is present. Only dispersion of the estimates

[1] No matter how the program effects are distributed, less than one percent of programs can be 10τ above or below the average. This follows from Chebyshev's (1867) inequality, which proves that for any distribution with a finite SD τ, the probability of an observation's being more than $k\tau$ from the mean, in either direction, cannot exceed $1/k^2$ and is typically much less.

beyond the tails of the null distribution should be taken as evidence of heterogeneity—and in Figure 1b that extra dispersion is very slight.

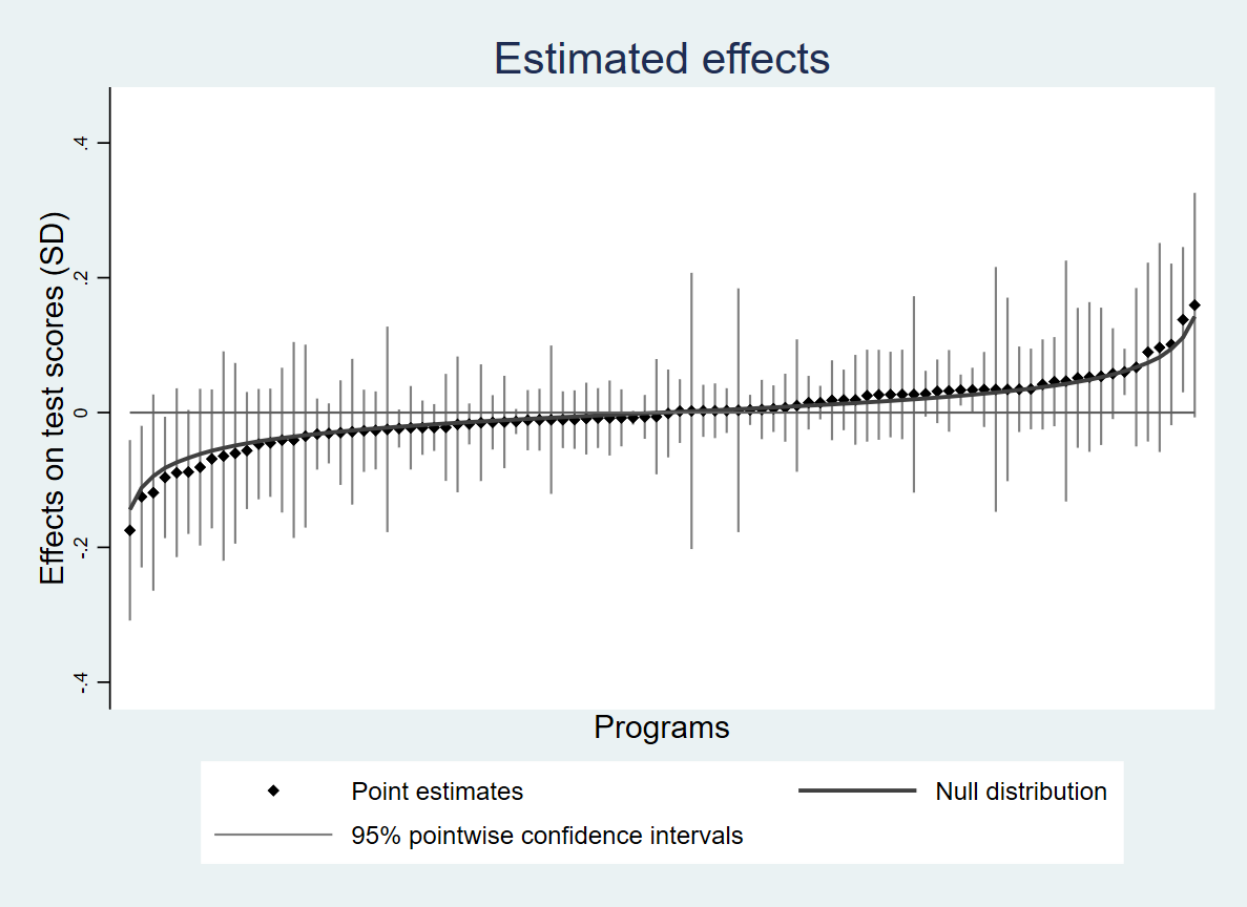

Figure 1b. The same estimates as in Figure 1a, but now compared to a null distribution showing how dispersed the estimates would be if there were no differences between programs. The actual estimates are barely more dispersed than the null distribution. The null distribution was calculated and drawn using the user-written Stata command ***caterpillar*** *(Bellows & von Hippel, 2017).*

Recommendation 3: Correct for multiple inferences to reduce false discoveries

Earlier we rejected the global null hypothesis that there was no heterogeneity among programs ($Q(91)=121$, p=.02). A common misconception is that, once the global null hypothesis has been rejected, any individually significant program effects can be taken seriously (Bloom & Michalopoulos, 2013). This is not correct. Even if we know heterogeneity is present, it may be challenging to identify the specific programs that are different.

In Figure 1a, we flagged 7 of 92 programs as differing significantly from the average, but most of these programs were likely *false discoveries*—average programs that happened by chance to produce statistically significant estimates (Benjamini & Hochberg, 1995). The fact that 7 estimates were significant does not mean that 7 programs were truly different. To the contrary,

since we tested 92 programs, each at a 5 percent significance level, we would expect an average of 4 or 5 significant results (5 percent of 92) even if all true program effects were the same.

The problem that we have just described is known as *multiple inferences* (Westfall et al., 2011), and there are several ways to correct it. Table 2a sorts the 92 *p* values from largest to smallest, then illustrates three simple corrections that correct the *p* values by increasing them:[2]

- The Bonferroni (1936) correction multiplies each *p* value by the number of inferences—here 92.
- The Holm (1979) procedure multiplies the smallest *p* value by 92, the next-smallest by 91, and so on.
- The Benjamini-Hochberg (BH) procedure multiplies the smallest *p* value by 92, the next-smallest by 92/2, the next smallest by 92/3, and so on.

After correction, we can report as significant any corrected *p* value that is less than a threshold value α, most often set to 5 percent. All three procedures round corrected *p* values down to 1, and the Holm and BH corrections also make adjustments to ensure that the corrected *p* values are still in descending order.[3]

Table 2a. Significance of program effects from Figure 1a, corrected for multiple tests

				Bonferroni correction		Holm correction		Benjamini-Hochberg correction	
Program	Estimate	SE	Uncorrected p	Multiplier	p	Multiplier	p	Multiplier	p
1	.002	.105	.9818	92	1	1	1	1	.982
2	.003	.092	.9706	92	1	2	1	92/91	.981
…	…	…	…	…	…	…	…		…
86	.034	.017	.0445	92	1	86	1	92/7	.585
87	-.096	.046	.0361	92	1	87	1	92/6	.554
88	-.125	.054	.0199	92	1	88	1	92/5	.366
89	.138	.055	.0122	92	1	89	1	92/4	.281
90	-.175	.068	.0105	92	.967	90	.946	92/3	.281
91	.033	.012	.0045	92	.413	91	.408	92/2	.207
92	.060	.018	.0006	92	.055	92	.055	92	.055

Note. After correcting for multiple tests, only one of 92 effects (with p=.055) approaches any conventional threshold for statistical significance. P values can be corrected using the p.adjust function in R or the ***qqvalue*** *command for Stata (Newsom, 2010).*

After any of the three corrections, only one of the seven significant differences in Figure 1a survived, with a borderline significant *p* value of 0.055. That program had an estimated effect

[2] Here we implement the corrections by adjusting the *p* values and comparing them to a fixed significance level. An equivalent approach is to leave the *p* values alone but adjust the significance level (Westfall et al., 2011).

[3] The Holm procedure "steps up," increasing a corrected *p* value if the *p* value below it is larger; the BH procedure "steps down," reducing a corrected *p* value if the *p* value above it is smaller. BH-corrected *p* values are sometimes called *q* values.

of just 0.06 SD, and even that was likely an overestimate, as we'll see a little later. All other programs had corrected *p* values greater than 0.2, providing little convincing evidence that the programs really differed from the average.

Although the three corrections led to similar conclusions in this example, they do have two differences:

- The corrections differ in their *power* to detect true treatment effects. The BH correction is most powerful, the Holm correction is next most powerful, and the Bonferroni correction is the least powerful.
- The corrections also differ in the *risk* that they control. The BH correction controls the *false discovery rate*—or the fraction of significant results that are false discoveries. The Bonferroni and Holm corrections, by contrast, control the *familywise error rate*—the probability of making *even a single false discovery*.

For example, if we report as significant any adjusted *p* value less than 0.05, then the BH correction ensures that only 5 percent of significant results will be false discoveries, while the Holm and Bonferroni corrections ensure that there is only a 5 percent chance that *any* of the significant effects is a false discovery.

Which correction should we use? The Bonferroni correction unnecessarily sacrifices power, so it is better to use the Holm or BH correction. Which correction to use depends on how many false discoveries you can tolerate. If you don't mind 5 percent of discoveries being false, use the BH correction. If you want to limit the risk of making *any* false discoveries, use the Holm correction.

Routine correction for multiple inferences, using either the Holm or the BH correction, would go a long way toward reducing reports of spurious heterogeneity.

Table 2b. Comparing corrections for multiple inferences.

Correction	Power to detect true effects	Risk controlled at a 5% significance level
Benjamini-Hochberg	High	*False discovery rate.* 5% of significant results will be false discoveries
Holm	Lower	*Familywise error rate*. 5% risk that *even one* significant result will be a false discovery
Bonferroni	Lowest	

Corrections don't just increase *p* values. They also widen confidence intervals.

A common misconception is that multiple inferences are only a problem if we rely on *p* values to test hypotheses. Some literature may unwittingly invite this misunderstanding by using the terms *multiple tests* or *p hacking*, which we have avoided in favor of the broader terms *multiple inferences* or *multiplicity*.

If the problem of multiple inferences were limited to *p* values, then we could sidestep the problem by relying instead on confidence intervals. Unfortunately, confidence intervals do not obviate the problem of multiple inferences. In fact, there is a one-to-one correspondence between

confidence intervals and hypothesis tests: rejecting the null hypothesis at $p<.05$ is equivalent to saying that the null hypothesis is not inside a 95 percent confidence interval. In Figure 1a, for example, the 7 programs whose uncorrected *p* values are less than 0.05 are exactly the 7 programs whose 95 percent confidence intervals do not cover 0.

Correcting hypothesis tests is equivalent to widening confidence intervals. Returning to our example of comparing 92 teacher preparation programs, Figure 1c shows uncorrected 95 percent confidence intervals, along with Bonferroni 95 percent confidence intervals, which are wider because they correct for 92 inferences. Each uncorrected has a 95 percent chance of covering its program's true effect—but that means there is only a 1 percent chance that all 92 intervals cover all 92 effects ($.95^{92}=.01$). By contrast, the Bonferroni intervals ensure at least a 95 percent probability that all the intervals cover all the program effects.

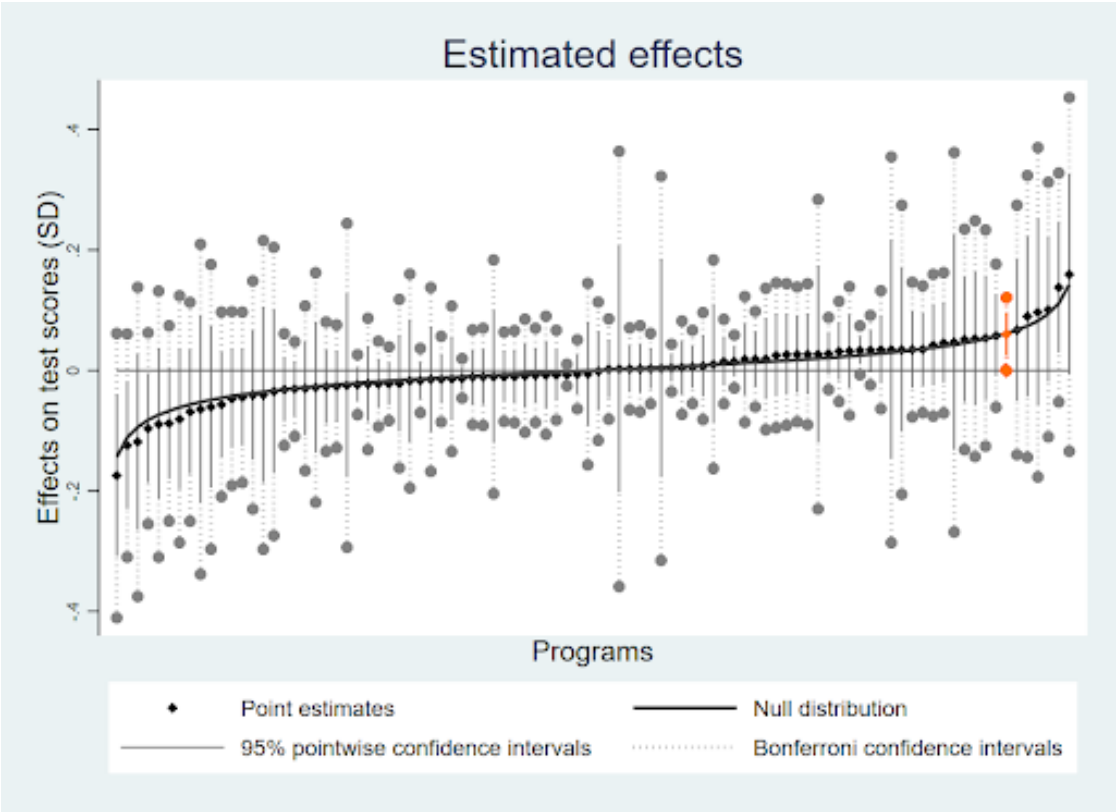

Figure 1c. The same estimates as in Figures 1a and b, but with confidence intervals lengthened by a Bonferroni correction for multiple inferences. After correction, only a single program, highlighted in orange, comes close to differing significantly from the average (p=0.055).

