AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

http://dx.doi.org/10.1037/met0000368

Disentangling Effect Size Heterogeneity in Meta-Analysis: A Latent Mixture Approach

Nan Zhang
American University

Mo Wang
University of Florida

Heng Xu
American University

Abstract

An important task of meta-analysis is to observe, quantify, and explain the heterogeneity across the reported effect sizes of primary studies. A primary issue that challenges this task is the myriad of subtle factors that could have contributed to the observed heterogeneity. We leveraged the recent advances in theoretical machine learning to develop a novel latent mixture-based method for disentangling effect-size heterogeneity in meta-analysis. Mathematical analysis and simulation studies were carried out to demonstrate that, when the observed heterogeneity stems from more than 1 factor, our method can attain a substantially higher statistical power than the traditional methods for moderator analysis without requiring researchers to make judgment calls on which factors to consider or correct for in analyzing the observed heterogeneity. We also conducted a case study with real-world data to show how our method may be used to address long-standing inconsistencies in the literature.

Translational Abstract

An important task of meta-analysis is to explain the heterogeneity among primary studies. However, it is often a challenge for researchers to delineate the myriad of subtle factors that could have contributed to the observed heterogeneity. We leveraged the recent advances in theoretical machine learning, specifically the efficient decomposition of Gaussian mixture distributions, to develop a novel latent mixture-based method for disentangling heterogeneity in meta-analysis. As demonstrated by mathematical analysis and simulation studies for moderator estimation, our method can attain substantially higher statistical power than the traditional methods without requiring researchers to make judgment calls on which factors to consider or correct for in analyzing the observed heterogeneity.

Keywords: meta-analysis, moderator analysis, mixture model, latent class analysis

Supplemental materials: http://dx.doi.org/10.1037/met0000368.supp

A key task of meta-analysis is to observe and assess the heterogeneity[1] among primary studies. When the observed heterogeneity exceeds what could be explained by artifactual factors (e.g., sampling error), it becomes important for a researcher to theorize and test *what could have caused the residual heterogeneity* (Thompson, 1994). Such investigations have led to scientific breakthroughs in many disciplines such as psychology (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006), epidemiology (Berlin, 1995), and medicine (Higgins, Thompson, Deeks, & Altman, 2003). For example, upon inspecting the unexplained heterogeneity in the effect of feedback interventions (FI) on performance, Kluger and DeNisi (1996) developed the feedback intervention theory (FIT) that predicts several novel moderators for the FI-performance relationship, such as FI cues and task characteristics, which have never been examined in any primary

Nan Zhang, Kogod School of Business, American University; Mo Wang, Warrington College of Business, University of Florida; Heng Xu, Kogod School of Business, American University.

Nan Zhang and Heng Xu was supported in part by the National Science Foundations under Grant 1850605 and by the Defense Advanced Research Projects Agency under Grant HR00111920023. Mo Wang's work on this research was supported in part by the Lanzillotti-McKethan Eminent Scholar Endowment. No material presented in the article has been previously disseminated.

Correspondence concerning this article should be addressed to Nan Zhang, Kogod School of Business, American University, 4400 Massachusetts Avenue Northwest, Washington, DC 20016. E-mail: nzhang@american.edu

[1] In meta-analysis, heterogeneity often refers to the variation of "real" effects across primary studies. Note that the meaning of "real" should not be confused with "ground truth," because methodological variations across studies are considered part of heterogeneity in some fields (Higgins et al., 2003) but not others (Hunter & Schmidt, 2004). This difference is moot in this article, as our goal is not to measure heterogeneity but to explain what caused the unexplained part of it. Thus, we use "*heterogeneity*" to refer to any variation *not* accounted for by factors already considered in a meta-analysis (e.g., sampling errors, corrections of measurement artifacts, etc.).

study. These moderators were then tested on a meta-analytic dataset through a process that falls under the umbrella term of *moderator analysis* (Hunter & Schmidt, 2004), and were found to account for a substantial portion of the observed but previously unexplained heterogeneity.

Despite the importance of disentangling heterogeneity, actually doing so in a meta-analysis can be challenging. A notable reason is the myriad of subtle factors that could have contributed to the observed heterogeneity and therefore affected the outcome of moderator analysis. Some well-known factors include the need for finer gradations[2] of variable coding (Hunter & Schmidt, 2004, p. 180), the presence of certain availability biases (McShane, Böckenholt, & Hansen, 2016), or the potential influence of questionable research practices (Simonsohn, Nelson, & Simmons, 2014). For example, Kluger and DeNisi (1996) had to consider excluding over 15% of the existing observations (91 out of 607) after noting that they were all reported by one researcher (Mikulincer) yet their inclusion sharply increased the overall heterogeneity.[3] The presence of these diverse factors poses a dilemma for meta-analysis researchers. Failure to consider them could drastically reduce the statistical power of moderator analysis (Stone-Romero & Anderson, 1994), which is already known to be lower than desired due to the often small number of primary studies (Hunter & Schmidt, 2004, p. 70). Yet identifying these factors and determining which ones to consider (and correct for) is not only difficult but often subjective. As a result, when attempting to disentangle heterogeneity in meta-analysis, researchers often had to resort to ad hoc procedures that vary considerably even within a field (Fletcher, 2007; Naaktgeboren et al., 2014), and rely on judgment calls to make important decisions such as whether the observed yet unexplained heterogeneity is "natural" or warrants further investigation (Higgins et al., 2003).

The goal of this article is to ease this heterogeneity-disentanglement process in meta-analysis by developing an analytical method that helps researchers attain a substantially higher statistical power in moderator analysis *without* having to rely on judgment calls about which factors to consider or correct for in analyzing the observed heterogeneity. Specifically, we leverage a recent breakthrough in theoretical machine learning to develop *latent mixture-based moderator analysis*, a novel "data-driven" meta-analytic method that serves as a preprocessing step for moderator analysis. The conceptual underpinning of our method is akin to latent-variable mixture modeling (McLachlan & Basford, 1988; McLachlan & Peel, 2004) in analyzing individual-level data (e.g., Bauer & Curran, 2003; Wang & Hanges, 2011). That is, we consider the input data (in the case of meta-analysis, the effect sizes reported in primary studies) as samples drawn from a *mixture* of multiple Gaussian distributions (each of which is referred to as a *mixture component*) rather than a single Gaussian distribution. Unlike traditional moderator-analysis methods that require researchers to theorize the factors that cause the heterogeneity among mixture components, our method starts by deploying an automated algorithm called *mixture decomposition* to infer the nature of each component from the input data *regardless of* the underlying factor[4] responsible for creating the component. As we will elaborate later, a researcher can then leverage the auto-decomposed mixture components to examine a hypothesized moderating effect in a more effective manner.

Before presenting an example that demonstrates the utility of our mixture-based method, we first briefly address an intriguing question surrounding its novelty: *Why* was the use of mixture modeling in meta-analysis exceedingly rare,[5] despite its extensive use in analyzing individual-level data and the wide recognition in meta-analysis that the distribution of reported effect sizes often has little resemblance with a single Gaussian distribution (Higgins, Thompson, & Spiegelhalter, 2009; Micceri, 1989)? A likely reason is the *computational challenge* associated with decomposing a mixture distribution when its components largely *overlap* with each other. Consider an intuitive illustration in Figure 1: While a casual inspection can reveal the two "cleanly" separated mixture components in Figure 1a, when the components overlap with each other like in Figure 1b, the decomposition becomes much less obvious. This challenge is exemplified in meta-analysis given the small effect sizes in fields like social psychology (Richard, Bond, & Stokes-Zoota, 2003) and management (Paterson, Harms, Steel, & Credé, 2016) and the limited number of primary studies, which together make a "clean" separation of multiple mixture components less likely than in individual-level data. Note that the dominating solution for mixture modeling in psychology and other social sciences, the expectation maximization (EM) algorithm, is known to have degraded accuracy when the mean difference between two mixture components is smaller than their standard deviations (Redner & Walker, 1984). The main reason behind this problem is the need for EM to accurately infer the mixture component each input sample belongs to (i.e., the "M" step) before estimating the parameters for each component (i.e., the "E" step), yet inferring the exact component is obviously infeasible for samples in the overlapping region, like an effect size of 0.05 in Figure 1b. Indeed, for the same reason, numerous EM-like algorithms in the statistics and computer science literature (e.g., Dasgupta, 1999) have a "separation requirement" for properly identifying the mixture components.

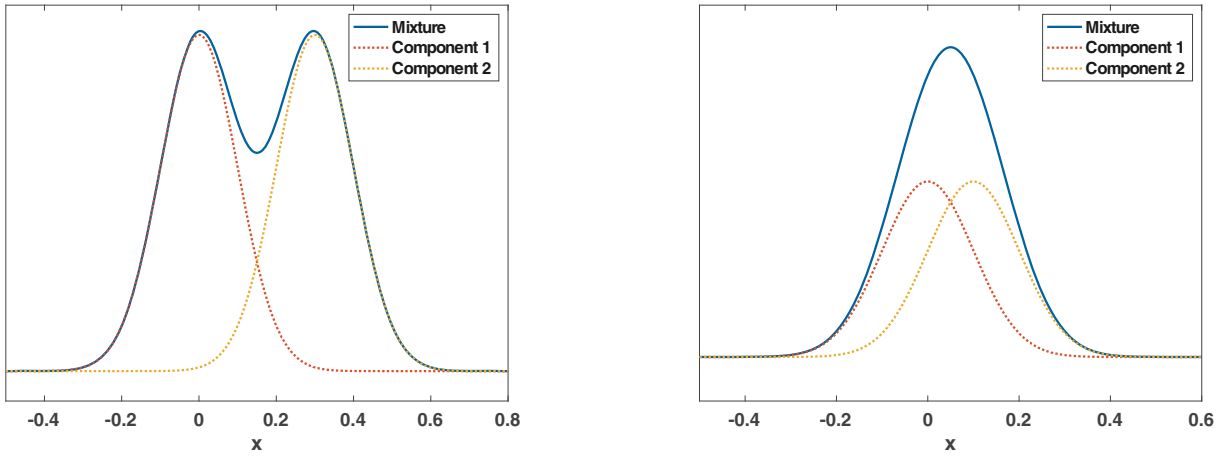
This computational challenge was addressed in 2010 by a trio of breakthroughs in theoretical machine learning (Belkin & Sinha, 2010; Kalai, Moitra, & Valiant, 2010; Moitra & Valiant, 2010). While their respective solutions differ, an idea they share is to *directly* infer the specifics of each mixture component (i.e., its mean, standard deviation, and weight in the mixture distribution) *without* having to first determine the component affiliation for each input sample. By doing so, the state-of-the-art algorithms can now accurately and efficiently decompose a mixture distribution even when the components overlap almost entirely with each other (Bandi, Bertsimas, & Mazumder, 2019; Belkin & Sinha, 2015; Kalai, Moitra, & Valiant, 2012). Drawing from this technical breakthrough, we develop in this article a novel mixture-decomposition algorithm specifically for meta-analysis, and dem-

[2] e.g., from high/low to a numerical scale.

[3] Perhaps because they all feature a study-level design that differs considerably from the rest of the primary studies (Kluger & DeNisi, 1996).

[4] Note that such a factor could be the core of moderator analysis, e.g., if it were a moderator variable that caused the observed effect sizes to follow different distributions. Or the factors could be an extraneous one that requires correction, e.g., if a mixture component emerged solely due to questionable research practices.

[5] A few notable exceptions (Nord et al., 2017; Schlattmann et al., 2015) were all in medicine-related fields for limited purposes such as outlier detection.



a) Non-overlapping mixture components (easy to decompose)

b) Largely overlapping mixture components (difficult to decompose)

Figure 1. Illustration of the computational challenge facing mixture modeling in meta-analysis. The left figure depicts two (almost) nonoverlapping mixture components (Gaussian distributions with mean 0 and 0.3, standard deviation 0.1). A visual inspection can easily decompose the observed mixture distribution (solid line) into its two components (dotted lines). In contrast, the right figure depicts two largely overlapping components (Gaussian distributions with mean 0 and 0.1, standard deviation 0.1). The two distributions now form only a single peak, making the decomposition a nontrivial challenge. See the online article for the color version of this figure.

onstrate the usage of this algorithm for disentangling the observed heterogeneity in moderator analysis.