Alternatives to multiple test corrections: Cross-validation and lower significance levels

The Holm and BH corrections have excellent properties, but both assume that we know how many hypotheses were tested. In our example, there were clearly 92 hypotheses, but in other settings it can be hard to know how many hypotheses were tested. For example, an investigator might test 92 hypotheses but not report all of them.

A simple approach is to lower the conventional threshold for statistical significance from 0.05 to 0.005 (Benjamin et al., 2017). This adjustment is arbitrary but easy to implement. It is equivalent to a Bonferroni correction for 10 inferences.

Another approach is to check whether evidence for heterogeneity replicates in a different sample (De Rooij & Weeda, 2020). Such out-of-sample validation is routine in genomics (Hewitt, 2012; Johnston et al., 2013), in machine learning (Burzykowski et al., 2023), and in business applications such as credit scoring (Mushava & Murray, 2024).The replication sample can come from the same population, but sampling from a different population allows investigators to check whether findings generalize beyond the original population.

For example, in our evaluation of teacher preparation programs, we estimated program effects for each grade level—4th, 5th, 6th etc. (von Hippel et al., 2016). The correlation between 4th and 5th grade estimates of the same program's effect was only 0 to 0.4, confirming that estimates were noisy and that most of the variance in estimates lay between 4th and 5th grade estimates of the same program's effect. When this within-program variance was subtracted, the remaining heterogeneity between programs was only 0 to 0.04 SD in student test scores.

Recommendation 4: Shrink estimated effects

By now you may be persuaded that only one estimated effect, at most, differs significantly from zero, and that program only differs from the average by 0.06 SD—and yet Figures 1a-c give the visual impression that some programs have effects of nearly 0.2 SD. That is because Figures 1a-c show the *estimated* effects, whose variance is inflated by estimation error. The *true* effects are much smaller.

Figure 1d estimates the true effects by ***shrinking*** the estimated effects toward the mean (toward zero in this example). The point estimates $\hat{\beta}$ were shrunk toward 0 by a factor of

$$w = \frac{\hat{\tau}^2}{\hat{\tau}^2 + s^2}$$

where $\hat{\tau}$ is an estimate of the heterogeneity SD and s is an estimate of $\hat{\beta}$'s standard error. These shrunken estimates are variously known as empirical Bayes estimates (Morris, 1983) with a normal prior, or James-Stein (1961) estimates, or best linear unbiased predictions (BLUPs) in random effects models (Robinson, 1991).

After shrinkage, the point estimates range from -0.02 to +0.02 SD. Standard errors also shrink, but by a factor of $\sqrt{w}$ rather than w. Because the standard errors shrink less than the point estimates, fewer shrunken estimates differ significantly from 0. In this example, there were 7 significant effects before shrinkage, but only 2 were significant after shrinkage (both with p=0.03).

The shrunken estimates may seem absurdly small, and indeed they are somewhat biased toward zero (von Hippel et al., 2016).[4] Nevertheless, shrinkage brings the estimated effects closer to the true effects in the sense that the shrunken estimates minimize the mean of the squared estimation errors (Morris, 1983; Robinson, 1991; James & Stein, 1961).

Notice that estimates shrink more if their standard error s is larger. That is, shrinkage can be seen as a way to remove the excess variance that comes from estimation error.

[4] Because the shrunken estimates are biased, the expression "best linear *unbiased* prediction" (BLUP) is somewhat misleading.

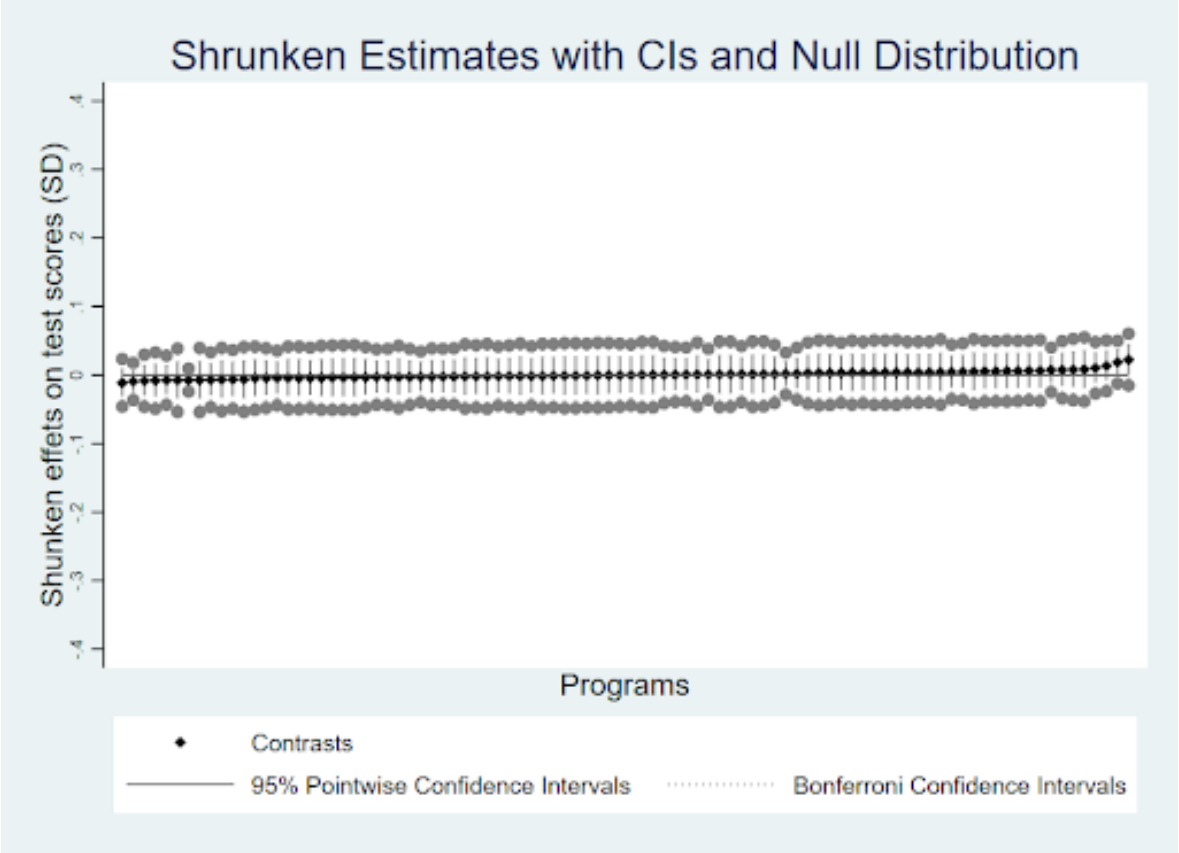

Figure 1d. Shrunken empirical Bayes estimates of the program effects. Estimates were shrunk using the ***metan*** *command and graphed using the* ***caterpillar*** *command—both user-written commands for Stata (Fisher et al., 2006/2024; Bellows & von Hippel, 2017). We have preserved the scaling of the Y axis to emphasize how much the estimated effects have shrunk.*

If the sample were larger, then the standard errors would be smaller, and the shrinkage would be less severe. But that does not mean that the shrunken estimates would necessarily be large. In a larger sample, the point estimates would be smaller *before shrinkage* because they would be less inflated by estimation error. Less shrinkage would be applied because less was needed, but the shrunken estimates would still be rather small.

<u>Shrunken estimates still need correction for multiple inferences</u>

A common misconception is that shrunken Bayesian estimates are exempt from the problem of multiple inferences (e.g., Gelman et al., 2012). This is not correct.[5] Even shrunken estimates are prone to false discoveries because of multiple inferences (Efron, 2012). Shrunken

[5] The usual Bayesian priors do not shrink estimates nearly enough to solve the problem of multiple inferences. Addressing multiple inferences in a Bayesian context requires specialized priors that account for the number of effects to be estimated and the prior probability that each effect is zero (e.g., Chang & Berger, 2021; Scott & Berger, 2006; Westfall et al., 1997). There is a substantial Bayesian literature on multiple inferences (Berry & Hochberg, 1999), including work on the estimation of heterogeneous effects (Berger et al., 2014). In short, Bayesians too must confront the problem of multiple inferences.

confidence intervals still have a 5 percent chance of not covering the true effect, so among 92 shrunken confidence intervals, we can expect some false discoveries.

In Figure 1d, we have widened the shrunken confidence intervals with the Bonferroni correction (although the Holm or BH correction would be preferable). After corrections, none of the shrunken estimates differs significantly from zero. Although some heterogeneity is likely present, it is very small and hard to pin on specific programs.

Recommendation 5: Use large samples—not just overall, but per estimate

In light of our failure to find more than one convincing program effect, you might imagine that the sample size was too small. In fact, the study involved over 5,000 teachers and 200,000 students (von Hippel & Bellows, 2018a). Why then did so few programs produce statistically significant differences?

One reason is that, although the sample was large overall, the samples for many individual programs were quite small. Although the largest program had almost 900 teachers, half the programs had fewer than 30 teachers each, and 20 programs had fewer than 10 teachers each. The problem of trying to estimate effects from small subsamples is common in the search for heterogeneity, and often overlooked when the overall sample size is large, as it is here.

If you are designing your own sample, you can deliberately oversample subgroups that are small in the population (Tipton et al., 2019). This was not an option in our study of teacher preparation programs, which already included every teacher trained by every program, but we could have accumulated more teachers from small programs if the study had run for a few more years. With only two years of data, it was not realistic to estimate the effects of small programs, and even more years of data might not have produced statistically significant results if the true differences between programs were as small as our analysis suggests.

It is also easy to be fooled by the fact that the sample has many individual observations (here students), and lose sight of the fact that those observations are not independent but clustered within larger units (here teachers). When analysis accounts for clustering,[6] the effective sample size may be several times smaller than it appears (Kalton, 1983).

Summary: How we un-fooled ourselves about unexplained heterogeneity

This section started with an example where there appeared to be substantial heterogeneity among programs. Seven programs had effects that differed significantly from the average, and estimated effects approached 0.2 SD for some programs. Or so it seemed at first.

But we were fooling ourselves. In fact, the program estimates were barely more dispersed than they would have been if all programs had equal effects. After correcting for multiple tests, we found that at most one of the programs had an effect that differed significantly from the average, and that effect was estimated as just 0.06 SD before shrinkage and a mere 0.02 SD afterward. We also realized that our sample, despite having over 200,000 children, nevertheless relied on small subsamples from many programs—providing inadequate power to compare many program effects.

These results illustrated how easy it is to fool ourselves about unexplained heterogeneity and demonstrated the steps needed to un-fool ourselves. In this example, being careful led us to conclude that little true heterogeneity was present—a disappointing finding, but one that might

[6] Psychologists commonly account for clustering with random effects (Yarkoni, 2022), while economists typically use cluster-robust standard errors (Cameron & Miller, 2015).

be common if our recommendations were followed routinely. On the other hand, if heterogeneity still seems substantial after our recommended steps, there would be more reason than usual to hope that the heterogeneity is real and might replicate.

How not to fool ourselves about explained heterogeneity

Our recommendations for unexplained heterogeneity also apply to efforts to explain heterogeneity with moderators. If instead of estimating a set of program effects, we were estimating a set of subgroup effects, or a set of treatment-by-moderator interactions, we would still recommend (1) decomposing the variance of estimates into signal and noise, (2) visualizing the distribution of estimates and comparing it to the null distribution, (3) correcting tests and confidence intervals for multiple inferences, (4) shrinking estimated effects toward zero, and (5) checking the number of observations needed to estimate each effect with adequate power.

But there are more ways to fool ourselves when we try to explain heterogeneity through moderators. As we remarked in the introduction, only one in five moderator-by-treatment interactions has replicated when tested in new data (Altmejd et al., 2019; Open Science Collaboration, 2015). This section will unpack the ways that we can fool ourselves when trying to explain heterogeneity.

Our presentation will support the following recommendations:

6. Estimate *interactions* instead of subgroup effects.
7. Don't test interactions without adequate power.
8. Don't test interactions at a high significance level—e.g., don't treat interactions as significant if p is less than .10 but greater than .05.
9. Don't look for too many interactions—
10. —especially if you found no main effect.
11. Preregister your interactions.
12. Motivate interactions by hypothesizing a mechanism that predicts which groups will experience larger or smaller effects.
13. Interpret covariate-treatment interactions cautiously if the covariates may be confounded with unseen variables that really moderate the effect.

We also repeat recommendation 3—correct for multiple inferences—which has additional benefits when we seek to explain heterogeneity through interactions. Finally, we resolve a longstanding debate about centering interactions in linear models. Specifically, we make the following recommendation:

14. You can mean-center variables if you like, but centering won't increase the power to detect an interaction.

We will motivate and illustrate our recommendations with several concrete examples.

Recommendation 6: Estimate interactions instead of subgroup effects

Investigators searching for heterogeneity often estimate treatment effects separately for different subgroups. But this practice, by itself, cannot prove that heterogeneity exists. Some *estimated* effects will inevitably be larger for one group than for others, but that does not prove that the *true* effect is larger as well. If an estimated effect is significant for one group and

insignificant for another, that does not prove heterogeneity either: “the difference between ‘significant’ and ‘insignificant’ is not [necessarily] significant” (Gelman & Stern, 2006).

The most general way to test whether treatment effects vary across subgroups is to estimate a subgroup-by-treatment interaction. The results of such an interaction test can be disappointing, as it is often harder to obtain a significant interaction than it is to obtain a significant result for a subgroup (Sainani, 2010; Wallach et al., 2017).

An example occurred in Project STAR, a block-randomized experiment in which students and teachers in 79 Tennessee schools were assigned at random to three conditions: a small class (averaging 15 students), a regular-sized class (averaging 23 students), and regular-sized classes with a teacher’s aide. Small classes significantly raised kindergarten test scores, while teachers’ aides did not.

Many researchers have claimed that the effect of class size was larger for black students and for students whose low family incomes qualified them for free school lunches (e.g., Finn et al., 1990; Jackson & Page, 2013; Krueger & Whitmore, 2001; Nye et al., 2000; Schanzenbach, 2006). But these claims have only been supported by subgroup analyses. When we test them with interactions, we find no significant evidence that the small-class effect varied with race or income. Table 3 uses data from Project STAR to compare effects on black and white students. When we analyze students of different races separately, the estimated class-size effect appears twice as large for black students (0.31 SD) as for white students (0.15 SD), but when we combine black and white students in a single analysis, the race-by-class-size interaction is not significant (p=0.12). So we cannot reject the null hypothesis that the true effect of class size was similar for black and white students. An interaction between class size and free-lunch status (not shown) was also nonsignificant, so we cannot reject the null hypothesis that the effect of class size was similar for children of higher and lower incomes as well.

Table 3. The effect of small classes on black and white students’ kindergarten reading scores: Results from Project STAR

Treatments	White students only	Black students only	Black and white students
Small class	0.15**	0.31**	0.15**
	(0.05)	(0.10)	(0.05)
Teacher’s aide	-0.01	0.14	-0.02
	(0.04)	(0.09)	(0.04)
Covariate			
Black student			-0.51***
			(0.08)
Interactions			
Small class × Black student			0.18
			(0.11)
Teacher’s aide × Black student			0.16
			(0.10)
Students	3,903	1,858	5,761
Schools	64	63	79

**** p<0.001, ** p<0.01, * p<0.05. Linear regression models with school fixed effects. Robust school-clustered standard errors in parentheses. Analysis is limited to black and white students.*

In light of our null interactions, it seems less surprising that later policies that reduced class sizes in California and Florida also found no evidence that the effects were larger for black children, Hispanic children, or children with low family incomes (Chingos, 2012; Jepsen & Rivkin, 2009).

Interaction tests should allow for other differences between groups. It is possible, for example, that the residual variance is different for black and white students; Table 3 allows for that by estimating standard errors that are robust to unequal residual variance (a.k.a. heteroskedasticity). If additional covariates were present, they might have different slopes for different groups; we could allow for that by including group-by-covariate interactions.

Alternatives to interactions

There are other ways to test for treatment-effect differences between subgroups. A popular approach is to estimate an effect in each subgroup separately and then compare the estimated effects with the following *Z* test (Paternoster et al., 1998):

$$Z = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{{s_1}^2 + {s_2}^2}}$$

Here $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated effects in group 1 and group 2, while s_1 and s_2 are the corresponding standard errors. In Project STAR, comparing the small-class effect across black and white students, this formula yields a *Z* statistic of 1.4 with a nonsignificant *p* value of .15—similar but not identical to the interaction test.

Although the *Z* test produced reasonable results here, it is limited because it assumes that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are uncorrelated. In Project STAR, the estimates are correlated because some black and white students attended the same schools and had the same teachers. Evidently the correlation here was negligible, but if the correlation were larger the *Z* test could suggest a significant difference when the interaction test did not.

There are other ways to test whether effects differ across subgroups. In *structural equation models*, for example, it is common to divide the sample into multiple groups and test the constraint that the parameters are equal across subgroups (Becker et al., 2013). In *seemingly unrelated regression*, one estimates regressions for two subgroups simultaneously, allowing the error terms to be correlated, and test the null hypothesis that certain coefficients are equal (Moon & Perron, 2006). In *multilevel models*, one can use regressors at level 1 to predict the effects of regressors at level 1—which reduces to a cross-level interaction when the levels are combined into a single equation (Bryk & Raudenbush, 1992). But interactions are probably the most general method, and the easiest to use in practice. Many researchers will find that they can simply estimate a subgroup-by-treatment interaction and only report the subgroups separately if the interaction is significant.

Recommendation 7: Don't test interactions without adequate power

The power to detect interactions is typically low. The most basic reason is simply that most interactions are small: the median effect size for published interactions is just 0.05 SD (Aguinis et al., 2005; Baranger et al., 2022). A recent study estimated that, even if an interaction were 0.2 SD, the median power of a typical psychology study to detect it would be only 18

percent (Sommet et al., 2023). If in fact the typical interaction is closer to 0.05 SD, then the typical power to detect it may be well below 18 percent.

Another reason for low power is that some interactions are identified by a small subset of data. For example, one study reported that black teachers were more likely to assign black children to programs for gifted students (Grissom & Redding, 2016). Although the analytic sample contained approximately 6,000 children, only 10 or so were black children who were assigned to gifted programs by black teachers.[7] With identification hinging on such a small subgroup, power was likely low. The result did not replicate in a later sample (Morgan & Hu, 2023).

Again, if you can design your own sample, you can deliberately oversample small subgroups that you suspect will have effects that differ from the average (Tipton et al., 2019). But if you are conducting secondary analyses, using samples that have already been collected, you may simply not have adequate power to detect many interactions that might interest you.

Low power increases the false discovery rate

Why is low power a problem? Tests with low power may seem conservative since they have little chance of producing a statistically significant result. Yet paradoxically, low power increases the false discovery rate—the proportion of statistically significant interactions that are spurious (Button et al., 2013; Higginson & Munafò, 2016; Ioannidis, 2005).

To understand this apparent paradox, look at the upper left corner of Figure 2. Here two interactions are tested at a significance level of 5 percent. One interaction is false (i.e., null or zero); the other interaction is true (i.e., nonzero though possibly small). We graph the false discovery rate as a function of power.

Remember that power is the probability that the *true* interaction will produce a statistically significant finding, while the significance level is the probability that the *false* interaction will produce a statistically significant finding. So if the power is much greater than the significance level, then the true interaction has a much greater chance than the false interaction of producing a significant result—and the false discovery rate will be low. But if the power is closer to the significance level, then the true and false interactions will have more similar chances producing a significant result—and the false discovery rate will be high.

For example, if power is 80 percent (a common target in power analysis), then the false discovery rate is only 6 percent. But if power is only 18 percent (a typical value when trying to detect interactions), then the false discovery rate rises to 22 percent or more. That is, 1 in 5

[7] The study authors did not report this number, but it was easy to calculate since only 3 percent of students were black children assigned to black teachers, and only 5 percent of students were assigned to gifted programs in any year (Grissom & Redding, 2016, Table 3). 5 percent of 3 percent of 6,000 is about 10 students. The effective sample size was further reduced by the fact that the sample was clustered, so that the 10 students likely had fewer than 10 teachers in fewer than 10 schools. The reported standard errors neglected clustering, and if standard errors were clustered it is unclear whether the interaction between student race and teacher race would have been significant.

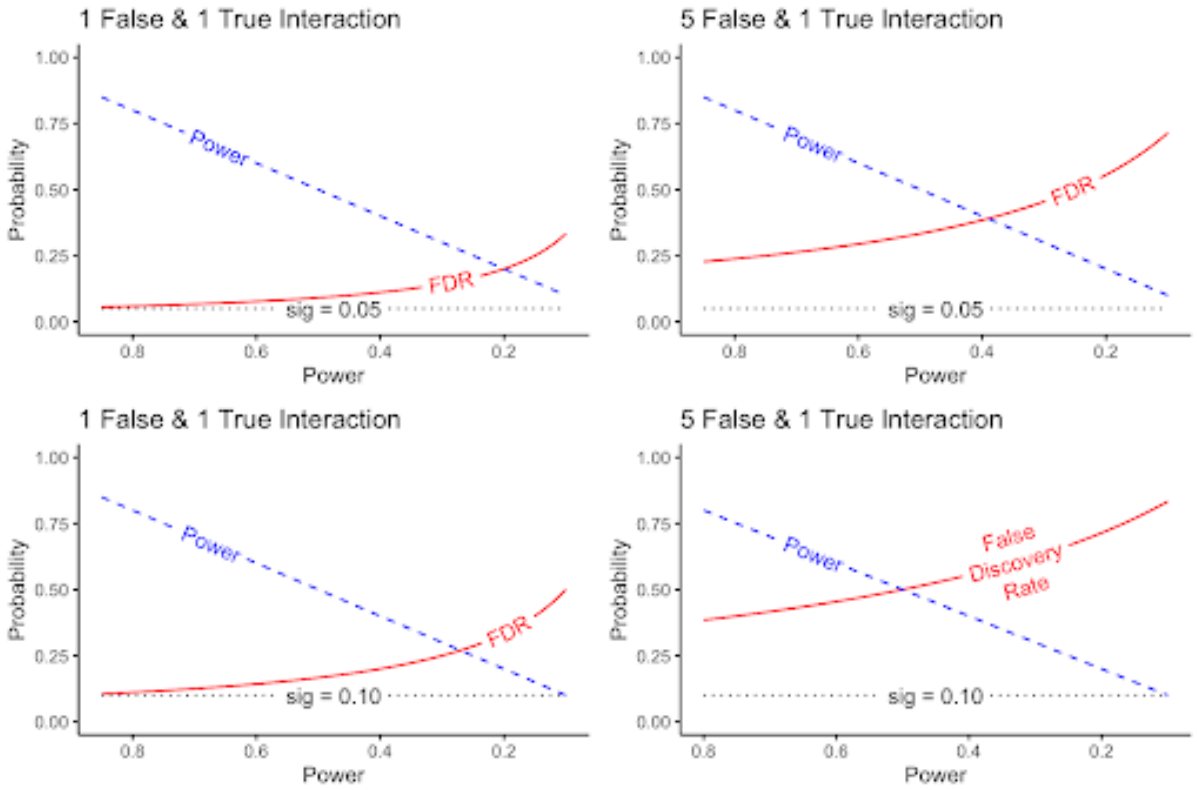

Figure 2. As power declines toward the significance level, the false discovery rate (FDR) rises. The FDR is greater if several false interactions are tested for each true one or if a higher level is used to determine significance. This figure graphs the false discovery rate, which is aF/(aF+(1-β)T), where a is the significance level, 1-β is the power, and T and F are the numbers of true and false interactions.

Recommendation 8: Don't test at a high significance level (e.g., 10 percent)

The principle illustrated above is that the false discovery rate is high if power is close to the significance level. As we have shown, this means that the false discovery rate rises if power is low. But it also means that the false discovery rate rises *if the significance level is high.*

A significance level of 5 percent is usually required to put an asterisk—*p<.05—next to a result in psychology, sociology, or public health. But these fields often allow a dagger (†) next to a "borderline significant" result with †p<.10, and in economics *p<.10 is sufficient for an asterisk.

Figure 2 (bottom left) shows that using a significance level of 10 percent makes the problem of false discoveries worse. Now if one false interaction is tested for each true one, at 18 percent power the false discovery rate is 36 percent. That is, about 1 in 3 significant interactions will be spurious.

Recommendation 9: Don't look for too many interactions

So far we have assumed that investigators test only one false interaction for each true one. But investigators often test many interactions, and the false discovery rate grows if a large fraction of tested interactions are false. This highlights the problem of multiple inferences.

The right side of Figure 2 shows what happens if we test 5 false interactions for each true one. Now iI the power to detect a true interaction is 18 percent (as is typical), the false discovery rate will be 58 percent if investigators use a 5 percent significance level, and 74 percent if they use a significance level of 10 percent. That is, most significant interactions will be spurious.

Figure 3 looks at the same problem through a different lens, shifting focus to the familywise error rate—the probability of at least one false discovery. The figure graphs the familywise error rate as a function of the significance level and the number of false interactions tested.