We use the example in Figure 2 to illustrate *how* a mixture-based method can increase the statistical power of moderator analysis without requiring researchers to explicitly identify and address the factors (besides the hypothesized moderators) that contribute to the observed heterogeneity. Note from Figure 2a and 2b that the observed effect-size distributions for moderator levels LOW and HIGH have the exact same mean and standard deviation. As a result, a traditional meta-analytic moderator analysis likely returns a *null* result, because these methods (e.g., metaregression; Thompson & Higgins, 2002; Q-statistic; Hedges & Pigott, 2004) are designed to compare the *mean* effect size across moderator levels. Nonetheless, even a casual inspection of Figure 2a and 2b would reveal that the moderator likely *does* affect the observed effect size, as it produces a bimodal distribution when moderator is HIGH and a unimodal one otherwise. One likely explanation here is the presence of a latent factor that contributes to the observed heterogeneity but is not captured by the moderator analysis. For example, perhaps like in the aforementioned case (i.e., Kluger & DeNisi, 1996), one component in Figure 2b is formed by primary studies adopting an unusual study-level design, and thus should have been excluded or addressed through a hierarchical moderator analysis (Hunter & Schmidt, 2004, p. 424). Alternatively, perhaps the coding of the moderator variable should have been finer-grained, and the two components in Figure 2b (i.e., moderator = HIGH) should have been coded differently (e.g., as "HIGH" and "VERY HIGH" instead). Regardless of the underlying reason, the presence of this latent factor "masks" the real moderating effect and drastically reduces the statistical power of moderator analysis.

Now consider how our latent mixture-based method can help address this issue without requiring a researcher to identify the latent factor. Given the overall distribution of the reported effect sizes (i.e., solid line in Figure 2c), our method first calls upon an automated mixture decomposition algorithm to decompose this distribution into its three components, all dotted lines in Figure 2c. Note that the algorithm takes only the overall effect-size distribution[6] as input, and is thus oblivious to, and unaffected by, the moderator variable(s). In other words, if a moderator variable truly had no effect on the reported effect size, then the moderator level should be independent of the *component affiliation* of the effect size (i.e., the posterior probability for the effect size to belong to each mixture component). In the case of Figure 2, we could easily see this is not the case because those effect sizes with moderator = LOW are more likely to belong to Component 2, while those with moderator = HIGH more likely belong to Components 1 or 3. This simple observation would allow our method to achieve a higher statistical power than the traditional methods without requiring a researcher to explicitly identify the latent factor (that separates the two components within HIGH).

As can be seen from this example, a key benefit of using mixture decomposition in moderator analysis is to *isolate* the effect of the latent factor by disentangling the various mixture components. To understand why, consider how the latent factor affects the statistical power of traditional moderator-analysis methods (e.g., metaregression), which are designed to compare effect sizes across moderator levels. With these methods, the heterogeneity introduced by the latent factor (i.e., increasing the effect size from

[6] After proper corrections for sampling error, artifact variability, etc.

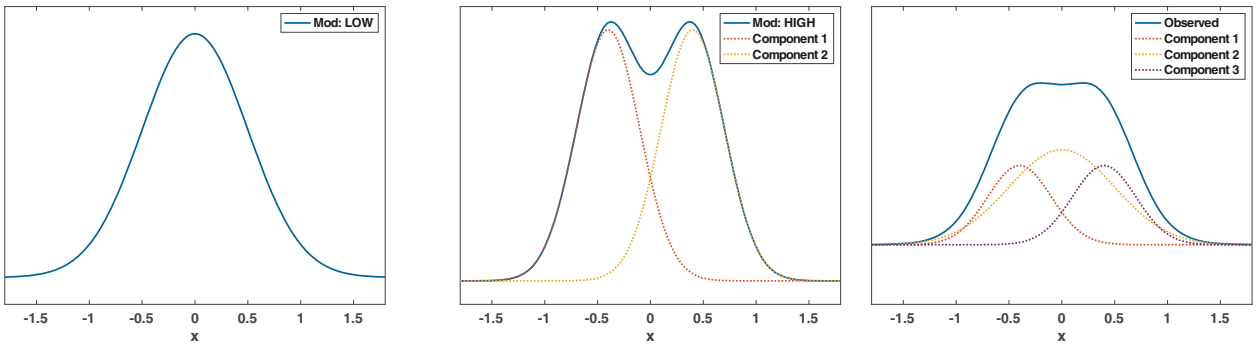


Figure 2. An illustration of how mixture decomposition increases the statistical power of moderator analysis. The left figure depicts the distribution of effect sizes when the value of a moderator variable is LOW. It is a Gaussian distribution with mean 0 and standard deviation 0.5. The middle figure depicts the effect-size distribution when the moderator level is HIGH. It is a mixture of two Gaussian distributions with mean −0.4 and 0.4, respectively, with each having a standard deviation of 0.3. Note that the mean of this mixture distribution is 0 and its standard deviation is $\sqrt{0.3^2 + 0.4^2} = 0.5$, both exactly the same as the case where moderator = LOW. The right figure depicts the overall distribution of observed effect sizes when half of the primary studies have moderator = LOW and the other half have moderator = HIGH. It is a mixture of three Gaussian distributions, with Components 1 and 3 corresponding to moderator = HIGH and Component 2 with moderator = LOW. In all three figures, the solid line represents the (probability density function of the) observed distribution, while the dotted lines represent its mixture components (if applicable). See the online article for the color version of this figure.

Component 1 to Component 3 in Figure 2c when moderator = HIGH) is mixed with, and therefore masks, the true heterogeneity caused by the moderator variable (i.e., lowering the effect size from Component 2 to Component 1 in Figure 2c when moderator changes from LOW to HIGH). This causes the traditional methods to fail unless the researcher identifies the latent factor. With our mixture-based method, the focal measure of comparison (across moderator levels) is *not* the raw effect sizes but their component affiliations. This change creates a *distinction* between the heterogeneity introduced by the latent factor (i.e., changing the affiliation from Component 1 to Component 3 when moderator = HIGH) and the true heterogeneity caused by the moderator variable (i.e., changing the affiliation from Component 2 to Component 1 when moderator changes from LOW to HIGH), so the two sources of heterogeneity are no longer mixed with each other. As we will elaborate in the article, this newly introduced distinction is what enables our method to properly assess the moderating effect without requiring researchers to explicitly identify the latent factor.

The rest of the article is organized as follows. First, we review the related literature from two perspectives: (a) the history and recent advances in mixture modeling, and (b) the methodological development in meta-analysis as related to disentangling the observed heterogeneity. We then introduce our latent mixture-based moderator analysis method and describe results from Monte Carlo simulation studies that compare the statistical power of our new method with the traditional approaches. We also present a case study with real-world data, using our method to address a long-standing inconsistency on how intrateam trust affects team performance. The article finishes with a discussion of the limitations of our method and the potential directions for its future development.

Literature Review

Gaussian Mixture Model

The statistical theory underlying mixture modeling has been known since the 19th century:[7] If there is heterogeneity in the data-generation process, then the observed data, in our case effect sizes reported in primary studies, are bound to be drawn from a *mixture* of multiple distributions (Böhning, 1999). Such a mixture distribution would have samples drawn with probability w_i from its i-th component distribution ($i \in [1, m]$), where $w_1 + \ldots + w_m = 1$. While there is no inherent constraint in mixture modeling on what type of distribution each component follows, most existing work considered the case where the component distributions are Gaussian (see Redner & Walker, 1984 and the citations within). The factor separating the Gaussian components can be obvious in certain cases (e.g., gender for the distribution of adult heights), but latent in others (e.g., for the famous Pearson's crab data; Pearson, 1894). This naturally leads to the research question of how to *decompose* the observed mixture distribution into its Gaussian components, a problem often referred to as *latent variable mixture modeling*.

Most classic solutions to the problem make assumptions about the underlying component distributions. Unfortunately, these assumptions are often too strong to hold in the context of meta-analysis. For example, EM (Dempster, Laird, & Rubin, 1977), the most popular method for mixture modeling in psychology and social sciences, is long known for requiring a "proper separation"

[7] See discussions of Pearson's work in Améndola, Faugère, and Sturmfels, (2016).

between different components (Hosmer, 1973). Even though the minimum necessary degree of separation was not fully established for the EM algorithm due to its iterative nature (Balakrishnan, Wainwright, & Yu, 2017), a well-understood rule-of-thumb is that EM cannot produce useful results when the mean difference between two components is less than their standard deviation (Redner & Walker, 1984, p. 213). This issue persists in a more recent line of research pioneered by Dasgupta (1999), which provides rigorous guarantees on the accuracy of mixture decomposition under a formalized separation assumption[8] between all pairs of mixture components (e.g., Achlioptas & McSherry, 2005; Sanjeev & Kannan, 2001; Vempala & Wang, 2004). As discussed in the introduction, the small effect sizes studied in a meta-analysis, coupled with the limited number of primary studies, makes such separation assumptions unlikely to hold. It is noteworthy that a long-standing line of research in statistics, known as the *method of moments* for mixture decomposition (Hopkins & Li, 2018; Lindsay, 1989; Lindsay & Basak, 1993; Wu & Yang, 2018), does not have this separation assumption but instead requires all components to share the same standard deviation. Unfortunately, this assumption of homoscedasticity is not immediately justifiable and has indeed been shown to often not hold in meta-analyses (Hedges & Olkin, 1985, pp. 11–12; Bonett, 2008).

Statisticians have long known that *none* of these assumptions is required for solving the mixture decomposition problem. Teicher (1961) provided an *identifiability proof* showing that no matter the shape and form of the mixture distribution, with enough samples, one can always learn every parameter (i.e., mean, standard deviation, and weight) of every Gaussian component to an arbitrary precision, so long as no two components are exactly the same. However, the identifiability proof in theory does not imply a methodological design in practice, and it was not clear until a trio of breakthroughs in 2010 (Belkin & Sinha, 2010; Kalai et al., 2010; Moitra & Valiant, 2010) addressed *how* one could drop the separation assumption yet still accurately recover the mixture components in practice, based on a reasonable number of samples and with limited computational resources.

While each of the breakthrough and their more recent follow-up work (e.g., Bandi et al., 2019) features a different algorithm for mixture decomposition, these algorithms share a common feature, in that none of them makes any separation assumption about the mixture components. They also share a common design scheme, in that they all infer the parameters of the Gaussian components by finding a parameter combination that minimizes a predetermined *statistical distance* between the mixture distribution computed from the parameter combination and the actually observed distribution. This distance metric varies from one work to another. For example, Kalai, Moitra, and Valiant (2010) considered the distance between two vectors, each consisting of the first $4k - 2$ statistical moments of the a distribution (where k is the number of mixture components); Belkin and Sinha (2010) and Moitra and Valiant (2010) used the distance between the probability densities of the two distributions; Bandi, Bertsimas, and Mazumder (2019) used a variation of the total variation distance metric (Levin, Peres, Wilmer, Propp, & Wilson, 2017), while Daskalakis and Kamath (2014) used the Kolmogorov–Smirnov distance metric (Daniel, 1990) between two cumulative density functions. Regardless of the statistical-distance metric used, these algorithms represent a breakthrough for mixture decomposition in practice because the number of samples they require is only *polynomial*, not exponential, to the inverse of estimation error for the component parameters (Kalai et al., 2012). Thus, we follow the same design scheme in our mixture-based moderator analysis, but customize the design of the statistical distance metric and the algorithm design (i.e., for finding the optimal parameter combination) according to the specific requirements and limitations of meta-analysis.

Disentangling Heterogeneity in Meta-Analysis

In the meta-analysis literature, many researchers have hinted or argued that the underlying effect sizes likely form a mixture of multiple distributions. For example, in studying the choice overload effect, Chernev, Böckenholt, and Goodman (2010) noted, and Simonsohn, Nelson, and Simmons (2014) concurred, that many primary studies were "*designed to document*" how the direction of an effect can be reversed by adjusting a moderator variable, effectively making the overall set of reported effect sizes a mixture of two distributions, one with a positive mean and the other negative. Hunter and Schmidt (2004) also recognized that, when a moderator variable takes on a continuum of values in primary studies, the distribution of effect sizes could be a mixture of many distributions, each with a different mean and variance. In addition, effect sizes reported in "outlier" primary studies have been treated as being drawn from a distribution with a larger variance than the rest of the studies (Beath, 2014), making the overall distribution a mixture of both.

In terms of moderator analysis to disentangle the mixture of effect-size distributions, our specific focus in this article is what is known as the task of *moderator estimation* (i.e., to determine how much of the observed heterogeneity can be attributed to a hypothesized moderator variable; Steel & Kammeyer-Mueller, 2002). There are other important tasks besides moderator estimation in moderator analysis. For example, the task of moderator detection aims to determine whether there exists a substantial amount of unexplained heterogeneity that could be attributed to one or more moderators (Whitener, 1990). The task of hierarchical moderator analysis is essential when there are multiple hypothesized moderators, so as not to confound their effects (Hunter & Schmidt, 2004, p. 424). As we will elaborate in the Discussion section, we leave the study of how to leverage mixture modeling in these tasks to future work.