- If we test interactions at a significance level of 5 percent, the risk of at least one false discovery is 23 percent if we test 5 false interactions, 50 percent if we test 10 false interactions, and 64 percent if we test 20 false interactions.
- If we test at a significance level of 10 percent, the risk of at least one false discovery is worse—41 percent if investigators test 5 false interactions, 65 percent if we test 10 false interactions, and 89 percent if we test 20 false interactions.

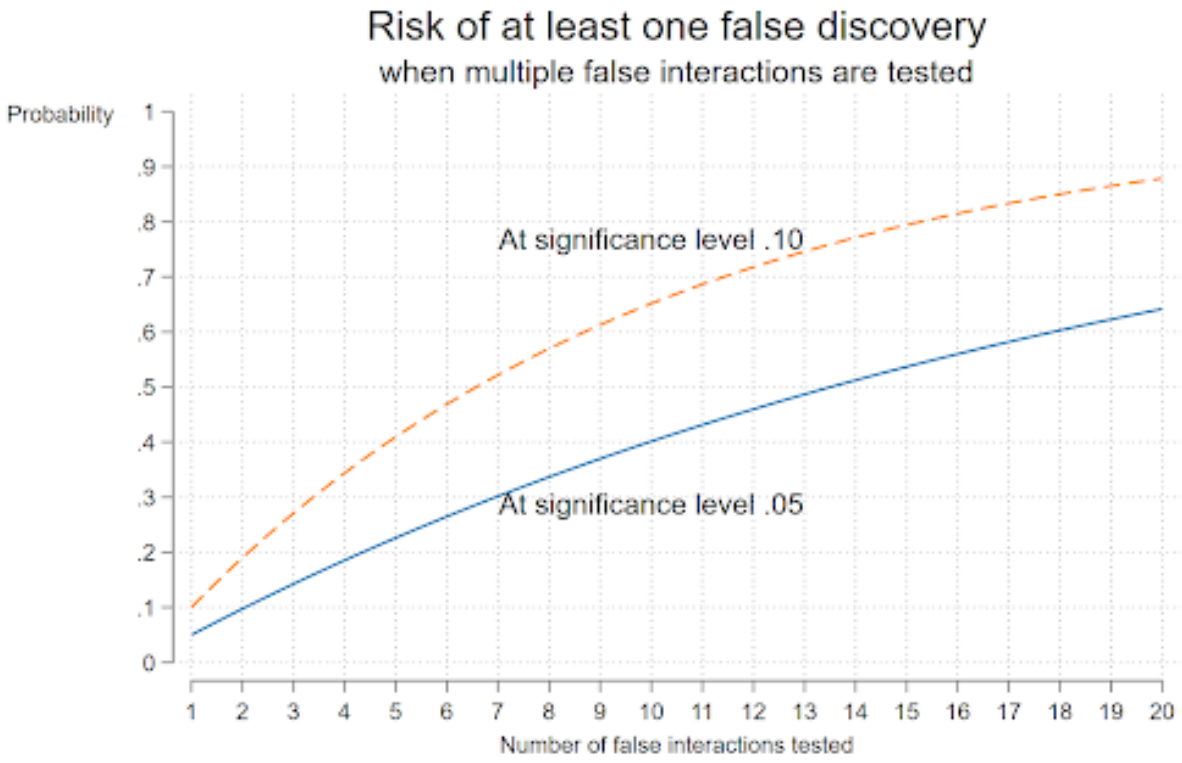

Figure 3. As more false interactions are tested, the risk of a false discovery rises. This figure graphs the probability of at least one false discovery, which is $1-(1-a)^k$, where a is the significance level and k is the number of false interactions tested.

These observations are important, because a single significant result, if highlighted in a study abstract, can be enough to get a study published. And investigators can practically guarantee a single significant result by testing enough interactions.

It is easy to find studies where 5 interactions—or 20 or more—have been tested. For example, a treatment can interact with gender (2 categories or more), race/ethnicity (2 to 6 categories), weight status (overweight, obesity, normal weight), age, educational level, and many other characteristics. The number of potential interactions grows if three- and four-way interactions are permitted—as they often are, at least implicitly.

For example, in estimating the effect of physical education classes on body weight, one study reported no effect except for boys in fifth grade (Cawley et al., 2013)—implying a three-way interaction between treatment, gender, and age. Another study reported an effect only for girls in kindergarten and first grade, and only if they started the study overweight—implying a four-way interaction between treatment, gender, age, and initial weight status (Datar & Sturm, 2004).

Neither interaction replicated in later data (Bednar & Rouse, 2020), and both may have been false discoveries.

Recommendation 10: Resist the temptation to test too many interactions if there is no main effect

Investigators who fail to find a significant main effect may be especially tempted to search for interactions. Such "no-but heterogeneity" does exist, but it is rare. True heterogeneity is generally smaller when there is little or no main effect (Olsson-Collentine et al., 2020).

In other words, investigators may be most tempted to search for heterogeneity when it is least likely to be present — "grasping for a positive straw in a negative haystack" (Berry, 2012). Investigators who search for no-but heterogeneity will test more false interactions and conduct lower-powered tests of true interactions. The result will be a higher false discovery rate than would be typical if a strong and significant main effect were present.

Recommendation 3 (again): Correct for multiple inferences

Our recommendation to avoid testing too many interactions stems from concerns about multiple inferences. We already recommended correcting for multiple inferences in our section on unexplained heterogeneity, but corrections have additional virtues when we try to explain heterogeneity through interactions.

Multiple test corrections do not just correct interactions that have already been tested—they also provide an incentive to avoid testing too many interactions in the first place. Once we agree to correct for multiple tests, then every interaction we test raises the bar that an interaction must clear to achieve significance. For example, if we test only 5 interactions, then we must multiply the smallest uncorrected p value by 5, and as long as the smallest uncorrected p value is less than .01, p will still be less than .05 after correction. But if we test only 5 interactions, then we must multiply the smallest uncorrected p value by 20, and unless the smallest uncorrected p value is less than .0025, every p value will be greater than .05 after correction.

By contrast, if we don't correct for multiple inferences, then every test increases our chances of finding a significant result—but many significant results will be spurious.

Recommendation 11: Preregister interactions and don't report interactions selectively

Many investigators report interactions selectively—testing many interactions but reporting only a few in their article, or reporting all interactions in their tables but mentioning only a few in their discussion and perhaps highlighting just one in their abstract. Multiple test corrections may be inadequate if they only correct for the number of interactions that are reported and not for the number of tests that were carried out.

The answer to this problem is preregistration. Before testing any interactions, we should preregister (for example on the Open Science Framework) which interactions we plan to test, and why. Later corrections should correct for all the interactions that were preregistered and not just those that were reported.

Recommendation 12: Motivate interactions with causal mechanisms

One way to reduce the number of false interactions tested is to make sure that every interaction has a strong and clear motivation. The motivation should be linked to the mechanism by which the treatment is believed to work (Rohrer & Arslan, 2021). When we have a clear and consistent theory about why an intervention should work at all, they are better prepared to make clear and targeted predictions about when and for whom it will work best or worst. By contrast, when the treatment is a black box with no clear mechanism, it is harder to explain why the effects would be heterogeneous.

An example of a well-theorized interaction occurs in the literature on early reading, where several studies have tested the hypothesis that phonics instruction is only useful for readers who lack fluent phonics skills. Students who have already mastered phonics will benefit little from further phonics instruction and can make better progress through independent and meaning-focused activities. Several small to moderate-sized observational and randomized studies have confirmed the interaction between phonics instruction and initial skill level, and investigators have developed a screening test that can prescribe an effective mix of basic and advanced reading activities according to a student's initial skill level (Connor & Morrison, 2016; Connor et al., 2004, 2009). Nevertheless, one large observational study failed to find any interaction between teachers' instructional practices and children's initial reading level (Chiatovich & Stipek, 2016). As this study illustrates, even well-motivated interactions do not always replicate, but they should offer us a better chance of avoiding false discoveries.

By contrast, many investigations simply interact the treatment with whatever demographic characteristics happen to be available, such as age, race, poverty, income, or geographic location (Cintron et al., 2022). Although there are behaviors—such as attraction or discrimination—where demographics may truly be the causal moderators of interest, in many settings demographics are only weak proxies, at best, for the underlying social, psychological, or biological processes that might moderate a treatment's effects.

For example, class size research rarely specifies the mechanism by which small classes should help anyone learn (Pedder, 2006), and without a clear mechanism it is hard to motivate the hypothesis that small classes would be better for black children (the hypothesis that failed in Table 3) or children from families with low incomes. It is not hard to improvise a theory—maybe children in smaller classes get more differentiated instruction, or maybe classmates distract them less—but such theories would not motivate interactions between class size and race or poverty. Instead, if the mechanism were that smaller classes reduce distraction, we might

expect larger effects for children with ADHD; or if smaller classes let teachers differentiate instruction, we might expect larger effects in classrooms where children's initial skill levels were more varied.

Recommendation 13: Interpret interactions cautiously when covariates may be confounded

Note that diverse skill levels and distracting environments might be more common in classrooms serving larger proportions of black students or students from low-income families. But such classrooms might not be more distracting or more varied in skill levels, and even if they were, race and poverty would be weak proxies at best—only slightly correlated with the psychological and instructional variables that actually moderate the effects of class size. When data are limited and mechanisms are vague, it is possible that apparent moderators in our study are actually proxies for unseen variables that really moderate the effect (Rohrer & Arslan, 2021). We find ourselves back in Cronbach's (1975) "hall of mirrors."

Note that the possibility of confounding disappears when a moderator is itself a treatment assigned at random. For example, in a psychiatry study that randomizes two treatments—say medication and therapy—there would be little ambiguity if there were an interaction suggesting that patients who received both medication and therapy benefited more than patients receiving either medication or therapy alone. But if a randomized treatment interacts with a trait or behavior that is not randomized —for example if therapy were more effective for one gender than for another—then there is always the possibility that the observed moderator is confounded with some other variable that truly moderates the effect (Tipton et al., 2019). This is part of what Cronbach meant by the "hall of mirrors."

Recommendation 14: Center variables if you like, but it won't increase the power to detect interactions

Power to detect interactions is often low. As we wrote earlier, the main reasons are that most interactions are small and many interactions are identified by small subsamples. But another reason is often suggested: an interaction *XZ* may be nearly *collinear* (i.e., strongly correlated) with one or both of the component variables *X* and *Z*. Collinearity can indeed inflate standard errors and reduce power. Nevertheless, collinearity is not a reason for low power to detect interactions, and steps to reduce collinearity do not increase the power to detect interactions.

To reduce collinearity, some scholars recommend mean-centering *X* and *Z* (i.e., subtracting their means) before calculating the interaction *XZ* (Aiken & West, 1991; Iacobucci et al., 2016; Robinson & Schumacker, 2009). Yet other scholars disagree, claiming that centering does nothing to reduce collinearity or shrink standard errors (Echambadi & Hess, 2007; McClelland et al., 2017). This disagreement has confused many applied researchers, who are left wondering whether they should center interacting variables or not.

Table 4 clarifies what does and does not change when interacting variables are mean-centered. Using data from Project STAR, the model regresses children's school absences *Y* on class size *Z* and county flu prevalence *X*. The interaction *XZ* tests the hypothesis that, since children can catch flu from their classmates, having a small class might reduce absences more in counties where flu prevalence was high (von Hippel, 2021). The table presents two versions of the model, one where *X* and *Z* are mean-centered and one where they are not.

Table 4. Linear regression predicting school absences, with flu prevalence centered or uncentered

	Uncentered	Centered
Class size (1=small, 0=large)	0.19	0.22
	(0.22)	(0.20)
Local flu prevalence (in percentage points)	-0.91	-0.45+
	(1.13)	(0.27)
Interaction: Small class × Local flu prevalence	0.06	0.06
	(0.15)	(0.15)
R^2	0.0438	0.0438

*†p<0.10, *p<0.05, **p<0.01. The model, fully described in von Hippel (2021), included child random effects, school fixed effects, and teacher-clustered standard errors (in parentheses).*

Notice that centering did not change the interaction at all. The interaction's point estimate, standard error, and *p* value are exactly the same whether variables are centered or not. These results are typical, and they mean that centering does not increase the power to detect an interaction (Afshartous & Preston, 2011; Iacobucci et al., 2016). Centering also did not change the model fit (R^2), which was 0.0438 whether variables were centered or not.

But centering did change the estimates of the main effects. The estimated effect of class size changed a little, and the estimated effect of flu prevalence changed a lot. The coefficient of flu prevalence was cut in half, and its standard error and *p* value were cut by a factor of 4, turning a clearly nonsignificant estimate (p=0.39) into one that was closer to significant (p=0.09). The reason for the increased power was reduced collinearity. Before centering, the interaction had a 0.96 correlation with flu prevalence, but centering reduced this correlation to 0.00; that is why the standard error for the effect of flu prevalence shrank by a factor of 4. Centering also reduced the correlation between class size and the interaction, but not as much, from 0.21 to 0.02; that is why the standard error of the class size effect shrank only a little.

But centering did not just change the *estimate* of the main effects; it also changed what the main effects *represent*. Before mean-centering, the main effect of class size represented the effect of small classes in a community with *no* flu. But after mean-centering, the class size slope represented the effect of small classes in a community with *average* flu prevalence. In a model with no interaction, these slopes would be the same, but in a model with interactions they can be different because the effect of class size may be different at different levels of flu prevalence. The shape of the regression surface is the same, but centering means that the main effects are estimated at the middle of the surface rather than the edge.

So should investigators center interacting variables or not? We would often recommend centering even though it doesn't increase power to detect an interaction. The advantage of centering is that it reduces the standard errors and *p* values of main effects, while allowing researchers to interpret them as the main effect at the average value of the covariates. In many settings (but not all), investigators will prefer this interpretation over the alternative of estimating main effects at a covariate value of 0.

Unreliability usually is not a major reason for low power

Another reason sometimes given for interactions' low power is that interactions are *unreliable*—that is, because much of the interaction's variance is due to measurement error

(Jaccard & Wan, 1995). There is a kernel of truth to this. The reliability of an interaction *XZ* is $\rho_X \rho_Z$, where ρ_X and ρ_Z are the reliability of *X* and *Z*—so an interaction between two 50% reliable variables is only 25% reliable. But this only matters in observational studies where variables are measured poorly. In experiments where one of the variables is a treatment assigned at random, that variable will typically be measured with little error and the problem of unreliability will be minor.

Summary: How we un-fooled ourselves about explained heterogeneity

In this section, we reviewed claims that several treatments had heterogeneous effects: the claim that small classes are more effective for children who are black or have low incomes, that physical education reduces weight only for children of certain ages and genders, and the claim that matching students' race to teachers' race increases the chances that black students will be classified as gifted. In each case, we found that the result did not replicate and that the original study either did not estimate the relevant interaction or lacked adequate power to detect the interaction if it was present.

More broadly, our calculations suggested that a combination of inadequate power, inadequate theoretical motivation, and multiple inferences have produced a situation where 50 to 89 percent of significant interactions may be false discoveries. Although this sounds grim, it is not unrealistic. In the social sciences, only about 20 percent of significant interactions can be replicated (Altmejd et al., 2019; Open Science Collaboration, 2015). If research practices were typically sound, significant interactions would be more common in large samples, where power is high, but instead significant interactions are more common in small samples, where power is low (O'Boyle et al., 2019). These considerations suggest that many false interactions are being tested, many non-significant interactions are not being reported, and many interactions reported as significant are probably false.

We suggested that investigators could reduce false discoveries by estimating interactions instead of subgroup effects, limiting estimation to interactions that have adequate power and clear theoretical motivation, and correcting for multiple tests. If these recommendations are followed, future attempts to explain heterogeneity should be more trustworthy and replicable.

How Not To Let Models and Measurement Fool Us About Heterogeneity

Our recommendations so far can help us not to fool ourselves about heterogeneity. Yet nearly all our recommendations so far pertain to general research *practices*—such as avoiding underpowered tests and correcting for multiple inferences.

Even if we follow sound *practices*, however, our *models* and *measurements* can sometimes fool us about heterogeneity. We are most easily fooled when models are non-linear or data might be measured on a non-interval scale. In this section, we will make 4 final recommendations:

15. Check whether your interaction is present on every scale that might be used to measure the outcome variable *Y*
16. Check for nonlinearity, floor, and ceiling effects.

17. When using logistic regression, check whether the interaction is present on both a probability and a log-odds scale.
18. When interpreting logistic, ordinal, or interval regression in terms of latent variables, check whether the interaction is still present if the latent variable has different variances for different subgroups.

We will motivate these recommendations with concrete examples and illustrations.

Baseline misalignment: It is hard to compare effects on groups who start at different levels

Before making our recommendations, we should say that they all stem in part from the problem of *baseline misalignment*. Baseline *alignment*—the requirement that the treatment and control groups have similar *Y* levels before treatment—is a common prerequisite for causal inference (Sekhon, 2009). If the treatment and control group start at similar *Y* levels, but finish at significantly different *Y* levels, then it is clear that there is a main effect of treatment.

When we search for heterogeneity, we often compare effects on groups whose *Y* values are *not* aligned at baseline. For example, we might compare the effects of an educational intervention on subgroups with higher or lower initial test scores, or the effect of a job training program on subgroups with higher or lower initial wages. When the groups being compared start at different *Y* levels—when they are *misaligned* at baseline—it can be challenging to tell whether effects on different subgroups are different. The challenges are especially acute when models are non-linear or the *Y* variable can be measured in different ways.

Recommendation 15: Check whether the interaction is present on every scale

When groups are not aligned at baseline, interactions can be sensitive to the *scale* on which the outcome *Y* is measured. On some scales, it may appear that the treatment effect is larger for group 1; on others, it may appear that the effect is larger for group 2. On still other scales, it may appears that the effect is equal for both groups (Domingue et al., 2022; Rohrer & Arslan, 2021).

As a simple example, suppose that a job training program increased group 1's average earnings from $20,000 to $30,000 and group 2's average earnings from $30,000 to $40,000. On a dollar scale, there is no group-by-treatment interaction; treatment increased both group's earnings by the same amount: $10,000. But on a percent scale, there is a group-by-treatment interaction; treatment improved group 1's earnings by 50 percent and group 2's earnings by only 33 percent. The effect on group 1 also appears greater if we use the logarithm of earnings.

Another example occurs in the literature on summer learning. Many studies have asked whether children in poverty fall behind other children more quickly during summer or during school—that is, does the effect of summer on learning interact with student poverty? The answer, it would seem, could shed light on whether schools make inequality better or worse. Yet the empirical evidence has been maddeningly inconsistent (von Hippel, 2019; von Hippel & Hamrock, 2019; von Hippel et al., 2018; Workman et al., 2022), and one reason is that results are sensitive to the scale on which learning is measured.

Figure 4 shows results from a longitudinal study where reading tests were scored on two different scales (von Hippel & Hamrock, 2019). One scale measured how many questions children answered correctly; the other scale estimated children's ability using item response

theory (IRT). On the number-right scale (left), children in poverty fell behind other children during first grade, but on the IRT ability scale (right), children in poverty caught up during first grade. The children were the same; the tests were the same—but changing the measurement scale flipped the time-by-poverty interaction from positive to negative (von Hippel & Hamrock, 2019, tbl. A5).

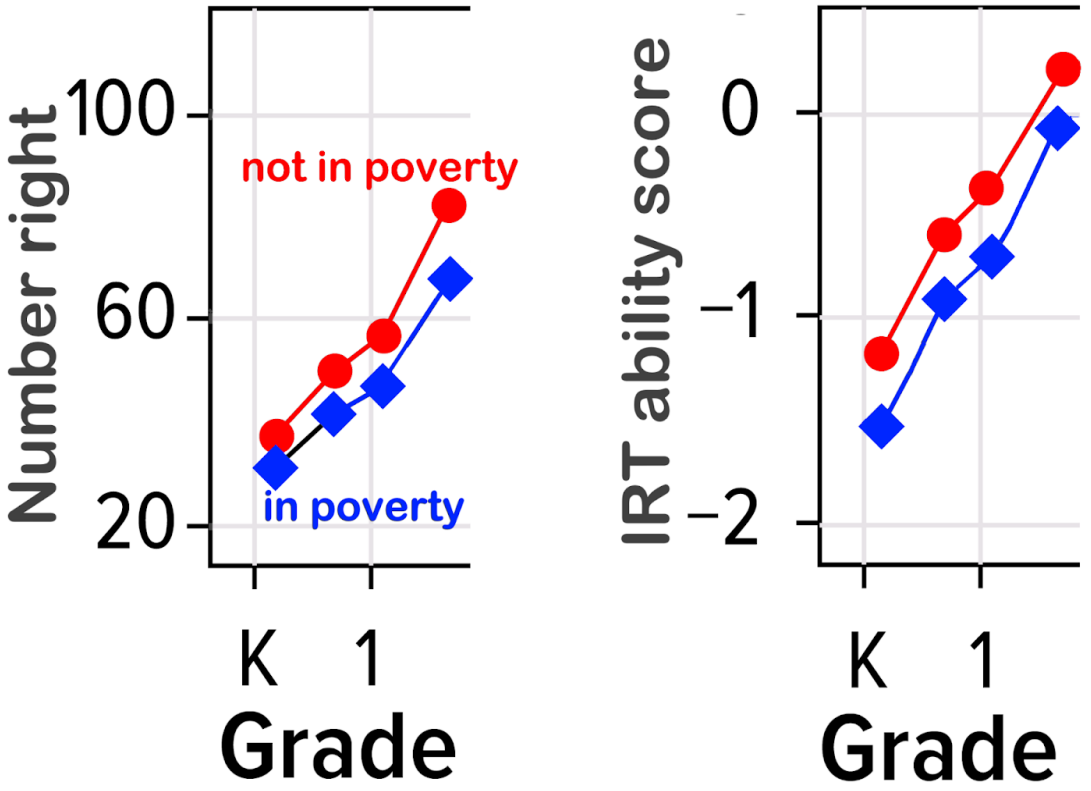

Figure 4. When scored on a number-right scale, children in poverty appear to fall behind during first grade. When scored on an IRT ability scale, children in poverty appear to catch up during first grade. Adapted from von Hippel and Hamrock (2019, Figure 1).

The interaction flipped not just because the scale changed, but also because children in poverty did not start at the same place on the scale as other children. In addition, the relationship between the two scales was nonlinear. If the two scales had had a linear relationship—for example, if one measured the number right and the other measured the percent right—then it would not have been possible to flip the interaction.

This example illustrates the importance of checking whether an interaction is robust to changes in scaling. If an interaction is sensitive to scaling, as this one is, we have two choices:

- We can argue that it is unclear whether the interaction is positive or negative, and refrain from interpreting it further.
- Or we can argue that one scaling is better than another, and that the better scale illustrates the true interaction.

In this example, we could argue that the IRT scale measures students' skills better than the number-right scale (von Hippel & Hamrock, 2019). On the other hand, studies using different IRT scales can still produce inconsistent and non-replicable results (Workman et al., 2022). Because of scaling and other issues, it remains unclear whether achievement gaps grow or shrink during school and summer, and whether schools make inequality better or worse.

Recommendation 16: Check for nonlinearity, floor effects, and ceiling effects

Interactions are often estimated with models that assume effects are linear. If effects are actually nonlinear, then fitting a linear model can produce the illusion of interaction where none exists.

Nonlinearity is not a concern when both the treatment and the moderator are discrete, but when either the treatment or moderator is continuous, then nonlinearity can be a concern.

Figure 5 gives 2 examples of nonlinear relationships in cognitive psychology. The left panel illustrates how reaction time decreases with practice. Reaction time decreases quickly at first, and then more slowly as it nears the floor of 0, approximating an exponential curve or power law (Heathcote et al., 2000).

The right panel illustrates how accuracy improves with practice. Accuracy improves quickly at first, then more slowly as it approaches the ceiling of 100 percent, approximating a logistic or sigmoid curve.

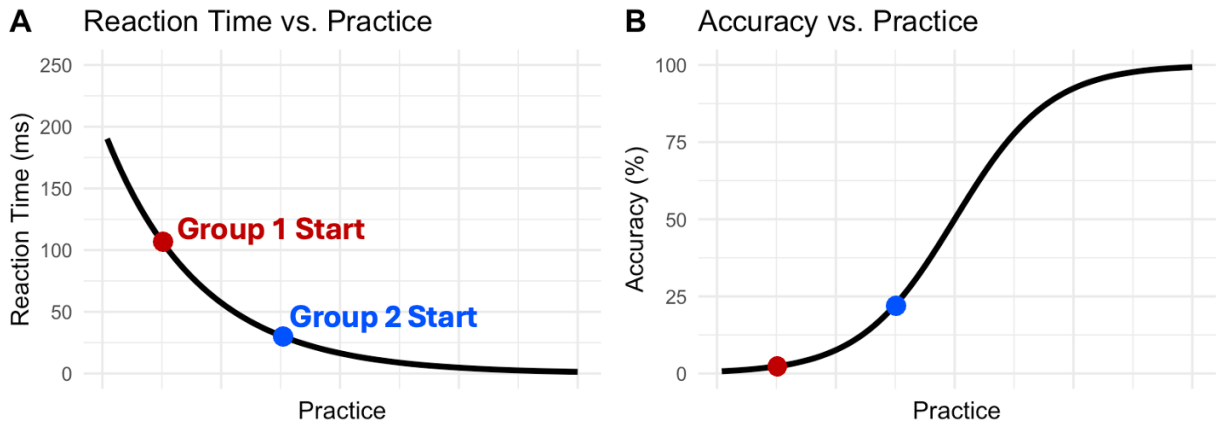

Figure 5. Common examples of nonlinearity in learning data. Group 1 is raw beginners, while group 2 has prior experience. Both groups follow the same nonlinear curve, but if an incorrect linear model is fit, it will appear that there is a group-by-practice interaction because one group progresses more quickly than the other after starting the experiment.