For the specific task of moderator estimation, there has been a long line of methodological research developing numerous methods over multiple decades, and an almost equally long line of meta-analysis research that draws on these methods to advance scientific fields such as psychology (e.g., Judge & Piccolo, 2004; Kluger & DeNisi, 1996; Liu, Huang, & Wang, 2014). When the moderator variable is continuous, a popular method for moderator analysis is metaregression (Glass, 1977), which regresses the observed effect sizes on the coded study-level characteristics such as moderator variables. When the moderator variable is dichotomous or categorical, one can perform a *subgroup analysis* by first partitioning the primary studies into subgroups according to their moderator levels, and then conducting a separate meta-analysis for each subgroup (for comparisons). In subgroup analysis, methods

[8] See Appendix A for a detailed review of recent results in theoretical computer science.

such as Q-statistic (Hedges & Pigott, 2004) can be used to test the statistical significance of the differences between subgroups.

Unsurprisingly, many variations of these moderator-analysis methods exist, with their superiority over each other long debated, (at least partially) due to their differences on the assumed model of effect sizes. For example, both fixed- and random-effect based methods have been criticized, the former for its assumption of effect-size homogeneity (Hunter & Schmidt, 2004, p. 201) and the latter for the fact that effect sizes reported in the literature are often a convenience sample rather than a random sample of the population (Schulze, 2004, p. 41). Similarly, there have been prolonged debates on the use of ordinary linear squares versus weighted linear squares in metaregression (Hedges & Olkin, 1985, Chapter 8, pp. 168–190; Steel & Kammeyer-Mueller, 2002), with the former being criticized for its homoscedasticity assumption of sample sizes and the latter for its potential sensitivity to outliers. Regardless of the specific method, a well-recognized universal challenge for the meta-analytic moderator analysis is that its statistical power tends to be low due to two main reasons: One is the small number of primary studies, and the other is that much of the observed heterogeneity could be caused by factors other than the hypothesized moderators (Hunter & Schmidt, 2004, p. 70).

Situating our latent mixture-based method in the methodological literature of moderator analysis, it is important to note that, while our method is designed to boost the statistical power of moderator analysis by alleviating the second issue discussed above (i.e., heterogeneity caused by factors other than the hypothesized moderators), it cannot completely solve the issue because sampling errors (and other artifactual variances) are bound to form a substantial portion of the observed variation across studies, and cannot be reduced without having more primary studies.

Latent Mixture-Based Moderator Analysis

This section provides an overview of our moderator estimation method *assuming* the existence of a mixture decomposition algorithm that can properly decompose a mixture of Gaussian distributions into its respective components. We refer readers to Appendix A to Appendix D for a detailed discussion of how we addressed the unique challenges for meta-analysis and the design of our mixture decomposition algorithm.

Input and Output of the Mixture Decomposition Algorithm

Because we treat the mixture decomposition algorithm as a black box in this section, it is important to define its input and output. The input, denoted by In_{M}, is the same as in bare-bone moderator analysis. That is, given m primary studies, In_{M} consists of their (reported) effect sizes $r(es_i)$, sample sizes N_i, and estimated standard errors (i.e., sampling error) $r(se_i)$, i.e.,

$$In_{\mathrm{M}}\text{: } \{r(es_1), N_1, r(se_1)\}, \{r(es_2), N_2, r(se_2)\}, \ldots, \{r(es_m), N_m, r(se_m)\}.$$

Note that $r(es_i)$ or $r(se_i)$ could be directly reported in a primary study or derived by the meta-analysis researcher. For example, when the effect size is the Pearson's correlation coefficient, $r(se_i)$ could be estimated as $(1 - r(es_i)^2)/\sqrt{N_i - 1}$. For the sake of simplicity, we assume the m samples to be independent, and defer to future work the analysis of interstudy correlations, for example, between those conducted by the same researchers, with similar study-level characteristics, and so forth, which is often addressed through a multilevel (Cheung, 2014) or multivariate (Gleser & Olkin, 2009) meta-analysis. For the same reason, we do not consider cases with missing data, for example, when a primary study does not report a sample size N_i or a standard error $r(se_i)$. An important distinction with regard to notation is that we use $r(\cdot)$ to denote the *reported* values and reserve es_i to represent the (latent) *ground-truth* effect size for the i-th primary study.

Given In_{M} as input, the goal of the mixture decomposition algorithm is to infer the individual components that together form the distribution of es_i. Thus, the output of the algorithm, denoted by Out_{M}, contains the parameter estimates for the components:

$$Out_{\mathrm{M}}\text{: } \{w_1, \mu_1, \sigma_1\}, \{w_2, \mu_2, \sigma_2\}, \ldots, \{w_k, \mu_k, \sigma_k\},$$

where k is the number of mixture components either specified by the researcher or estimated by the algorithm, and w_i, μ_i, σ_i are the estimated weight, mean, and standard deviation for the i-th component, respectively. With an accurate mixture decomposition algorithm, we have

$$F(G) \approx w_1 \cdot F_{\mathbb{N}}(\mu_1, \sigma_1^2) + w_2 \cdot F_{\mathbb{N}}(\mu_2, \sigma_2^2) + \ \ldots \ + w_k \cdot F_{\mathbb{N}}(\mu_k, \sigma_k^2), \tag{1}$$

where G is the distribution of es_i (i.e., a mixture of multiple Gaussian distributions according to our model), $F(G)$ is the probability density function of G, and $F_{\mathbb{N}}(\mu_i, \sigma_i^2)$ represents the probability density function of a Gaussian distribution with mean μ_i and standard deviation σ_i. While the order of the mixture components is not important, for the sake of consistency, we sort the components in an increasing order of μ_i (and an increasing order of σ_i for tie-breakers).

Use of Mixture Components in Moderator Analysis

Having defined the input In_{M} and output Out_{M} of the mixture decomposition algorithm, we now discuss how Out_{M} can be used in moderator estimation. First, a meta-analysis researcher could *directly* benefit from learning Out_{M}, as it provides a holistic view of the effect-size distribution and can be used to identify obvious outliers (Nord, Valton, Wood, & Roiser, 2017; Schlattmann, Verba, Dewey, & Walther, 2015) or to inspect whether the distribution is apparently bimodal (e.g., the example of choice overload in Simonsohn et al., 2014). Nonetheless, for our purpose of moderator estimation, in order to disentangle the hypothesized moderating effect from other heterogeneity-contributing factors, we need to further infer the component affiliation of each primary study. The purpose, as illustrated with the example in Figure 2, is to inspect whether a change of moderator level "moves" a study from one component to another. While it might be tempting to deterministically associate each es_i with one component, recall from the previously discussed pitfall of the EM algorithm that different mixture components in a meta-analysis often overlap considerably with each other, making the deterministic assignment unlikely to be accurate. To address this challenge, we capture the component affiliation probabilistically rather than deterministically. Specifically, we compute from Out_{M} the posterior probability for es_i to belong to a mixture component, say the j-th one denoted by C_j:

$$\Pr\{es_i \in C_j \mid r(es_i)\} = \frac{P(r(es_i) \mid es_i \in C_j) \cdot Pr\{es_i \in C_j\}}{P(r(es_i))}. \tag{2}$$

A key observation enabling our latent mixture-based method is that In_M (i.e., the input to the mixture decomposition algorithm) contains no information about the moderator variable to be tested. Thus, if a moderator variable V truly had no effect on the effect size es_i, then V would be independent of $\Pr\{es_i \in C_j \mid r(es_i)\}$ for any $j \in [1, k]$. As such, just like how traditional moderator analysis compares effect sizes across moderator levels or correlates the effect sizes with the moderator variable, if we replace the effect size $r(es_i)$ in such analysis with $\Pr\{es_i \in C_j \mid r(es_i)\}$, the traditional methods should still not be able to reject their null hypotheses with probability higher than the significance level if the moderator V had no effect on es_i. Consider a simple example where $k = 2$, that is, the mixture decomposition algorithm returns two components C_1 and C_2. Because there is always $\Pr\{es_i \in C_1 \mid r(es_i)\} + \Pr\{es_i \in C_2 \mid r(es_i)\} = 1$, we can simply replace the effect size in existing moderator-analysis methods with $\Pr\{es_i \in C_2 \mid r(es_i)\}$ (or, equivalently, $\Pr\{es_i \in C_2 \mid r(es_i)\} - \Pr\{es_i \in C_1 \mid r(es_i)\}$). Examples here include using it as the regressand in metaregression (and regresses it over the moderator variables and potentially other control variables), comparing its mean across different levels of a categorical moderator variable, and so forth. Regardless of the method used, if the results indicate that a moderator variable V is a significant predictor of the posterior probability, then we should reject the null hypothesis[9] because the only way for V to affect the posterior probability is by affecting the effect size es_i.

Figure 3 illustrates such an example, where a change of the moderator level (from LOW to HIGH) varies not only the mean (from 0 to 0.1) but also the standard deviation (from 1 to 0.1) of the effect-size distribution. As can be seen from Figure 3b, if a method directly compares the mean effect size (i.e., $\text{mean}(es_i)$) between the moderator levels, one might not be able to properly reject the null hypothesis given the overlapping confidence intervals. In contrast, once we replace es_i with $\Pr\{es_i \in C_2 \mid r(es_i)\}$, like in Figure 3c, the moderating effect can be clearly identified from the vastly different component affiliations between the two moderator levels. This example demonstrates how our method can substantially increase the statistical power of moderator analysis by leveraging the mixture decomposition algorithm.

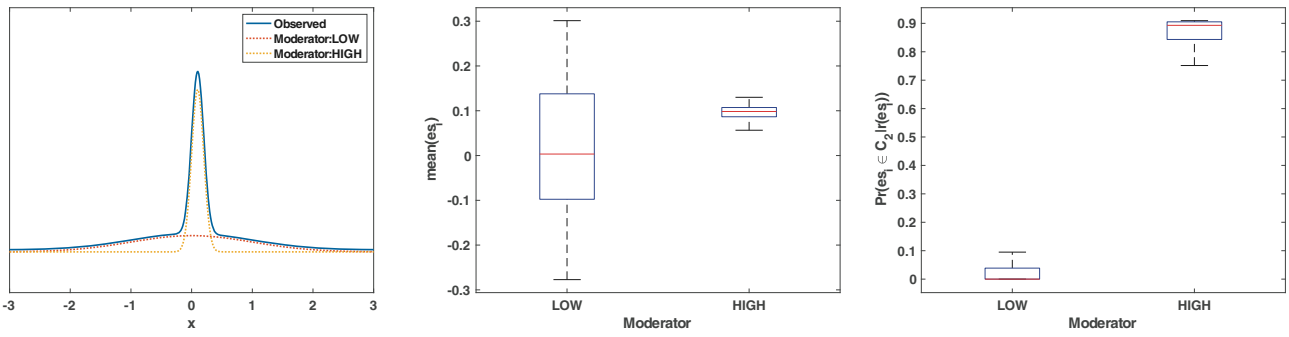
Mixture of More Than Two Components

When the mixture decomposition algorithm returns more than two components, the design of moderator analysis is subtler because it is possible for the hypothesized moderating effect to only affect a subset of the components. To this end, a seemingly simple solution is to design a holistic test of correlation between the moderator variable V and the affiliation probabilities for *all* components, for example, by regressing V over a $(k - 1)$-dimensional vector $\Pr\{es_i \in C_1 \mid r(es_i)\}, \ldots, \Pr\{es_i \in C_{k-1} \mid r(es_i)\}$. Unfortunately, this solution has two drawbacks: First, it is no longer "transparent" to the moderator analysis method. For example, while this solution is clearly a variation of metaregression, it is unclear how to integrate this solution with the method of subgroup analysis. The second, and more important, drawback is that this solution increases the likelihood of moderator analysis *capitalizing on chance* and returning a false positive result. Just like how running a metaregression with too many moderator variables might capitalize on sampling error in moderator analysis (Higgins & Thompson, 2004; Hunter & Schmidt, 2004), inspecting too many mixture components could have the same effect when one of the components by chance produces a high correlation between $\Pr\{es_i \in C_j \mid r(es_i)\}$ and V.