Such nonlinear relationships can create an illusion of interaction if we compare groups who start at different points on the curve. For example, we might compare the effects of practice on two groups—a group of raw beginners and a group with prior experience in the task. It may appear that the more experienced group improves less from practice—a group-by-treatment interaction—but the real issue is that the more experienced group is starting on a flatter part of the curve, nearer the floor or ceiling. They improve less because less improvement is possible, not because practice is less effective for them.

This highlights the necessity of plotting your data and not just assuming the data fit a linear model.If you find evidence of nonlinearity, it is better to use a curve that fits the data—say an exponential, power, or logistic curve—rather than assuming linearity. If no simple mathematical function fits the data well, then you can fit a flexible curve using splines, for example, or generalized additive models (Simonsohn, 2024).

The issue here is similar to the issues that arise from scaling. Indeed, the relationships in Figure 5 can be linearized if we transform the scale of the outcome. For example, if we take the log of reaction time, then its relationship with practice will be closer to linear, and it may be possible to estimate an interaction effect reasonably accurately with a linear model. Likewise, we can often linearize the relationship between practice and accuracy by applying a logistic transformation to accuracy. We will say more about the logistic transformation in the next section.

Recommendation 17: In logistic regression, check if interactions occur in both probabilities and log odds

Perhaps the most common example of interactions being sensitive to scaling, nonlinearity, floors, and ceilings occurs when modeling the effect of some treatment on the chances of a binary outcome *Y*=1 such as graduating high school, voting, or becoming overweight (Domingue et al., 2022; Embretson, 1996; Kang & Waller, 2005). Two models are commonly used.

A *linear probability model* predicts the probability *p* of *Y*=1 as a linear function of *X*, *Z*, and the interaction *XZ*:

$$p = \alpha_0 + \alpha_X X + \alpha_Z Z + \alpha_{XZ} XZ$$

The coefficients represent the effect on the probability of graduation, etc. of a one-unit increase in *X*, *Y*, or the interaction *XZ*.

Alternatively, a *logistic regression model* predicts the log odds *ln((p/(1-p))* as a linear function of the same variables:

$$ln((p/(1-p)) = \beta_0 + \beta_X X + \beta_Z Z + \beta_{XZ} XZ$$

The logistic model is harder to interpret, but has the advantage of never predicting out-of-bounds values because log odds have no ceiling or floor.

If all the probabilities are in the middle of the range—say, between 0.2 and 0.8—then the linear and logistic models will give compatible results (Long, 1997; von Hippel, 2015). The logistic regression coefficients β will have an approximately linear relationship with the linear probability coefficients α, and if one model shows a significant interaction, then so will the other. The models will also give compatible results if the probabilities are all in a narrow range near the ceiling (a probability of 1) or near the floor (a probability of 0) (von Hippel, 2017).

But if the starting probabilities are near the ceiling for one group, and not for the other, then whether you see an interaction can depend on whether you put the results on a probability scale or a log odds scale. In fact, interaction terms in logistic and linear models can disagree not just in significance, but also in sign (Ai & Norton, 2003).

Figure 6 illustrates the problem with models that predict high school graduation from motivation and cognitive ability (Ganzach et al., 2000).

- A linear model, predicting the probability of graduation, shows a negative interaction effect between a motivational "treatment"[8] and ability, suggesting that treatment was more beneficial for those with lower ability.
- But a logistic model, using a log odds scale, showed a *positive* interaction, suggesting that the treatment was most helpful to those with *high* ability.

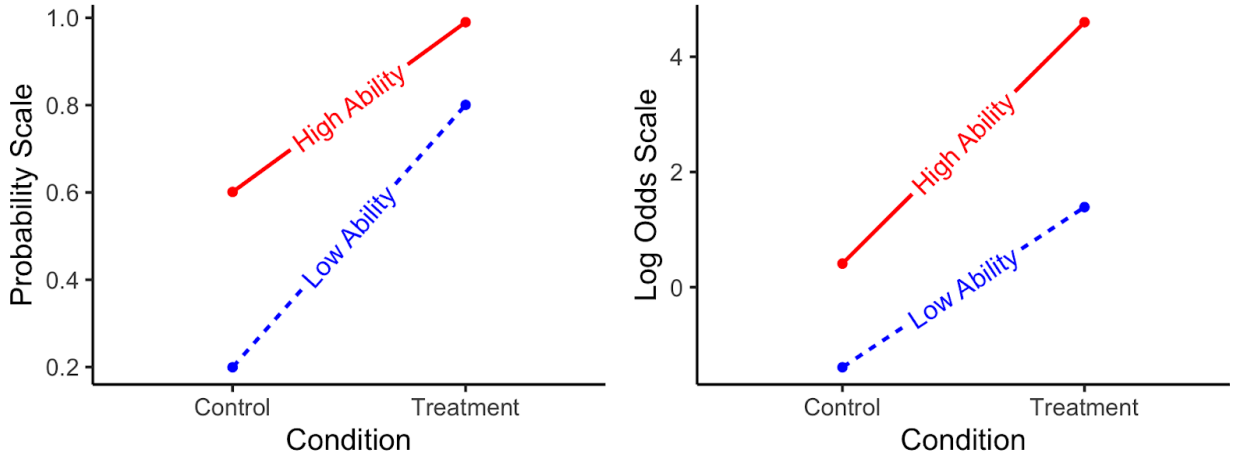

Figure 6. The appearance of interactions depends on scale. Both graphs show the effects on graduation of a motivational intervention given to students with high vs. low ability. The results are the same in both graphs, but the scale of the Y axis is different. The left panel measures graduation on a probability scale, and suggests a negative treatment-by-ability interaction: the benefit of the motivational intervention appears greater for students with low ability. The right panel measures graduation on a log-odds scale, and suggests a positive interaction, with a larger treatment benefit for students with high ability. (This example is inspired by Ganzach et al. [2000].)

This paradoxical result occurred for three reasons:

- First, the high- and low-ability groups did not start at the same level. If the groups started with the same probability of graduation, the difference between the probability and log odds scale would not matter; the interaction would have the same sign either way. But since the two groups started at different points, the choice of scale mattered for the interaction.
- Second, the log odds scale is a nonlinear transformation of the probability scale. If the two scales had a linear relationship, the change of scale would not affect the sign of the interaction.
- Third, the log odds scale is more sensitive to increases or decreases in probabilities near the ceiling of 100%. In this example, the high-ability treatment group had almost a 100% graduation rate. If neither group's graduation rate went above 80% or so, then again the interaction would not be so sensitive to the scale.

[8] In Ganzach et al.'s (2000) original example, there was no motivational treatment, only the students' self-reported motivation level. We have changed this to a motivational treatment, in order to preserve a clear distinction between the treatment variable and the moderator (ability).

Which scale should investigators choose—probabilities or log odds? Some researchers argue that probabilities are a more "natural" metric. While these researchers do not necessarily advocate the linear probability model, they do argue that the results of a logistic regression model should be transformed back to a probability scale before any judgment about interactions is made (Long & Mustillo, 2021; Mize, 2019).

Probabilities are indeed a more natural scale in many settings, but they also raise interpretive problems because of ceiling and floor effects. On a probability scale, higher education interventions will often seem to have smaller effects on students whose probability of graduation would be near 100% without intervention. But does this mean that the interventions are less "effective" for such students, or does it simply reflect the fact that those students' outcomes have little room for improvement on a probability scale?

And probabilities are not always the most natural scale. There are some behaviors, such as gambling, where it is more natural to think in terms of odds or log odds.

There are also settings where the purpose of logistic regression is not to estimate probabilities at all. We will discuss those settings next.

Recommendation 18: In nonlinear models, allow for unequal latent variances

There is another way to interpret logistic regression. Instead of interpreting it just as a probability model, we can view it as an attempt to estimate the parameters of a linear regression model:

$$Y^* = \beta_0 + \beta_X X + \beta_Z Z + \beta_{XZ} XZ + e^*$$

where Y^* is a latent variable that we cannot observe directly; instead, we can only observe whether Y^* exceeds some threshold (Long, 1997). For example, suppose that Y^* is adult body mass index (BMI), but all we observe is a dummy variable Y indicating whether the adult is overweight ($Y^* \geq 25$) or not ($Y^* < 25$). Then we can estimate the parameters of the linear regression model for Y^* by fitting a logistic regression model to Y.

It would appear that this approach can be used to estimate interactions. For example, if Z is a weight loss intervention and X is a dummy indicating whether the patient is male or female, then if we conduct a logistic regression of overweight Y on X, Z, and the interaction XZ, the coefficient of the interaction will tell us whether the intervention reduced BMI Y^* more among men or among women—even if we cannot observe Y^* directly but can only observe the dummy variable Y representing overweight.

But there is a catch. The approach assumes that the residuals e^* of the latent linear regression have a logistic distribution with the same variance for men and for women. If this assumption is violated, then the approach can give misleading results (Allison, 1999). You can get a significant interaction when actually the treatment has the same effect on both groups. Or you can get no interaction when actually the treatment has different effects in each group.

Figure 7 illustrates the problem. Groups A and B have the same average Y^* value before treatment and the same, somewhat higher, average Y^* value after treatment. So the treatment's effect on Y^* is the same for both groups; there is no interaction. However, because group B has a larger residual variance than group A, the probability of exceeding the threshold increases faster for group B than for group A. If you use logistic regression to model the probability of exceeding the threshold, you will get a substantial and possibly significant coefficient for the interaction XZ.

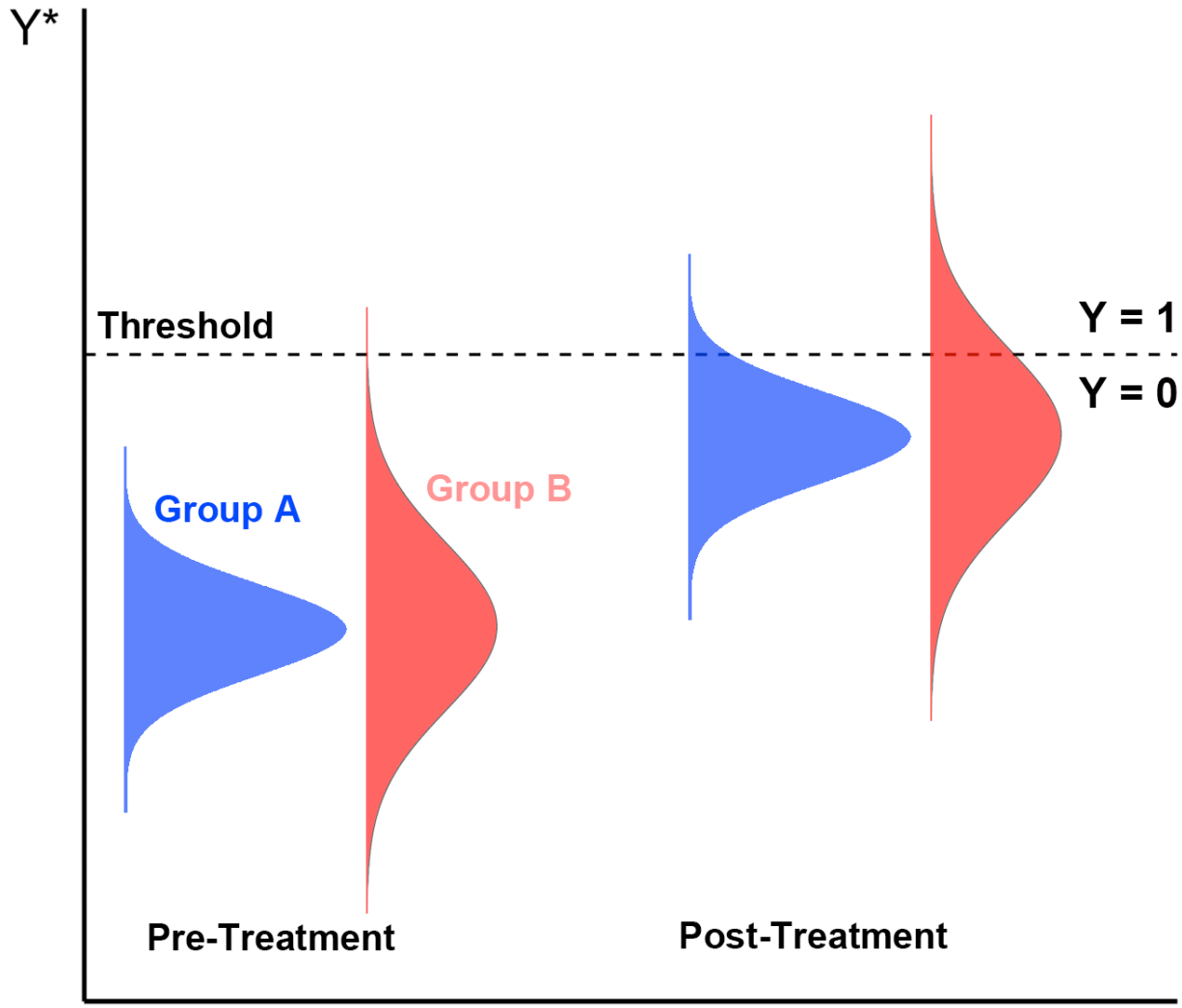

Figure 7. Differences in latent variance can create the illusion of interaction. Here the latent means change by the same amount in group A and group B. But because the groups have unequal variances, the probability of exceeding the threshold increases by only about 10 percent for group A and by 30 percent for group B.