Fortunately, this issue of false positive is unlikely to be of serious concern in practice due to a technical reason: the sample size (i.e., the number of primary studies) available in a meta-analysis is rarely enough to support a mixture decomposition that produces more than three components. We refer readers to Appendix A for a detailed discussion of the technical results related to this current limitation, and the promising recent developments that could potentially address the limitation in the future. To summarize the existing results, we note that, when no separation assumption is made for the mixture components, the sample size required by mixture decomposition increases *exponentially* with the number of mixture components (Hardt & Price, 2015; Moitra & Valiant, 2010). This exponential growth challenges the feasibility of extracting a large number of components from a real-world dataset with only a limited number of samples. Reflecting this concern, most existing work on mixture decomposition (without the separation assumption) focused *exclusively* on the case of *two* mixture components (e.g., Daskalakis & Kamath, 2014; Hardt & Price, 2015; Kalai et al., 2010); and *all* simulation or experimental results we could find in the literature (e.g., Bandi et al., 2019; Li & Schmidt, 2017) tested only two or three mixture components when the sample size is below 1,000. Because a meta-analysis rarely covers more than a few hundred primary studies, it appears unlikely for a mixture of more than three components to emerge from mixture decomposition. Thus, while we strive to keep the design of the mixture decomposition algorithm generic to the number of components, we focus on utilizing two or three mixture components for moderator analysis in this paper, and leave a thorough study of the generic case to future work.

When the mixture decomposition algorithm returns three components, we can, somewhat surprisingly, adopt the same simple solution as the two-component case. The reason why the moderating effect can still be identified from $\Pr\{es_i \in C_2 \mid r(es_i)\}$, that is, the posterior probability for es_i to belong to what is now the "*middle*" component (i.e., with mean between C_1 and C_3), can be explained by contradiction. Consider the case where the hypothesized moderator variable V has two levels, LOW and HIGH. Suppose the idea of choosing the middle component fails. In other words, the hypothesized moderator V had a significant moderating effect, yet $\Pr\{es_i \in C_2 \mid r(es_i)\}$ stayed roughly constant between $V =$ LOW and HIGH. In this case, a change of V would have to "move" an effect size between C_1 and C_3, simply because the sum of the posterior probability for all three components has to be 1. Note that C_1 and C_3 represent the more "extreme" effect sizes at different sides of C_2. As such, if primary studies with $V =$ LOW and HIGH differ significantly on their affiliations with C_1 and C_3, then we will likely observe a significant difference between the mean effect sizes of the two moderator levels. That is, the traditional moderator analysis methods likely work well anyway. In contrast, when the effect of V is to "move" an effect size between C_2 and the other components, as illustrated by the example in the

[9] In this case, the NULL hypothesis with our method is that es_i follows the same distribution regardless of v_i.



a) Mixture distribution

b) Moderator analysis using raw effect sizes

c) Moderator analysis using component affiliations

Figure 3. An illustration of how our latent mixture-based moderator analysis works. The left figure depicts the case where the effect-size distribution is Gaussian with mean 0 and standard deviation 1 when the moderator is LOW, and mean 0.1 and standard deviation 0.1 when the moderator is HIGH. The solid line represents the observed mixture distribution. The middle figure shows the box plot for the mean effect sizes given 40 primary studies at each moderator level. The right figure shows the box plots for $\Pr\{es_i \in C_2 \mid r(es_i)\}$ used in our method that clearly separates the two moderator levels. See the online article for the color version of this figure.

introduction, then the moderating effect might be missed by the traditional analysis (but properly identified by our latent mixture-based method) because a move from C_2 to the two components on both sides of it may not change the mean effect size significantly.

In summary, once the mixture decomposition algorithm returns two or three components in Out_{M}, our latent mixture-based method is a simple two-step process: First, based on Out_{M}, our method computes for each primary study its posterior probability of belonging to the second mixture component (i.e., $\Pr\{es_i \in C_2 \mid r(es_i)\}$ for $i = 1, \ldots, m$). Second, we replace es_i with $\Pr\{es_i \in C_2 \mid r(es_i)\}$ before calling any traditional method for moderator estimation. The detailed design of the mixture decomposition algorithm is discussed in Appendix A, with parameter setup discussed in Appendix B, the ideas for reducing the computational overhead in Appendix C, and the handling of the three-component case in Appendix D.

Simulation Studies Using the Latent Mixture-Based Method

Overview

We conducted two simulation studies to determine how well our latent mixture-based method performs compared with the traditional methods. The simulation studies focused on two outcomes. One was the comparison of statistical power (i.e., 1 − Type II error rate) attained by different methods given the same Type I error rate. The other was the comparison of their statistical power under various known meta-analytic conditions. This allowed for the analysis of how the underlying effect-size distribution impacts performance of the methods.

The main difference between the two simulation studies was the number of mixture components in the effect-size distribution. The first study focused on the two-component case akin to the example in Figure 3, while the second study focused on the three-component case akin to the example in Figure 2. The purpose for designing these two studies was to demonstrate two different use cases of the latent mixture-based method. In Study 1, the hypothesized moderator variable was the only factor affecting the effect-size distribution. Thus, the superiority of our method stems from the component affiliation of an effect size being a better "signal" for moderator estimation than the effect size itself. In Study 2, there was a latent factor besides the moderator variable that affects the effect-size distribution. As such, the superiority of our method now stems from its ability to disentangle the latent factor from the true moderating effect.

We implemented all methods using Node.js and R; used a variety of statistical packages (through an R-to-Node.js wrapper), including the jStat library for JavaScript, R, and the metafor library for R (Viechtbauer, 2010); and produced all figures using MATLAB.[10] In the rest of this section, we first describe the simulation design in each study, before discussing the specifics of the moderator estimation methods tested in both studies.

Simulation Study 1: Two Components

For the first study, we created two levels for the moderating variable: LOW and HIGH, assigned a Gaussian effect-size distribution to each level, and systematically varied the parameters of the two distributions in order to examine the influence of varying a moderating effect on the outcome of the moderator analysis. Specifically, when the moderator is LOW, we simulated the effect sizes with a Gaussian distribution of mean 0 and standard devia-

[10] The download link and a brief description of an R package for our latent mixture-based method is available in the online supplemental materials.

tion 1. When the moderator is HIGH, we created three levels[11] for the mean of the effect-size distribution: 0.1, 0.3, and 0.5; and four levels for its standard deviation: 0.1, 0.3, 0.5, and 1. To simulate different proportions of primary studies with moderator being LOW or HIGH, we created three levels of weight composition: 25% (i.e., 25% HIGH and 75% LOW), 50%, and 75%. To examine the influence of the number of primary studies, we created three different levels: 40, 70, and 100.

Overall, the simulation design consisted of 108 unique conditions or a 3 (mean of HIGH) × 4 (standard deviation of HIGH) × 3 (weight) × 3 (number of primary studies) factorial design. Besides these simulation conditions, we also examined separately the null-effect case where the effect-size distribution for HIGH is exactly the same as LOW, that is, with mean 0 and standard deviation 1. For the null-effect case, the only meaningful factor is the number of primary studies, because the mean and standard deviation for HIGH are both fixed, and the weight composition has no influence on the observed effect-size distribution. For each simulation condition, we first calculate the number of primary studies with each moderator level according to the sample size and the weight specified in the simulation condition, and then generate the effect sizes for each moderator level independently at random, according to the corresponding Gaussian distribution specified in the simulation condition.

Simulation Study 2: Three Components

For the second study, we again set two levels for the moderating variable (i.e., LOW and HIGH) but mapped their effect-size distributions to three instead of two Gaussian components. While LOW always featured a Gaussian effect-size distribution, the distribution for HIGH was a mixture of two Gaussian components. For the sake of clarity, we fixed the mean of the two Gaussian components for HIGH at −1 and 1, respectively, and referred to them as the *left* and *right* components. Meanwhile, we created three levels for the mean of the LOW component: 0.1, 0.2, and 0.3, and referred to it as the *middle* component. We created four levels for the standard deviation of each component: 0.1, 0.3, 0.5, and 1; and three levels for the number of primary studies: 40, 70, and 100. The number of primary studies was always evenly distributed between LOW and HIGH and, within HIGH, evenly distributed (or differing by 1 if needed) between its two components. Overall, this second study consisted of 36 unique conditions or a 3 (mean of LOW) × 4 (standard deviation) × 3 (number of primary studies) factorial design.

Moderator Estimation Methods Tested

In each simulation study, we compared our latent mixture-based method against two widely used methods for moderator estimation: metaregression (Thompson & Higgins, 2002) and Q-statistic (Borenstein, Hedges, Higgins, & Rothstein, 2011, pp. 107–125). For all three methods, we tested a model with the moderator variable (i.e., HIGH/LOW) being the only study-level characteristic variable that potentially predicts the reported effect sizes. For example, with metaregression, the vector of reported effect sizes was regressed on the moderator variable (and an intercept term). Note that, while numerous variations for metaregression have been developed in the literature (e.g., Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015), the differences among many variations were moot in our simulations because we assumed all primary studies to have the same sample size. For example, all primary studies bear the same weight regardless of whether the meta-analytic model assumes fixed or random effects. Consequently, the results of ordinary least square and weighted least square became equivalent. Several alternatives for significance testing in metaregression, like the Knapp and Hartung (2003) test, also became equivalent with the standard Wald-type test. We tested other alternative designs of metaregression, such as the permutation method for significance testing (Higgins & Thompson, 2004; Viechtbauer et al., 2015), but did not find significant differences in terms of statistical power and Type I error rates. For Q-statistic, because the between-study variance is, by design, different for the two moderator levels in most cases, we used separate (rather than pooled) estimates of effect-size variance when computing the Q-statistic (Borenstein et al., 2011, p. 167). For each of the three methods, we calculated its statistical power from the simulation results according to a significance level of $\alpha = .05$, and its Type I error rate based on the outcomes of the null-effect cases discussed in the simulation design, again when setting $\alpha = .05$.

Simulation Study Results

Comparisons of Statistical Power

Tables 1 and 2 compare the statistical power rates attained by the three methods in the two simulation studies, respectively. Table 1 also includes the Type I error rates of all three methods. Note that, while Table 2 directly shows the statistical power of

[11] As these two distributions (i.e., when moderator being LOW and HIGH) are critical factors for both mixture decomposition and moderator analysis, it is necessary that we briefly explain our decision to only vary the distribution of HIGH while keeping the LOW distribution constant. A key rationale here is that all three moderator-analysis methods being examined are linear-invariant, meaning that their outcomes do not change when their input effect sizes undergo linear transformations. Similarly, their outputs also stay the same when we swap the labels of LOW and HIGH for the moderator variable. As such, for any given pair of Gaussian distributions, without affecting the outcome of moderator analysis, we can always consider the distribution with the larger standard deviation as the LOW distribution, and then normalize it through linear transformation to make the mean of the distribution 0 and the standard deviation 1, as assumed in our simulations. Specifically, the linear transformation is to transform a value x to $f(x) = (x - m)/s$, where m and s are the (original) mean and standard deviation of the distribution, respectively. Note that, once we apply the same linear transformation f to the other (i.e., HIGH) distribution, the HIGH distribution will always have a standard deviation between 0 and 1, justifying our decision to create four levels within this range for the standard deviation of HIGH. If the mean of the HIGH distribution is negative after applying f, we can always multiple f by −1 (i.e., applying another linear transformation) to make its mean positive without changing the LOW distribution. Because of this, we only need to consider nonnegative values for the mean of HIGH. The reason why we only simulated values under 0.5 for the mean of HIGH was simply because all three methods being examined achieved near-perfect accuracy once the mean of HIGH rose above 0.5. Similarly, we did not test the case where the mean of HIGH is 0 because metaregression and Q-statistic were not designed to identify the moderating effect when the mean effect size remains exactly the same under different levels of the moderating variable (Hedges & Pigott, 2004). Thus, we decided to focus on the simulation conditions where the mean of HIGH is between 0.1 and 0.5.

Table 1
Statistical Power and Type I Error of Latent Mixture-Based Moderator Analysis Versus Metaregression and Q-Statistic in Simulation Study 1

	Latent mixture-based moderator analysis				Metaregression				Q-statistic			
Simulation condition	Power Mean	Power Median	Power Stddev	Type I Error	Power Mean	Power Median	Power Stddev	Type I Error	Power Mean	Power Median	Power Stddev	Type I Error
Num of studies												
40	0.57	0.52	0.35	0.05	0.25	0.21	0.19	0.08	0.12	0.09	0.12	0.07
70	0.69	0.81	0.33	0.01	0.36	0.34	0.26	0.03	0.22	0.16	0.20	0.03
100	0.76	0.90	0.30	0.05	0.44	0.40	0.31	0.07	0.30	0.22	0.28	0.05
Mean HIGH												
0.10	0.60	0.73	0.39		0.11	0.08	0.09		0.03	0.02	0.03	
0.30	0.66	0.74	0.34		0.31	0.30	0.16		0.15	0.14	0.09	
0.50	0.75	0.88	0.26		0.63	0.62	0.22		0.46	0.46	0.19	
Stddev HIGH												
0.10	1.00	1.00	0.01		0.41	0.32	0.30		0.20	0.09	0.23	
0.30	0.89	0.94	0.15		0.38	0.32	0.29		0.20	0.10	0.22	
0.50	0.54	0.55	0.20		0.35	0.23	0.27		0.21	0.13	0.23	
1.00	0.25	0.22	0.19		0.26	0.23	0.20		0.26	0.23	0.20	
Weight HIGH												
0.25	0.60	0.56	0.34		0.23	0.15	0.25		0.18	0.08	0.21	
0.50	0.71	0.86	0.33		0.39	0.35	0.28		0.26	0.17	0.26	
0.75	0.70	0.86	0.33		0.43	0.39	0.25		0.21	0.17	0.18	
Overall	0.67	0.78	0.33	0.04	0.35	0.29	0.27	0.06	0.22	0.14	0.22	0.05

each method, Table 1 displays the marginal statistics (mean, median, and standard deviation) of their statistical power due to the large number of simulation conditions in Study 1. Observe from the tables that, while the Type I error rates of all three methods are very close, the statistical power of our latent mixture-based method is significantly higher than the other two under almost all conditions.

Remarkably, metaregression achieved a statistical power of 0.8 or above in only one tenth of the conditions in Study 1 (11 out of 108, 10.19%), and no condition in Study 2 (out of 36). Q-Statistic never achieved this standard for any condition in either study. In comparison, our latent mixture-based method achieved a power of 0.8 in about half of the conditions in both studies (53 out of 108, 49.07% in Study 1; 17 out of 36, 47.22% in Study 2). Even when we relax the threshold on power from 0.8 to 0.5, metaregression only achieved so for fewer than one third of all conditions (31 out of 108, 28.70%) in Study 1, and only two conditions in Study 2 (out of 36, 5.56%). Q-Statistic did so for 17 conditions (out of 108, 15.74%) in Study 1, and again no condition in Study 2. In comparison, our latent mixture-based method achieved a power of 0.5 or above in 67.59% of the conditions in Study 1 (73 out of 108) and 75.00% in Study 2 (27 out of 36). Furthermore, the latent mixture-based method increased the statistical power by at least 100% (over both metaregression and Q-statistic) in 44 conditions (40.74%) in Study 1 and 28 conditions (77.78%) in Study 2. For 22 conditions (20.37%) in Study 1 and 21 conditions (58.33%) in Study 2, the latent mixture-based method achieved a power increase of at least 500% over both metaregression and Q-statistic.

Table 2
Statistical Power of Latent Mixture-Based Moderator Analysis Versus Metaregression and Q-Statistic in Simulation Study 2

		$mean_{LOW} = .1$			$mean_{LOW} = .2$			$mean_{LOW} = .3$		
Stddev	Num of studies	Latent mixture	Metaregression	Q-statistic	Latent mixture	Metaregression	Q-statistic	Latent mixture	Metaregression	Q-statistic
0.1	40	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
0.1	70	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.30	0.00
0.1	100	0.97	0.00	0.00	1.00	0.00	0.00	1.00	0.73	0.00
0.3	40	0.96	0.00	0.00	0.82	0.00	0.00	0.82	0.04	0.00
0.3	70	0.93	0.00	0.00	0.88	0.05	0.00	0.79	0.19	0.02
0.3	100	0.84	0.00	0.00	0.82	0.03	0.00	0.90	0.50	0.05
0.5	40	0.50	0.00	0.00	0.61	0.00	0.00	0.63	0.06	0.02
0.5	70	0.75	0.02	0.00	0.67	0.06	0.01	0.78	0.26	0.14
0.5	100	0.71	0.02	0.01	0.78	0.12	0.03	0.72	0.35	0.23
1.0	40	0.15	0.02	0.07	0.21	0.05	0.09	0.14	0.10	0.19
1.0	70	0.25	0.05	0.07	0.18	0.09	0.12	0.29	0.20	0.27
1.0	100	0.17	0.04	0.11	0.28	0.09	0.19	0.29	0.16	0.41

Note. Stddev is the standard deviation of effect sizes in each mixture component. Num of Studies is the total number of primary studies over both moderator levels (i.e., three mixture components). $mean_{LOW}$ is the mean effect size when the moderator level is LOW.

Impacts of Simulation Condition

We further investigated the impact of the simulation condition, specifically the effect sizes of the simulation factors, on the statistical power of moderator analysis. We focused on the power comparison between the latent mixture-based method and metaregression, given the considerably lower power of Q-statistic in the simulation (as discussed above, the power of Q-statistic remained below 0.8 in all simulation conditions). To this end, we conducted a five-way analysis of variance (ANOVA) over the results of Study 1, with the response variable being the statistical power, and the five factors being the method of moderator analysis (i.e., metaregression or the latent mixture-based method) and the four simulation factors, (i.e., the number of primary studies N and the mean, standard deviation, and weight of the HIGH component). Because each observation (i.e., each simulated dataset) provided only a binary response (i.e., a significant moderating effect was either detected or not, leading to a power of either 1 or 0), we conducted 300 replications per cell, resulting in 2 (method) $\times$ 3 (mean) $\times$ 4 (standard deviation) $\times$ 3 (weight) $\times$ 3 (N) = 216 different unique conditions and 64,800 different data sets analyzed in the ANOVA. This analysis was conducted in MATLAB Version R2020a.

Table 3 shows the results of the ANOVA. Given the large number of observations (64,800), 30 out of 31 effects being tested returned a significant p value (i.e., $p < .05$; the only exception, Method $\times$ N, returned $p = .0527$). Thus, we followed Steinley (2006) to focus only on the effects with a large effect size η^2, specifically those with $\eta^2 \geq 0.005$. In terms of main effects, the most significant ones, in the descending order of mean squares (i.e., also F), are *method*, *standard deviation*, *mean*, N, and *weight*. The statistical power was higher when (a) the latent mixture-based method was used, (b) the standard deviation of HIGH decreased, (c) the mean of HIGH increased, (d) the number of primary studies N increased, or (e) the weight of HIGH increased. The significant main effect of *method* confirmed the superiority of our latent mixture-based method over metaregression ($F(1, 64{,}584) = 12{,}338.37$, $p < .0001$). The reason behind the main effects of N and the *mean* and *standard deviation* (of the HIGH distribution) are also straightforward: The larger the number of primary studies was, or the larger the difference was between the means or standard deviations of the two components (LOW and HIGH), the easier it was for either moderator-analysis method to identify the moderating effect.

The last main effect of *weight*, however, is counterintuitive because it appears to suggest that increasing the number of primary studies with moderator being HIGH can somehow improve the accuracy of moderator analysis. This is obviously false because, to consider an extreme-case scenario, when the weight of HIGH became 99%, no moderator analysis would be able to accurately identify the moderating effect unless the number of primary studies were extremely large. Upon further investigation, the effect of weight was qualified by the significant two-way interaction between standard deviation and weight, $F(6, 64{,}584) = 133.86$, $p < .0001$. Figure 4a depicts this interaction effect. As can be seen in the figure, for either method, when the standard deviations of HIGH and LOW were the same (i.e., 1), the statistical power of moderator analysis was roughly the same when the weight was 0.25 or 0.75. This is because the two cases were interchangeable under a simple linear transformation,[12] and were therefore indistinguishable for any linear-invariant method like both methods being tested. In contrast, when the standard deviation of HIGH was smaller (e.g., 0.3), either method reached a higher statistical power when the weight of HIGH was larger. The reason here is that the standard deviation of HIGH was, by setup, always smaller than LOW. Thus, when the weight of HIGH increased, the overall standard deviation of the observed effect sizes was bound to decrease, naturally improving the power of moderator analysis.

Besides this interaction between *standard deviation* and *weight*, ANOVA also identified three other interactions: two 2-way interactions between *method* and *standard deviation*, $F(3, 64{,}584) = 2327.45$, $p < .0001$ and between *method* and *mean*, $F(2, 64{,}584) = 1424.54$, $p < .0001$, which were qualified by the three-way interaction of *method*, *mean*, and *standard deviation*, $F(6, 64{,}584) = 164.84$, $p < .0001$. As can be seen from Figure 4b, this three-way interaction can be interpreted by the *method-mean* interaction having different patterns across different levels of the *standard deviation*. In general, the power improvement of the latent mixture-based method over metaregression was more pronounced when the HIGH distribution had a smaller mean (see Figure 4c). Such improvement became even more pronounced when standard-deviation of HIGH was smaller (see Figure 4d). We further elaborate on these observations in the following discussions.

Observed Trends From Simulation Results

We draw two important conclusions from the simulation results: (a) the power improvement offered by the latent mixture-based method was particularly pronounced when the difference of mean effect sizes between moderator levels was small, or when the difference of their standard deviations was large; and (b) a limitation of the latent mixture-based method was that it offered little improvement over the traditional methods when the standard deviation of *every* mixture component is very large.

Improvement over existing methods. Concerning the first point, one can observe from Table 1 that, in Study 1, the latent mixture-based method offered particularly significant power improvements when the mean difference between LOW and HIGH was smaller or when their standard-deviation difference was larger. The same can be observed from Table 2 for Study 2, where HIGH has a significantly larger standard deviation due to the presence of the latent factor. A key reason why metaregression (and Q-statistic) did *not* perform well in these cases was actually the null hypotheses it was designed to test: *The group mean effect sizes are equal across different levels of the moderator variable* (Hedges & Pigott, 2004). When the moderator variable has two levels, this means that, to reject the null hypothesis, the confidence intervals for the mean effect sizes at both levels need to be narrow. Therefore, as long as one moderator level had reported effect sizes with a large standard deviation, its wide confidence interval would likely overlap with the other (no matter how narrow a confidence interval the other level had), making it difficult to reject the null hypothesis. More formally, note that with metaregression, testing the significance of a moderating variable is

[12] This specific linear transformation is $f(x) = -(x - mean_{\text{HIGH}})$, where $mean_{\text{HIGH}}$ is the mean of the HIGH distribution. One can see that after applying this linear transformation to the LOW and HIGH distributions in the 0.25 case, the output distributions become exactly the same as the HIGH and LOW distributions in the 0.75 case, respectively.

Table 3
ANOVA With Statistical Power as the Response

Source	*df*	*SS*	*MS*	*F*	η^2
Method	1	1645.79	1645.79	12338.37	0.10
SD	3	2042.13	680.71	5103.25	0.13
Mean	2	1237.68	618.84	4639.42	0.08
N	2	370.69	185.35	1389.54	0.02
Weight	2	301.59	150.80	1130.52	0.02
Method × *SD*	3	931.36	310.45	2327.45	0.06
Method × Mean	2	380.03	190.02	1424.54	0.02
SD × Weight	6	107.13	17.86	133.86	0.01
Method × Mean × *SD*	6	131.93	21.99	164.84	0.01
Total	64,799	16194.24			

Note. *df* = degree of freedom; SS = sum of squares; MS = mean squares; *SD* = standard deviation (of the HIGH component); *N* = number of primary studies. $p < .001$ for all rows in the table. Only main effects and interactions with effect size $\eta^2 \geq .005$ were included in the table. All *F*s had 64,584 denominator degrees of freedom.

essentially testing whether its corresponding regression coefficient is zero (Thompson & Higgins, 2002). Given two moderator levels with standard deviations σ_1, σ_2 and weights w_1, w_2, respectively, the standard deviation for the regression coefficient is of the form[13]

$$s_b = \sqrt{\frac{\sigma_1^2 w_1 m + \sigma_2^2 w_2 m + s^2 m}{(m-2)(w_1 w_2^2 m + w_2 w_1^2 m)}}, \quad (3)$$

where m is the total number of primary studies and s is the standard error for each primary study. This formula can be further simplified to $s_b = \sqrt{4(\sigma_1^2 + \sigma_2^2 + 2s^2)/(m-2)}$ when $w_1 = w_2 = 0.5$. Consistent with earlier discussions, so long as one moderator level had a large standard deviation, this large value would dominate the value of s_b, thereby significantly reducing the power of metaregression no matter how small the standard deviation of the other level was. Similarly, for Q-statistic (Hedges & Pigott, 2004), when $w_1 = w_2 = 0.5$, the expected value of the test statistic (over the randomness of all reported effect sizes) is

$$E(Q_B) = \frac{m \cdot (\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2 + 2s^2}, \quad (4)$$

where μ_1 and μ_2 are the mean effect sizes for the two moderator levels, respectively. Under the null hypothesis, Q_B follows the chi-square distribution with degree of freedom 1. Thus, when $\mu_1 \neq \mu_2$, the larger Q_B is, the higher the statistical power of the Q-statistic method will be. Again, so long as one moderator level had a large standard deviation, this large value would dominate the denominator of $E(Q_B)$, thereby significantly reducing the power of Q-statistic *no matter* how small the standard deviation of the other level was.