This problem here is distinct from the question of whether we use probabilities or log odds—either way, we can get misleading estimates for the interaction. It is also not a question of the two groups having different mean *Y** values before treatment. For this example, we made sure that they had the same means. The problem is just that they have different variances.

The problem of unequal variances is not limited to logistic regression, but crops up whenever we want to model an unobserved continuous variable *Y** from an observed variable that tells us only whether *Y** falls in certain categories or intervals. Examples occur in ordinal logistic regression, ordinal probit regression (Reardon et al., 2017; Williams, 2009, 2010), and interval regression (Bartlett, 2015). In all cases, assuming equal residual variances can lead to misleading estimates of interaction effects. The solution is to fit a more flexible model using software that allows for *heteroskedasticity*—i.e., the possibility that different groups have

different residual variances. Examples include the heteroskedastic ordinal probit model (Reardon et al., 2017) or the *het* option to the *intreg* interval regression command in Stata.

A related problem is that logistic and probit models assume that the residuals of the latent variable Y^* have a logistic or normal distribution, but some latent variables do not. For example, the logistic and normal distributions are both symmetric, but the distribution of BMI is noticeably skewed, and asymmetry in the BMI distribution may affect conclusions of models that break BMI into categories such as normal, overweight, and obesity. Some software (such as the *gintreg* command for Stata) can fit categorical models that allow the underlying variable to have a skewed distribution (e.g., a gamma distribution), which fits BMI better. That too is worth trying before claiming an interaction.

Conclusion: Will the New Heterogeneity Revolution Succeed?

The prospect of a new heterogeneity revolution is alluring. If the revolution succeeds, we will be better able to target the right interventions to the right people in the right places at the right time. Patients will get more effective drugs with fewer side effects; educational interventions will reach the students who need them without wasting resources on students who don't. In field after field, a heterogeneity revolution could provide greater benefits with lower costs (Bryan et al., 2021).

But will the new heterogeneity revolution achieve its potential? Our title is meant to evoke both a best-case and a worse-case scenario. The worst case was described in Pete Townshend's (1971) cynical lyrics to The Who's hit song, "Won't Get Fooled Again":

I'll tip my hat to the new Constitution,
Take a bow for the new revolution,
Smile and grin at the change all around,
Pick up my guitar and play,
Just like yesterday.
Then I'll get on my knees and pray
We don't get fooled again.

As discussed in the introduction, we have been through a heterogeneity revolution before, and the results were disappointing. Reliable examples of heterogeneity proved hard to find. Many interactions involved small groups. Most interactions were small and hard to interpret. Several generations of researchers have been disappointed by their attempts to harness heterogeneity in pursuit of a more robust and useful social science.

Even in recent studies, interactions have been far less replicable than main effects; in fact, only one in five interaction effects has replicated successfully in new data. Many of the practices and incentives that produced spurious interactions one or two generations ago are still alive and well today. Investigators seeking publication and funding still have strong incentives to claim interactions where none exist (Franco et al., 2014), especially if there is little or no main effect. Underpowered analyses and multiple inferences are prevalent. Statistical methods that produce illusions of heterogeneity remain common, and methods that are more robust and less prone to illusion are not as widely known.

Yet our title also evokes a more optimistic scenario. In a 1974 commencement address at Caltech, the physicist Richard Feynman expressed skepticism about the usefulness and

replicability of research in education, psychology, and criminology. Toward the end, Feynman (1974) articulated a key principle of scientific inquiry:

The first principle is that you must not fool yourself—and you are the easiest person to fool.
So you have to be very careful about that.
After you've not fooled yourself, it's easy not to fool other scientists.
You just have to be honest in a conventional way after that.

The social sciences may have reached a level of maturity where we are starting to learn how not to fool ourselves. There has been substantial progress in the 50+ years since the critiques of Cronbach, Feynman, and Townshend. We have more appreciation for the prevalence of non-replicable results and the research practices that produce them. We take larger samples and use better research designs with more developed statistical methods whose properties are better understood. Pre-registration and replication studies are becoming more common. Our goal in this paper was to collect the practices and methods that are necessary to make any present or future heterogeneity revolution more successful.

Because authors competing for publication may be reluctant to adopt practices that reduce their chances of finding significant interactions, we may need institutional reforms rather than simply trusting that individual authors will do the right thing. In clinical trials, regulators manage the risk of approving ineffective drugs by requiring statisticians to preregister analyses and correct for multiple inferences (Dmitrienko & D'Agostino, 2018). In genomics, a replication crisis led top journals to require authors to correct for multiple inferences and replicate analyses across multiple samples (Hewitt, 2012; Johnston et al., 2013).

Similar steps may be needed to address multiple inferences in psychology and other social sciences. If journals do not outright require authors to correct for multiple inferences, they can award badges to authors who do so voluntarily. Some journals already award badges to authors who pre-register their analyses—a practice that, among other things, helps to document how many hypotheses were tested and need to be corrected.

These steps will substantially increase the chances of a successful heterogeneity revolution. Yet we may have to temper our expectations for what such a revolution can achieve, at least in the short term. In light of past disappointments, it may be rash to hope that in a few years we will develop a richly textured description of exactly when, where, and for whom every intervention will work or fail. It seems more realistic to hope that we may come to understand just a few circumstances that make a big difference in the effectiveness of a few interventions. And that will require not just better methods, but humility about what those methods can realistically accomplish. The heterogeneity revolution must go hand-in-hand with the open science and replicability movements, with the combined aims of not only exploring sources of heterogeneity but making sure that claims about heterogeneity can stand the test of time.

References

Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education, 19*(3). https://doi.org/10.1080/10691898.2011.11889620

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94–107. https://doi.org/10.1037/0021-9010.90.1.94

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123–129. https://doi.org/10.1016/S0165-1765(03)00032-6

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. SAGE.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research, 28*(2), 186–208. https://doi.org/10.1177/0049124199028002003

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLOS ONE, 14*(12), e0225826. https://doi.org/10.1371/journal.pone.0225826

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Baranger, D. A., Finsaas, M., Goldstein, B., Vize, C., Lynam, D., & Olino, T. (2022). *Tutorial: Power analyses for interaction effects in cross-sectional regressions*. PsyArXiv. https://doi.org/10.31234/osf.io/5ptd7

Bartlett, J. (2015, February 20). Interval regression with heteroskedastic errors. *The Stats Geek*. https://thestatsgeek.com/2015/02/20/interval-regression-with-heteroskedastic-errors/

Becker, J-M., Rai, A., Voleckner, F., and Ringle, C. (2013). Discovering unobserved heterogeneity in structural equation models, *MIS Quarterly, 37*(3), 665-694. http://misq.org/discoveringunobserved-heterogeneity-in-structural-equation-models-to-avert-validity-threats.html

Bednar, S., & Rouse, K. (2020). The effect of physical education on children's body weight and human capital: New evidence from the ECLS-K:2011. *Health Economics, 29*(4), 393–405. https://doi.org/10.1002/hec.3990

Bellows, L., & von Hippel, P. T. (2017). *CATERPILLAR: Stata module to generate confidence intervals, Bonferroni-corrected confidence intervals, and null distribution* [Computer software]. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458360.html

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

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

Berger, J. O., Wang, X., & Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics, 24*(1), 110-129.

Berry, D. A. (2012). Multiplicities in cancer research: ubiquitous and necessary evils. *Journal of the National Cancer Institute, 104*(15), 1125-1133.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1), 215–227. https://doi.org/10.1016/S0378-3758(99)00044-0

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science*, *14*(2), 179-188.

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://doi.org/10.1037/xge0000558

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commericiali Di Firenze*, *8*, 3–62.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., Gan, K., Harvill, E., & Sarna, M. (2018). The Investing in Innovation Fund: Summary of 67 evaluations. Final Report. NCEE 2018-4013. In *National Center for Education Evaluation and Regional Assistance*. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED583834

Brand, J. E., Fletcher, J., & Torche, F. (Eds.). (2023). Disparate effects of disruptive events on children. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *7*(6).

Brown, W. A. (2019). *Lithium: A doctor, a drug, and a breakthrough*. Liveright Publishing.

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), Article 8. https://doi.org/10.1038/s41562-021-01143-3

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc.

Burzykowski, T., Geubbelmans, M., Rousseau, A. J., & Valkenborg, D. (2023). Validation of machine learning algorithms. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(2), 295-297. https://doi.org/10.1016/j.ajodo.2023.05.007

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Cawley, J., Frisvold, D., & Meyerhoefer, C. (2013). The impact of physical education on obesity among elementary school children. *Journal of Health Economics*, *32*(4), 743–755. https://doi.org/10.1016/j.jhealeco.2013.04.006

Chang, S., & Berger, J. O. (2021). Comparison of Bayesian and frequentist multiplicity correction for testing mutually exclusive hypotheses under data dependence. *Bayesian Analysis*, *16*(1). https://doi.org/10.1214/20-BA1196

Chebyshev, P. L. (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, 12, 177–184.

Chiatovich, T., & Stipek, D. (2016). Instructional approaches in kindergarten: What works for whom? *The Elementary School Journal*, *117*(1), 1–29. https://doi.org/10.1086/687751

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. *Economics of Education Review*, *31*(5), 543–562. https://doi.org/10.1016/j.econedurev.2012.03.002

Cintron, D. W., Adler, N. E., Gottlieb, L. M., Hagan, E., Tan, M. L., Vlahov, D., Glymour, M. M., & Matthay, E. C. (2022). Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences. *Annals of Epidemiology*, *70*, 79–88. https://doi.org/10.1016/j.annepidem.2022.04.009

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101-129. https://doi.org/10.2307/3001666

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 54–61.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305–336. https://doi.org/10.1207/s1532799xssr0804_1

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child\times instruction interactions on first graders' literacy development. *Child Development*, *80*(1), 77–100.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127. https://doi.org/10.1037/h0076829

Damme, K. S. F., & Mittal, V. A. (2023, March). *Call for papers: Managing clinical heterogeneity in psychopathology: Perspectives from brain research.* https://www.apa.org/pubs/journals/abn/managing-clinical-heterogeneity-psychopathology

Datar, A., & Sturm, R. (2004). Physical education in elementary school and body mass index: Evidence from the early childhood longitudinal study. *American Journal of Public Health, 94*(9), 1501–1506.

De Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science, 3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Dmitrienko, A., & D'Agostino, R. B. (2018). Multiplicity considerations in clinical trials. *New England Journal of Medicine, 378*(22), 2115-2122.

Domingue, B. W., Kanopka, K., Trejo, S., Rhemtulla, M., & Tucker-Drob, E. M. (2022). Ubiquitous bias and false discovery due to model misspecification in analysis of statistical interactions: The role of the outcome's distribution and metric properties. *Psychological Methods*. https://doi.org/10.1037/met0000532

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, *26*(3), 438–445. https://doi.org/10.1287/mksc.1060.0263

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212. https://doi.org/10.1177/014662169602000302

Feynman, R. P. (1974, June 14). *Cargo cult science.* Commencement, California Institute of Technology, Pasadena, CA. https://calteches.library.caltech.edu/51/2/CargoCult.htm

Finn, J. D., Achilles, C. M., Bain, H. P., Folger, J., Johnston, J. M., Nan Lintz, M., & Word, E.

R. (1990). Three years in a small class. *Teaching and Teacher Education*, *6*(2), 127–136. https://doi.org/10.1016/0742-051X(90)90030-9

Fisher, D., Harris, R., Bradburn, M., Deeks, J., Harbord, R., Altman, D., Steichen, T., Sterne, J., & Higgins, J. (2006). *METAN: Stata module for fixed and random effects meta-analysis* (Statistical Software Components S456798). Boston College Department of Economics. Revised July 15, 2024. https://ideas.repec.org/c/boc/bocode/s456798.html

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Ganzach, Y., Saporta, I., & Weber, Y. (2000). Interaction in linear versus logistic models: A substantive illustration using the relationship between motivation, ability, and performance. *Organizational Research Methods*, *3*(3), 237–253. https://doi.org/10.1177/109442810033002

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, *31*(1), 79–88. https://doi.org/10.1016/j.cpr.2010.09.007

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331. https://doi.org/10.1198/000313006X152649

Goldhaber, D., Kane, T., McEachin, A., Morton, E., Patterson, T., & Staiger, D. (2022). *The Consequences of Remote and Hybrid Instruction During the Pandemic* (w30010; p. w30010). National Bureau of Economic Research. https://doi.org/10.3386/w30010

Golino, H., Christensen, A. P., & Nesselroade, J. R. (2022). *Towards a psychology of individuals: The ergodicity information index and a bottom-up approach for finding generalizations*. https://doi.org/10.31234/osf.io/th6rm

Grissom, J. A., & Redding, C. (2016). Discretion and disproportionality: Explaining the underrepresentation of high-achieving students of color in gifted programs. *AERA Open*, *2*(1). https://eric.ed.gov/?id=EJ1194583

Haim, A., Prihar, E., & Heffernan, N. (2022). Toward improving effectiveness of crowdsourced, on-demand assistance from educators in online learning platforms. *Educational Data Mining Conference*. https://par.nsf.gov/biblio/10374331-toward-improving-effectiveness-crowdsourced-demand-assistance-from-educators-online-learning-platforms

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Hayes, S. C., Ciarrochi, J., Hofmann, S. G., Chin, F., & Sahdra, B. (2022). Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement. *Behaviour Research and Therapy*, *156*, 104155. https://doi.org/10.1016/j.brat.2022.104155

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185-207. https://doi.org/10.3758/BF03212979

Hedges, L. V., & Olkin, I. (2014). *Statistical Methods for Meta-Analysis*. Academic Press.