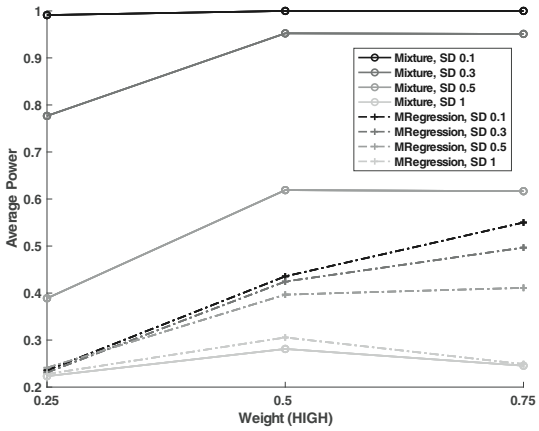
In contrast, our latent mixture-based method turned the uneven standard deviations across moderator levels, whether caused by the nature of the moderating effect (like in Study 1) or by the presence of a latent factor (like in Study 2), from a deficiency into an *asset* by leveraging the unevenness to better disentangle the mixture components. For example, consider a simulation condition in Study 1 where the mean, standard deviation, and weight of HIGH were 0.5, 0.1, and 0.5, respectively. We collected from the simulation results the component affinity of each effect size. For those with a moderator level of LOW, the mean was 0.39 and the standard deviation was 0.22. For those with HIGH, the mean was 0.72 and the standard deviation was 0.35. Simple calculation[14] would suggest that, to confirm a statistically significant difference on the mean, we only needed on average 12 primary studies for each moderator level. Consistent with this calculation, the simulation results showed that our latent mixture-based method achieved a statistical power of 1.00 when the total number of primary studies was only 40—significantly higher than the powers of 0.57 and 0.26 for metaregression and Q-statistic, respectively.

Limitation of the latent mixture-based method. Concerning the second point, one important limitation of the latent mixture-based method is that it offers little improvement when the effect-size distributions for all moderator levels have large standard deviations. For both studies, this limitation can be observed from the low statistical power of our approach when the standard deviation is 1.0 (in Tables 1 and 2).

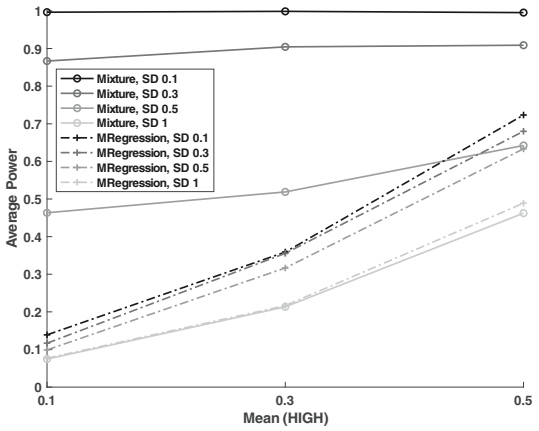
It is important to note that this lack of improvement was *not* solely caused by the increasing error in the mixture decomposition process. To verify this, we made a slight change to Study 1 by feeding the latent mixture-based method with the *exact* distributions of LOW and HIGH, and found the power and sensitivity of its output to stay essentially the same. Upon further investigation, we found that the root reason behind this lack of improvement appeared to be how we leveraged the decomposed mixture in moderator analysis. More specifically, recall that we replaced the regressand in metaregression with the (posterior) probability for a reported effect size to belong to the second mixture component. When all moderator levels had large standard deviations, this probability became approximately *monotonic*, with a close-to-linear relationship, with the value of the effect size. More specifically, in Study 1, when the mean, standard deviation, and weight of HIGH was 0.1, 1, and 0.5, respectively, assuming that the mixture decomposition algorithm can recover the exact distributions of LOW and HIGH, the posterior probability for an effect size x to belong to the HIGH component is

[13] LOW and HIGH were coded as 0 and 1, respectively, in the metaregression.

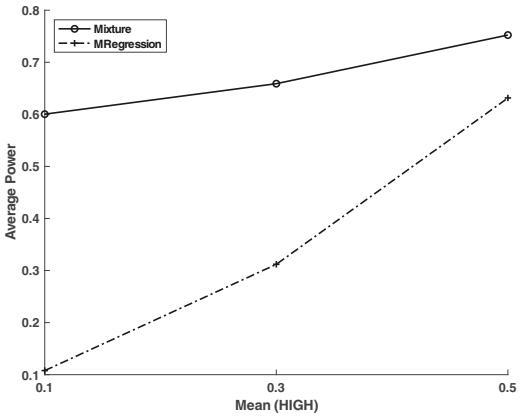
[14] Specifically, when each moderator level has a sample size of 12, the *t*-statistic for sample mean difference is $(0.72 - 0.39)/(0.22/\sqrt{12} + 0.35/\sqrt{12}) = 2.01 > 1.96$, which is the *t*-statistic corresponding to a significance level of $p = .05$ in a two-tailed *t*-test.



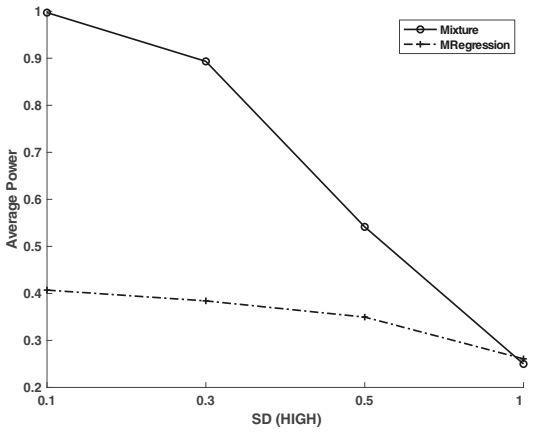
a) *standard deviation* and *weight*



b) *method*, *mean* and *standard deviation*



c) *method* and *mean*



d) *method* and *standard deviation*

Figure 4. Interaction effects. Mixture = our latent-mixture method; MRegression = the metaregression method; *SD* = standard deviation.

$$\begin{aligned}\Pr\{x \in \text{HIGH} \mid x\} &= \frac{e^{0.1x-0.005}}{1+e^{0.1x-0.005}} \\ &\approx \frac{0.995+0.1x}{1.995+0.1x} \\ &\approx \frac{0.995}{1.995}+0.025x, \end{aligned} \quad (5)$$

where the two approximations hold[15] when $-2 < x < 2$ (covering 95.4% of all effect sizes). Given the aforementioned linear-invariant property of the latent mixture-based method, this close-to-linear relationship between x and $\Pr\{x \in \text{HIGH} \mid x\}$ reduced the latent mixture-based method to metaregression, explaining this observed limitation.

A Case Study Using Real-World Data

While the simulation studies highlighted two use cases of the latent mixture-based method, there is yet another, arguably more prevalent, use case: when the effect sizes associated with every moderator level is heterogeneous, that is, consisting of multiple mixture components. This scenario often happens when effect sizes are derived from observational data, with multiple moderator variables affecting the focal relationship at once. In this case, there are almost certainly moderator variables whose values are not perfectly aligned with the decomposed mixture components. Using a real data set as a case study, we show how the latent mixture-based moderator-analysis method can better detect moderating

[15] Note that the first approximation assumes $e^{0.1x\,-\,0.005} \approx 1 + (0.1x - 0.005)$, a commonly used linear approximation of the exponential when the exponent is close to 0. While there is no universal consensus on an upper limit on the exponent for the approximation to hold, because $e^{-0.2} = 0.82$ and $e^{0.2} = 1.22$, we consider this approximation to be valid when $-2 < x < 2$. The second approximation assumes $0.0025x^2 \ll 0.05x$, which obviously holds when $-2 < x < 2$.

effects when each level of a moderator variable is corresponding to a (different) mixture of the decomposed components.

The Real-World Data Set

As a case study, we considered a problem extensively studied in psychology: whether trust matters more for the performance of virtual teams than face-to-face teams (i.e., whether team virtuality has a moderating effect on the relationship between intrateam trust and team performance). Interestingly, the two recent meta-analyses, published on the same issue of the *Journal of Applied Psychology*, drew different conclusions: While Breuer, Hüffmeier, and Hertel (2016) found it significant (i.e., the trust-performance relationship was stronger in virtual teams than face-to-face teams), De Jong, Dirks, and Gillespie (2016) found that the strength of the trust-performance relationship does *not* meaningfully differ between virtual teams and face-to-face teams. These different conclusions continued the long-standing inconsistency in the literature on whether the moderating role of team virtuality indeed exists (Alge, Wiethoff, & Klein, 2003; Muethel, Siebdrat, & Hoegl, 2012; Staples & Webster, 2008).

Throughout this case study, we used the same data and artifact-correction procedures as De Jong et al. (2016). Specifically, we first downloaded the data from its online supplemental materials, and then applied sampling-error correction and measurement-error correction according to the specifications within. Note that because De Jong et al. (2016) used artifact distributions, rather than individual artifacts, for measurement-error correction (according to the procedures in Hunter & Schmidt, 2004), this step actually had no effect on the outcome of our moderator analysis. We verified the correction of the procedure by successfully reproducing the statistics reported in De Jong et al. (2016), for example, the mean (corrected) effect-size difference between team virtuality being HIGH and LOW was 0.09, with 95% confidence interval being [−0.03, 0.20]. We then applied the latent mixture-based method on the data. To test the reliability of results, we conducted a variant of the "leave-one-out" analysis by randomly excluding from the input data a number of the primary studies. We tested three specific cases: excluding one study, 10% of studies, and 20%. For each case, we repeated the test 100 times and recorded the number of cases where latent mixture-based method identified a moderating effect with a significance level of $p = .05$.

Results and Discussion

Table 4 shows the results of the latent mixture-based method when applied over the entire input dataset, and compares them with the results in De Jong et al. (2016). As can be seen in the table, the latent mixture-based method did identify a significant moderating effect of team virtuality. Furthermore, the leave-one-out analysis confirmed that such an effect was not an artifact of one or a small number of primary studies. Specifically, when randomly excluding one, 10%, and 20% of the primary studies, the results still indicated a significant moderating effect in 98%, 79%, and 67% of the cases, respectively.

To understand *how* our latent mixture-based method was able to identify the moderating effect while the existing method cannot, we compared how traditional subgroup analysis and our method model the (subgroup) distributions for face-to-face teams (i.e., moderator = LOW) and virtual teams (i.e., HIGH). As can be seen in Figure 5b, our latent mixture-based model reveals that the two moderator levels differ quite significantly when the effect sizes are small, yet the difference diminishes when the effect sizes are large. Unfortunately, if one did not disentangle the mixture components but instead relied on the overall mean and standard deviation of the two subgroups, like in traditional subgroup analysis, this difference would no longer be recognizable, as shown in Figure 5a. To verify this explanation, we designed a simple test with the aim of "isolating" the effect of the right-most component. Specifically, we considered the following question: if we only considered primary studies that reported effect size below a certain threshold, say 0.2, we would have "zoomed in" to the part of the distribution where the moderator variable has a significant effect. In this case, would it be *more likely* for a moderator-analysis method to identify the moderating effect despite of the reduced sample size? As can be seen in Figure 5c, when the threshold increased, the *t*-score first rose until reaching the peak when the threshold was around 0.2, after which it declined, eventually falling out of the statistically significant range. This is remarkably consistent with our explanation that the increased resolution offered by our method is the reason why it detected the moderating effect.

In this case study, the latent mixture-based method not only identified the moderating effect, but also pinpointed *where* team virtuality likely had the strongest effect: when the effect size of the trust-performance relationship was relatively small. There are various explanations for why this could have happened. For example, it could have been caused by interactions between team virtuality and other team characteristics, as pointed out by De Jong et al. (2016) and De Guinea, Webster, and Staples (2012). It might also have been caused by a "ceiling effect" on how strong the trust-performance relationship could be. More research is needed for understanding the reason behind this observation. Finally, while we did not reanalyze the data in the other existing meta-analysis (Breuer et al., 2016), which had different inclusion-exclusion criteria and did identify a significant moderating effect for team virtuality, we would like to note that the mean effect size reported there for both face-to-face and virtual teams were *smaller than* those reported by De Jong et al. (2016): $\rho = 0.22$ versus 0.26 for face-to-face teams, 0.33 versus 0.35 for virtual teams. This is again consistent with the finding of the case study because, as discussed earlier, primary studies reporting smaller effect sizes likely demonstrate a stronger moderating effect for team virtuality.

General Discussion

In this section, we first discuss the research implications that can be drawn from our latent mixture-based method. We then review the limitations of our latent mixture-based method and the future research needed to address them.