Hedges, L. V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350
Hewitt, J. K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics, 42,* 1–2. http://dx.doi.org/10.1007/s10519-011-9504-z
Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186
Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLOS Biology*, *14*(11), e2000995. https://doi.org/10.1371/journal.pbio.2000995
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70. https://www.jstor.org/stable/4615733
Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences*, *121*(32), e2403490121. https://doi.org/10.1073/pnas.2403490121
Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods*, *48*(4), 1308–1317. https://doi.org/10.3758/s13428-015-0624-x
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124
Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*(2), 348–357. https://doi.org/10.1037/0033-2909.117.2.348
Jackson, E., & Page, M. E. (2013). Estimating the distributional effects of education reforms: A look at Project STAR. *Economics of Education Review*, *32*, 92–103. https://doi.org/10.1016/j.econedurev.2012.07.017
James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *1,* 361-379.
Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250. https://doi.org/10.3368/jhr.44.1.223
Johnston, C., Lahey, B. B., & Matthys, W. (2013). Editorial policy for candidate gene studies. *Journal of Abnormal Child Psychology, 41*(4), 511–514. https://doi.org/10.1007/s10802-013-9741-0
Kalton, G. (1983). *Introduction to survey sampling* (No. 35). Sage Publications.
Kang, S.-M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement*, *29*(2), 87–105. https://doi.org/10.1177/0146621604272737
Kosorok, M. R., & Laber, E. B. (2019). Precision medicine. *Annual Review of Statistics and Its Application*, *6*(1), 263–286. https://doi.org/10.1146/annurev-statistics-030718-105251
Krueger, A. B., & Whitmore, D. (2001). Would smaller classes help close the black-white achievement gap?. *Princeton University Industrial Relations Section*, Working Paper #451. http://arks.princeton.edu/ark:/88435/dsp01w66343627

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE.
Long, J. S., & Mustillo, S. A. (2021). Using predictions and marginal effects to compare groups in regression models for binary outcomes. *Sociological Methods & Research*, *50*(3), 1284–1320. https://doi.org/10.1177/0049124118799374
Marquart, J., Cen, E. Y., & Prasad, V. (2018). Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. *JAMA Oncology*, *4*(8), 1093–1098. https://doi.org/10.1001/jamaoncol.2018.1660
Maslej, M. M., Furukawa, T. A., Cipriani, A., Andrews, P. W., & Mulsant, B. H. (2020). Individual differences in response to antidepressants: A meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry*, *77*(6), 607. https://doi.org/10.1001/jamapsychiatry.2019.4815
Maslej, M. M., Furukawa, T. A., Cipriani, A., Andrews, P. W., Sanches, M., Tomlinson, A., Volkmann, C., McCutcheon, R. A., Howes, O., Guo, X., & Mulsant, B. H. (2021). Individual differences in response to antidepressants: A meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry*, *78*(5), 490. https://doi.org/10.1001/jamapsychiatry.2020.4564
McClelland, G. H., Irwin, J. R., Disatnik, D., & Sivan, L. (2017). Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016). *Behavior Research Methods*, *49*(1), 394–402. https://doi.org/10.3758/s13428-016-0785-2
McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376–390. https://doi.org/10.1037/0033-2909.114.2.376
McManus, R. M., Young, L., & Sweetman, J. (2023). Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231186615. https://doi.org/10.1177/25152459231186615
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
Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. J., … Duckworth, A. L. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, *600*(7889), 478–483. https://doi.org/10.1038/s41586-021-04128-4
Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., Bogard, J. E., Buttenheim, A., Chabris, C. F., Chapman, G. B., Choi, J. J., Dai, H., Fox, C. R., Goren, A., Hilchey, M. D., … Duckworth, A. L. (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, *118*(20), e2101165118. https://doi.org/10.1073/pnas.2101165118
Mize, T. D. (2019). Best practices for estimating, interpreting, and presenting nonlinear interaction effects. *Sociological Science*, *6*, 81–117. https://doi.org/10.15195/v6.a4
Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*(2), 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x
Moon, H. R., & Perron, B. (2006). Seemingly unrelated regressions. *The New Palgrave Dictionary of Economics*. https://doi.org/10.1057/978-1-349-95121-5_2296-1
Morgan, P. L., & Hu, E. H. (2023). Fixed effect estimates of student-teacher racial or ethnic matching in U.S. elementary schools. *Early Childhood Research Quarterly*, *63*, 98–112. https://doi.org/10.1016/j.ecresq.2022.11.003
Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47-55.
Murphy, K. R., & Russell, C. J. (2017). Mend it or end it: Redirecting the search for interactions in the organizational sciences. *Organizational Research Methods, 20*, 549–573. https://doi.org/10.1177/1094428115625322
Mushava, J., & Murray, M. (2024). Comprehensive credit scoring datasets for robust testing: Out-of-sample, out-of-time, and out-of-universe evaluation. *Data in Brief, 54*, 110262. https://doi.org/10.1016/j.dib.2024.110262
Newson, R. B. (2010). Frequentist q-values for multiple-test procedures. *The Stata Journal*, *10*(4), 568-584.
Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the tennessee class size experiment. *American Educational Research Journal*, *37*(1), 123–151. https://doi.org/10.3102/00028312037001123
O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2019). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and Psychology*, *34*(1), 19–37. https://doi.org/10.1007/s10869-018-9539-8
Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294
Öngür, D., & Bauchner, H. (2020). Notice of retraction: Maslej et al. individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry*. 2020;77(6):607-617. *JAMA Psychiatry*, *77*(8), 786. https://doi.org/10.1001/jamapsychiatry.2020.2026
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716
Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*(4), 859-866.
Pedder, D. (2006). Are small classes better? Understanding relationships between class size, classroom processes and pupils' learning. *Oxford Review of Education, 32*(2), 213–234. https://doi.org/10.1080/03054980600645396
Persad, G., Peek, M. E., & Emanuel, E. J. (2020). Fairly prioritizing groups for access to COVID-19 Vaccines. *JAMA*, *324*(16), 1601. https://doi.org/10.1001/jama.2020.18513
Phaf, R. H., & Rotteveel, M. (2023). An audience facilitates facial feedback: A social-context hypothesis reconciling original study and nonreplication. *Psychological Reports*, 003329412311539. https://doi.org/10.1177/00332941231153975
Prasad, V. (2016). Perspective: The precision-oncology illusion. *Nature*, *537*(7619), Article 7619. https://doi.org/10.1038/537S63a
Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic

ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3–45.
Reardon, S. F., & Stuart, E. A. (2017). Editors' introduction: Theme issue on variation in treatment effects. *Journal of Research on Educational Effectiveness*, *10*(4), 671–674. https://doi.org/10.1080/19345747.2017.1386037
Robinson, C., & Schumacker, R. E. (2009). Interaction effects: Centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints*, *35*(1), 6–11.
Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, *6*(1), 15-32. https://doi.org/10.1214/ss/1177011926
Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2), 251524592110073. https://doi.org/10.1177/25152459211007368
Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 451–469. https://doi.org/10.1111/j.1467-9868.2012.01047.x
Sainani, K. (2010). Misleading comparisons: The fallacy of comparing statistical significance. *PM&R*, *2*(6), 559–562. https://doi.org/10.1016/j.pmrj.2010.04.016
Schanzenbach, D. W. (2006). What have researchers learned from Project STAR?. *Brookings Papers on Education Policy, (9)*, 205-228. https://www.jstor.org/stable/20067282
Schudde, L. (2018). Heterogeneous effects in education: The promise and challenge of incorporating intersectionality into quantitative methodological approaches. *Review of Research in Education*, *42*(1), 72–92. https://doi.org/10.3102/0091732X18759040
Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*(7), 2144–2162. https://doi.org/10.1016/j.jspi.2005.08.031
Sekhon, J. (2009). The Neyman—Rubin model of causal inference and estimation via matching methods. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford Handbook of Political Methodology* (1st ed., pp. 271–299). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0011
Self, W. H. (2021). Comparative effectiveness of Moderna, Pfizer-BioNTech, and Janssen (Johnson & Johnson) vaccines in preventing COVID-19 hospitalizations among adults without immunocompromising conditions—United States, March–August 2021. *MMWR. Morbidity and Mortality Weekly Report*, *70*. https://doi.org/10.15585/mmwr.mm7038e1
Simonsohn, U. (2024). Interacting with curves: How to validly test and probe interactions in the real (nonlinear) world. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231207787. https://doi.org/10.1177/25152459231207787
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.
Smith, B., & Sechrest, L. (1991). Treatment of aptitude × treatment interactions. *Journal of Consulting and Clinical Psychology*, *59*(2), 233–244. https://doi.org/10.1037/0022-006X.59.2.233
Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How many participants do i need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231178728. https://doi.org/10.1177/25152459231178728

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, *24*(3), 97–118. https://doi.org/10.1257/jep.24.3.97

Szabo, L. (2018, September 13). Much touted for cancer, 'precision medicine' often misses the target. *Kaiser Health News*. https://khn.org/news/is-precision-medicine-the-answer-to-cancer-not-precisely/

Tipton, E., & Olsen, R. B. (2022). Enhancing the generalizability of impact studies in education. Toolkit. NCEE 2022-003. *National Center for Education Evaluation and Regional Assistance*.

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. De Leeuw, & B. West (Eds.), *Experimental Methods in Survey Research* (1st ed., pp. 435–456). Wiley. https://doi.org/10.1002/9781119083771.ch22

Townshend, P. (1971, June 25). *Won't Get Fooled Again* [Song recorded by The Who].

Viswesvaran, C., & Sanchez, J. I. (1998). Moderator Search in meta-Analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement*, *58*(1), 77–87. https://doi.org/10.1177/0013164498058001007

von Hippel, P. T. (2015, July 5). Linear vs. logistic probability models: Which is better, and when? *Statistical Horizons*. https://statisticalhorizons.com/linear-vs-logistic/

von Hippel, P. T. (2017, March 8). When can you fit a linear probability model? More often than you think. *Statistical Horizons*. https://statisticalhorizons.com/when-can-you-fit/

von Hippel, P. T. (2019, June 4). *Is summer learning loss real? How I lost faith in one of education research's classic results*. Education Next. https://www.educationnext.org/is-summer-learning-loss-real-how-i-lost-faith-education-research-results/

von Hippel, P. T. (2021). The effect of smaller classes on infection-related school absence: Evidence from the Project STAR randomized controlled trial. In *Annenberg Institute for School Reform at Brown University* (EdWorkingPaper). Annenberg Institute for School Reform at Brown University. https://eric.ed.gov/?id=ED613646

von Hippel, P. T., & Bellows, L. (2018a). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*. https://doi.org/10.1016/j.econedurev.2018.01.005

von Hippel, P. T., & Bellows, L. (2018b, May 8). Rating teacher-preparation programs: Can value-added make useful distinctions? *Education Next*. https://www.educationnext.org/rating-teacher-preperation-programs-value-added-make-useful-distinctions/

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science*, *6*(3). http://dx.doi.org/10.15195/v6.a3

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

Wallach, J. D., Sullivan, P. G., Trepanowski, J. F., Sainani, K. L., Steyerberg, E. W., & Ioannidis, J. P. (2017). Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Internal Medicine, 177*(4), 554-560. https://doi.org/10.1001/jamainternmed.2016.9125

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika, 84*(2), 419-427.

Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). Multiple comparisons and multiple tests using SAS. *SAS Institute*.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, *37*(4), 531–559. https://doi.org/10.1177/0049124109335735

Williams, R. (2010). Fitting heterogeneous choice models with Oglm. *The Stata Journal*, *10*(4), 540–567. https://doi.org/10.1177/1536867X1101000402

Winkler, R. (2024, January 31). 23andMe's fall from $6 billion to nearly $0. *Wall Street Journal*. https://www.wsj.com/health/healthcare/23andme-anne-wojcicki-healthcare-stock-913468f4

Workman, J. M., von Hippel, P. T., & Merry, J. (2022). Summer learning findings often fail to replicate, even in recent data. *Unpublished Manuscript*.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L., & Tahk, A. M. (2019). Moderation of classic social psychological effects by demographics in the U.S. adult population: New opportunities for theoretical advancement. *Journal of Personality and Social Psychology, 117*(6), e84–e99. https://doi.org/10.1037/pspa0000171