Research Implications

For the focal meta-analytic task studied in this article (i.e., mixture estimation), the main research implication of our latent mixture-based method is its ability to offer a substantially higher statistical power than the existing methods when the underlying effect sizes are not identically distributed, but instead form a mixture of multiple different distributions, with the heterogeneity

Table 4
Results for Moderator Analysis

Team virtuality	Studies	ρ	SD_ρ	SE_ρ	95% CI $\rho_1 - \rho_2$		c	SD_c	SE_c	95% CI $c_1 - c_2$	
High	26	.35	.17	.047	−.03	.20	.96	.02	.004	.01	.05
Low	56	.26	.24	.039			.93	.07	.010		

Note. ρ = mean corrected effect size; SD_ρ = standard deviation of ρ; SE_ρ = standard error of ρ; 95% CI ρ_1–ρ_2 = confidence interval of between-group effect-size difference; c = mean component affinity; SD_c = standard deviation of c; SE_c = standard error of c; 95% CI c_1–c_2 = confidence interval of between-group component-affinity difference.

between distributions caused by factors that could be known or unknown. Our method achieves this superiority by explicitly modeling the distribution of effect sizes with a Gaussian mixture model. To address a unique challenge for latent mixture modeling in meta-analysis—the inability for the EM algorithm and its variants to handle mixture distributions with mostly overlapping components—we leveraged a recent breakthrough in theoretical machine learning that enabled the accurate decomposition of mixture components arbitrarily close to each other. The results of such a decomposition can then be used to more effectively test a hypothesized moderating effect. A unique feature of our method is that it can be considered a preprocessing step *orthogonal* to the numerous methods developed (and debated) in the literature for moderator estimation. As such, researchers conducting meta-analysis can freely decide which existing method to use after obtaining the decomposed mixture components.

Within the scope of meta-analysis, the research implications of our work extend beyond the moderator estimation task discussed in this article, as many other meta-analytic tasks could also benefit from a proper understanding of the mixture composition of the effect-size distribution. For example, when applied to moderator detection, the mixture model could help determine whether an unexplained heterogeneity is "natural" or should be attributed to one or more unknown moderating effects. As an illustration, consider the case where the effect-size distribution is a mixture of two equal-weight Gaussians versus a mixture dominated by one Gaussian plus a low-weight component with extreme values. Apparently, the former is more likely to be explained by unknown moderators, while the latter could have the unexplained heterogeneity attributed to a small number of outlier studies. How to draw such inferences from various types of mixture compositions could be studied in future research.

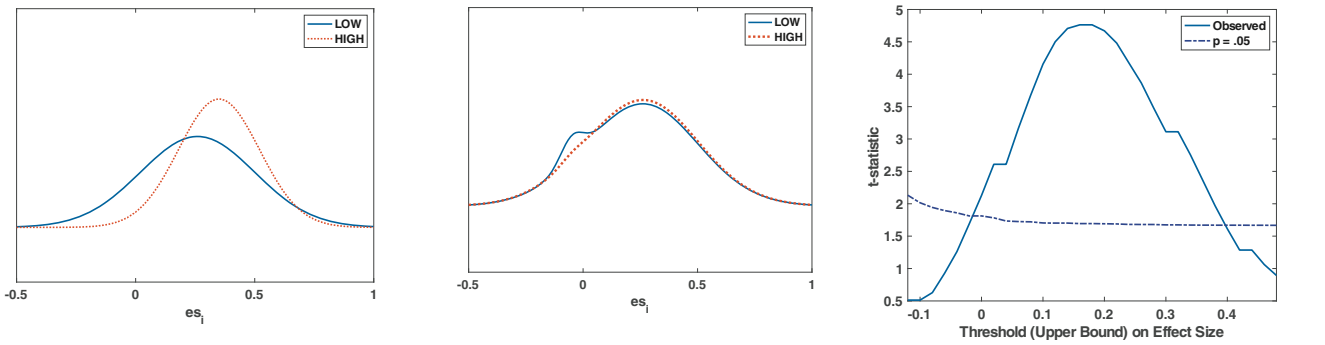
Mixture modeling could also help detect another common cause of unexplained heterogeneity: the need for finer gradations of variable coding (Hunter & Schmidt, 2004, p. 180). As discussed in the introduction, if one moderator level features a Gaussian distribution yet the other exhibits a two-component mixture, then a researcher should consider whether the second level could be further partitioned into finer gradations, and examine whether such finer coding is consistent with the two mixture components. Similarly, if each moderator level has only one Gaussian component, then finer coding is less likely to alter the outcome of moderator analysis.

Another well-known challenge in meta-analysis that could potentially benefit from mixture modeling is the detection of and correction for availability biases such as publication bias (Lin & Chu, 2018) and questionable research practices (QRPs). For example, QRPs such as data peeking (Simonsohn et al., 2014) are known to artificially increase effect size, essentially creating a mixture component with a larger mean. This difference in distribution has already been leveraged in existing research to detect QRPs, for example, through tools such as *p*-curve[16] (Simonsohn et al., 2014) and test of excess significance (Ioannidis & Trikalinos, 2007). Compared with these tools, mixture decomposition has the potential to not only detect the "questionable" component but also reveal the distribution of the "other" component that is not affected by the QRP.

Similarly, publication bias is known to lead to an excessive skewness in the effect-size distribution (Begg & Mazumdar, 1994), with the skewness being leveraged by existing research for detection (Ferguson & Brannick, 2012) and correction (Duval & Tweedie, 2000). Because a skewed distribution can be modeled as a mixture of two overlapping Gaussian distributions (Kalai et al., 2012), mixture decomposition also has the potential to detect and correct for such biases. How to generate robust estimates of the true effect-size distribution based on the output of mixture decomposition is an intriguing topic for future research.

Finally, the machine-learning technique leveraged by our method, which removes the mixture-class separation requirement of EM-based techniques, can also help improve latent mixture analysis with observed data (e.g., latent class analysis, latent profile analysis, growth mixture modeling). Poor mixture-class separation (i.e., a small distance between adjacent mixture components) has been shown to impact numerous aspects of latent mixture analysis, such as hindering the convergence of the computational process (Tofighi & Enders, 2008), requiring a substantially larger sample (Tueller & Lubke, 2010), making class enumeration (i.e., to determine the number of mixture components) more challenging (Depaoli, 2013), and so forth. Future research can study the use of the state-of-the-art mixture decomposition algorithms in latent mixture analysis with observed data. Such studies will likely need to address a number of unique challenges that were not sufficiently discussed in the computer science literature, such as how to enable class enumeration when a large class separation is no longer required by the mixture decomposition algorithm, how to determine if a non-Gaussian distribution should be decomposed into two largely overlapping Gaussian components

[16] Rigidly speaking, *p*-curve measures skewness of the *p*-value distribution, not the effect-size distribution. Yet the underlying mathematical principle is the same, as a (one-tailed) *p* value can be expressed as a monotonic function of the effect size.



a) Models based on a single Gaussian distribution

b) Models based on our mixture decomposition results

c) Changes of t-statistic with threshold on reported effect size

Figure 5. An illustration of our case study for examining the moderating effect of team virtuality in the relationship between intrateam trust and team performance. The left figure depicts the distributions being tested in traditional subgroup analysis. Because the outcome depends only on the mean ρ and standard deviation SD_ρ of the corrected effect sizes, the traditional method is essentially testing the mean difference between the two Gaussian distributions depicted in the figure (with mean ρ and standard deviation SD_ρ for each subgroup). The middle figure depicts what is tested in our method, that is, the component affiliation. Each line represents the estimated mixture distribution for the corresponding subgroup, generated by weighting the decomposed mixture components with the posterior distribution of component affiliation for effect sizes in the subgroup. The right figure depicts the change of the *t*-statistic between the mean effect sizes of studies with moderator being LOW and the other studies, as estimated by kernel density estimation when we only consider studies with reported effect sizes below a threshold, which varies from −0.12 to 0.50. The dotted line in the figure marks the *t*-score corresponding to $p = .05$. See the online article for the color version of this figure.

or be attributed to other factors such as poor measurement scaling (Bauer & Curran, 2003), and so forth.

Limitations and Future Directions

It is important to note several limitations to our study. First, the current method only tests the moderating effect of a single variable. While one could use the method in a hierarchical moderator analysis, as we will elaborate in this section, it may be advantageous to specifically design a latent mixture-based method for evaluating multiple moderator variables (and their interactions). Second, while we focused on the Gaussian mixture model, there may be real-world data sets that feature non-Gaussian components (e.g., Bauer & Curran, 2003), which call for future studies to investigate the potential use of non-Gaussian mixture models in meta-analysis. Finally, as discussed in the simulation results, our method does not perform well when all mixture components feature large standard deviations. We elaborate on these limitations and the corresponding future directions below.

Multiple moderator variables. When there are multiple moderator hypotheses to test, a well-known challenge to moderator estimation is there are often not enough primary studies to cover the many value combinations of moderator variables and support a fully hierarchical moderator analysis (Hunter & Schmidt, 2004, p. 424). Unfortunately, directly integrating our latent mixture-based method with hierarchical moderator analysis would worsen this problem. For example, if one first breaks out all primary studies into subgroups by the value of one categorical moderator variable before testing another moderator variable over each subgroup, then the smaller sample size in each subgroup could negatively affect not only the statistical power of traditional moderator analysis but also the accuracy of mixture decomposition (and therefore the statistical power of our method).

Interestingly, a potential solution to this challenge is to take inspiration from the various spectral algorithms recently developed for high-dimensional mixture decomposition (Anandkumar, Ge, Hsu, Kakade, & Telgarsky, 2014; Belkin & Sinha, 2015; Goyal, Vempala, & Xiao, 2014; Hsu & Kakade, 2013; Huang, Ge, Kakade, & Dahleh, 2015), and reduce the sample size required for mixture decomposition by taking into account both the effect sizes and the hypothesized moderator variables during the mixture decomposition process. For example, suppose there are two theoretically predicted moderator variables *A* and *B*. In analyzing the moderating effect of *B* (and understanding its interactions with *A*), we could adjust the target of mixture decomposition from a univariate distribution (i.e., the observed effect-size distribution) to a multivariate one representing the joint distribution of the observed effect size and *A*. As demonstrated by the recent spectral algorithms, if *A* is correlated with the effect size, then a two-dimensional mixture decomposition algorithm can leverage the correlation to reduce the sample size required for an accurate decomposition, simply because each sample now contains more information that can be used to guide the decomposition process. To this end, future research can study how to tailor the design of a high-dimensional mixture decomposition algorithm for the purpose of meta-analysis, how to study the moderating effect of *B* and its interactions with *A* based on the estimated mixture composition, and so forth.

Variations of mixture modeling. A limitation of our study is that it focused on *Gaussian* mixture modeling, which assumes each

component of the mixture to follow a Gaussian distribution. It is important to understand that there is *no* inherent constraint in mixture decomposition to assume each component to be Gaussian. Even though the vast majority of existing work on mixture decomposition made the Gaussian assumption (McLachlan & Peel, 2004), there were attempts to decompose mixture distributions into non-Gaussian components (e.g., Banfield & Raftery, 1993), or to distinguish between a mixture of multiple Gaussian and one non-Gaussian distribution (e.g., Nylund, Asparouhov, & Muthén, 2007). More generally, the field of unsupervised learning (e.g., clustering; Figueiredo & Jain, 2002) in machine learning could be considered as decomposing a (usually high-dimensional) mixture distribution into separate components without presumed distributions.

In terms of whether future research should drop the Gaussian assumption in studying mixture distributions in meta-analysis, there are two perspectives to consider. First, like in the famous case of Pearson's (1894) crab data, a skewed (thus non-Gaussian) distribution could be further decomposed into multiple Gaussian components. Indeed, as the literature of kernel density estimation (Silverman, 2018) suggests, *any* distribution can be expressed as a mixture of Gaussian distributions. From this perspective, it appears that the Gaussian assumption might not fundamentally limit the generalizability of mixture modeling. Nonetheless, there is also another perspective that the Gaussian assumption might unnecessarily increase the number of components we have to consider in a mixture model. For example, if we already know that each component likely follows a heavy-tailed distribution (e.g., Burton, 2012), then replacing the Gaussian assumption with a heavy-tailed distribution could substantially reduce the number of components in the mixture distribution, and thereby reducing the number of primary studies required for an accurate decomposition. To this end, future research could study what types of distributions other than Gaussian often emerge in a meta-analysis and revise the distributional assumptions in the mixture model accordingly.

Leveraging results from distribution testing in moderator analysis. Recall from the discussion of simulation results that a key limitation of our latent mixture-based method is its low power when all mixture components feature large standard deviations. In particular, we showed that the power stayed low even when the method had access to the exact mixture composition. This raised an intriguing feasibility question: Is it possible for moderator analysis to take advantage of an accurately decomposed mixture distribution, even when the mixture components have large standard deviations? To this end, we note a very active research area in theoretical computer science called *distribution testing* (see Canonne, 2017 for an excellent literature review), which may be especially useful for understanding whether a small number of samples are enough to distinguish between two distributions close to each other. For example, translating a landmark result in distribution testing (Chan, Diakonikolas, Valiant, & Valiant, 2014, Theorem 1.2) into moderator analysis, we know *no* method can possibly identify a moderating effect unless the number of primary studies exceeds a threshold that is inversely proportional to $d_{\mathrm{TV}}^{4/3}$, where d_{TV} is the total variation distance between the LOW and HIGH effect-size distributions. For example, with our Simulation Study 1, with equal number of primary studies in LOW and HIGH, there is

$$d_{\mathrm{TV}} = C \cdot \min\left(1, \max\left(\frac{\sigma_{\mathrm{LOW}}^2 - \sigma_{\mathrm{HIGH}}^2}{\sigma_{\mathrm{LOW}}^2}, \frac{\mu_{\mathrm{HIGH}} - \mu_{\mathrm{LOW}}}{\sigma_{\mathrm{LOW}}}\right)\right) \tag{6}$$

where μ_{LOW}, μ_{HIGH} represent the means and σ_{LOW}, σ_{HIGH} represent the standard deviations of the two components, and C is a constant proven to be between 1/200 and 9/2 (Devroye, Mehrabian, & Reddad, 2020). While this lower bound cannot tell us *exactly* how many primary studies are required,[17] it reveals the following insights that are consistent with our simulation results.

First, it shows that the large-standard-deviation case *should* have been much harder for moderator analysis than the other conditions. When $\sigma_{\mathrm{HIGH}} < 1$, d_{TV} was determined by the first input to max, with $d_{\mathrm{TV}} = 0.99C$, $0.81C$, and $0.75C$ when $\sigma_{\mathrm{HIGH}} = 0.1$, 0.3, and 0.5, respectively. When $\sigma_{\mathrm{HIGH}} = 1$, d_{TV} was determined by the second input instead, with $d_{\mathrm{TV}} = 0.1C$, $0.3C$, and $0.5C$ when $\mu_{\mathrm{HIGH}} = 0.1$, 0.3, and 0.5, respectively. Even the *largest* value when $\sigma_{\mathrm{HIGH}} = 1$ is smaller than the *smallest* value when $\sigma_{\mathrm{HIGH}} < 1$, confirming that the lower power of moderator analysis would likely hold even for the best possible meta-analysis method.

Second, it also appears to indicate that handling the case of $\mu_{\mathrm{HIGH}} = 0.1$ and $\sigma_{\mathrm{HIGH}} = 1$ requires more primary studies than what are typically available in a meta-analysis. To understand why, note that the value of d_{TV} in this case is 9.9 times lower than the case of $\mu_{\mathrm{HIGH}} = 0.1$ and $\sigma_{\mathrm{HIGH}} = 0.1$, meaning that the minimum number of primary studies it requires is roughly $9.9^{4/3} = 21.26$ times larger. Even when the "best" method took just *five* studies per moderator level to successfully identify the moderating effect in the latter case, the number of primary studies required in the former would be roughly $10 \times 9.9^{4/3} = 213$, larger than many existing meta-analyses in psychology and the social sciences.

Finally, it also points to potential methodological advances for moderator analysis. Chan et al. (2014) designed an algorithm that *achieved* this proven lower bound on sample size with an interesting two-step approach. Again translating the algorithm to the context of moderator analysis, the algorithm functions as follows. It starts by partitioning the possible values of the effect size into two subsets: One consists of "popular" values that are frequently reported for either LOW or HIGH. The algorithm uses a variant of the χ^2-test to determine whether LOW and HIGH differs significantly in terms of their observed frequencies on these popular values. Then, for the other subset of "unpopular" values, the algorithm calls upon another landmark result (Goldreich & Ron, 2011) in distribution testing to determine whether the two distributions differ on the subset. We do not elaborate on this second part because it is not important to our subsequent discussions. The reason why this algorithm is capable of minimizing the required number of primary studies follows directly from the above-derived value of d_{TV}. For example, under the simulation conditions, the "popular values" of effect size are naturally close to μ_{HIGH} and

[17] Almost all existing results in distribution testing uses the Big-O notation (Knuth, 1997, Section 1.2.11) in theoretical computer science—i.e., they focus on investigating the relationship between the required sample size (i.e., number of primary studies) and the characteristics of the underlying distributions (i.e., LOW and HIGH), and ignore constant factors in the derived results, making it impossible to derive or bound the exact number of primary studies that are required in a meta-analysis.

μ_{LOW}. Focusing on these values barely affects the numerator $\mu_{HIGH} - \mu_{LOW}$, yet can significantly reduce the denominator σ_{LOW} and thereby increase d_{TV}. When the resulting reduction on the lower bound offsets the reduction of sample size (because now only those primary studies reporting "popular values" are considered), the algorithm has the potential to more effectively detect the moderating effect when the total number of primary studies is relatively small. Because the focus of this article is to introduce mixture decomposition to meta-analysis, we leave further investigation of this design of moderator analysis to future research.

References

Achlioptas, D., & McSherry, F. (2005). On spectral learning of mixtures of distributions. In P. Auer & R. Meir (Eds.), *Learning theory. Lecture notes in computer science* (Vol. 3559, pp. 458–469). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/11503415_31

Alge, B. J., Wiethoff, C., & Klein, H. J. (2003). When does the medium matter? Knowledge-building experiences and opportunities in decision-making teams. *Organizational Behavior and Human Decision Processes, 91,* 26–37. http://dx.doi.org/10.1016/S0749-5978(02)00524-1

Améndola, C., Faugère, J. C., & Sturmfels, B. (2016). Moment varieties of Gaussian mixtures. *Journal of Algebraic Statistics*. Advance online publication. http://dx.doi.org/10.18409/jas.v7i1.42

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research, 15,* 2773–2832.

Anandkumar, A., Hsu, D., & Kakade, S. M. (2012). A method of moments for mixture models and hidden Markov models. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 23,* 33.1–33.34.

Anderson, J., Belkin, M., Goyal, N., Rademacher, L., & Voss, J. (2014). The more, the merrier: The blessing of dimensionality for learning large gaussian mixtures. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 35,* 1–30.

Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics, 45,* 77–120. http://dx.doi.org/10.1214/16-AOS1435

Bandi, H., Bertsimas, D., & Mazumder, R. (2019). Learning a Mixture of Gaussians via mixed integer optimization. *INFORMS Journal on Optimization, 1,* 185–264. http://dx.doi.org/10.1287/ijoo.2018.0009

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49,* 803–821. http://dx.doi.org/10.2307/2532201

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8,* 338–363. http://dx.doi.org/10.1037/1082-989X.8.3.338

Beath, K. J. (2014). A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Methods, 5,* 285–293. http://dx.doi.org/10.1002/jrsm.1114

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. http://dx.doi.org/10.2307/2533446

Belkin, M., & Sinha, K. (2010, June). *Toward learning Gaussian mixtures with arbitrary separation, annual conference on learning theory, Haifa, Israel*. Retrieved from https://www.learningtheory.org/colt2010/papers/082sinha.pdf

Belkin, M., & Sinha, K. (2015). Polynomial learning of distribution families. *SIAM Journal on Computing, 44,* 889–911. http://dx.doi.org/10.1137/13090818X

Berlin, J. A. (1995). Invited commentary: Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *American Journal of Epidemiology, 142,* 383–387. http://dx.doi.org/10.1093/oxfordjournals.aje.a117645

Böhning, D. (1999). *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others* (Vol. 81). Boca Raton, FL: CRC Press.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13,* 99–109. http://dx.doi.org/10.1037/1082-989X.13.2.99

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. New York, NY: Wiley.

Brannick, M. T., Yang, L. Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods, 14,* 587–607. http://dx.doi.org/10.1177/1094428110368725

Breuer, C., Hüffmeier, J., & Hertel, G. (2016). Does trust matter more in virtual teams? A meta-analysis of trust and team effectiveness considering virtuality and documentation as moderators. *Journal of Applied Psychology, 101,* 1151–1177. http://dx.doi.org/10.1037/apl0000113

Burton, C. (2012). Heavy tailed distributions of effect sizes in systematic reviews of complex interventions. *PLoS ONE, 7,* e34222. http://dx.doi.org/10.1371/journal.pone.0034222

Canonne, C. L. (2017). *Property testing and probability distributions: New techniques, new models, and new goals* (Doctoral dissertation). Columbia University, New York, NY.

Chan, S. O., Diakonikolas, I., Valiant, P., & Valiant, G. (2014). Optimal algorithms for testing closeness of discrete distributions. In C. Chekuri (Ed.), *Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1193–1203). Philadelphia, PA: Society for Industrial and Applied Mathematics. http://dx.doi.org/10.1137/1.9781611973402.88

Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences, 137,* 51–73. http://dx.doi.org/10.1016/S0025-5564(96)00075-2

Chernev, A., Böckenholt, U., & Goodman, J. (2010). Commentary on Scheibehenne, Greifeneder, and Todd choice overload: Is there anything to it? *The Journal of Consumer Research, 37,* 426–428. http://dx.doi.org/10.1086/655200

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19,* 211–229. http://dx.doi.org/10.1037/a0032968

Daniel, W. W. (1990). *Applied nonparametric statistics* (2nd ed.). Boston, MA: PWS-Kent.

Dasgupta, S. (1999). Learning mixtures of Gaussians. In P. Beame (Ed.), *Proceedings of the 40th Annual Symposium on Foundations of Computer Science* (pp. 634–644). Washington, DC: Institute of Electrical and Electronics Engineers Computer Society. Retrieved from https://dl.acm.org/doi/10.5555/795665.796496

Daskalakis, C., & Kamath, G. (2014). Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. *Journal of Machine Learning Research: Workshop and Conference Proceedings, 35,* 1183–1213.

De Guinea, A. O., Webster, J., & Staples, D. S. (2012). A meta-analysis of the consequences of virtualness on team functioning. *Information & Management, 49,* 301–308. http://dx.doi.org/10.1016/j.im.2012.08.003

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology, 101,* 1134–1150. http://dx.doi.org/10.1037/apl0000110

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B. Methodological, 39,* 1–22. http://dx.doi.org/10.1111/j.2517-6161.1977.tb01600.x

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18,* 186–219. http://dx.doi.org/10.1037/a0031609

Devroye, L., Mehrabian, A., & Reddad, T. (2020). The total variation distance between high-dimensional Gaussians. *arXiv*. Retrieved from https://arxiv.org/pdf/1810.08693.pdf

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

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17,* 120–128. http://dx.doi.org/10.1037/a0024445

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24,* 381–396. http://dx.doi.org/10.1109/34.990138

Fletcher, J. (2007). What is heterogeneity and is it important? *British Medical Journal, 334,* 94–96. http://dx.doi.org/10.1136/bmj.39057.406644.68

Foti, R. J., Bray, B. C., Thompson, N. J., & Allgood, S. F. (2012). Know thy self, know thy leader: Contributions of a pattern-oriented approach to examining leader perceptions. *The Leadership Quarterly, 23,* 702–717. http://dx.doi.org/10.1016/j.leaqua.2012.03.007

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5,* 351–379. http://dx.doi.org/10.3102/0091732X005001351

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York, NY: Russell Sage Foundation.

Goldreich, O., & Ron, D. (2011). On testing expansion in bounded-degree graphs. In O. Goldreich (Ed.), *Studies in complexity and cryptography. Miscellanea on the interplay between randomness and computation* (pp. 68–75). Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-22670-0_9

Goyal, N., Vempala, S., & Xiao, Y. (2014). Fourier PCA and robust tensor decomposition. In D. Shmoys (Ed.), *Proceedings of the 46th Annual ACM Symposium on Theory of Computing* (pp. 584–593). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/2591796.2591875

Hardt, M., & Price, E. (2015). Tight bounds for learning a mixture of two gaussians. In R. Rubinfeld (Ed.), *Proceedings of the 47th Annual ACM Symposium on Theory of Computing* (pp. 753–760). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/2746539.2746579

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

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445. http://dx.doi.org/10.1037/1082-989X.9.4.426

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Higgins, J. P., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23,* 1663–1682. http://dx.doi.org/10.1002/sim.1752

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A, 172,* 137–159. http://dx.doi.org/10.1111/j.1467-985X.2008.00552.x

Hopkins, S. B., & Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In M. Henzinger (Ed.), *Proceedings of the 50th Annual ACM Symposium on Theory of Computing* (pp. 1021–1034). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/3188745.3188748

Hosking, J. R. M. (1990). *L*-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society Series B. Methodological, 52,* 105–124. http://dx.doi.org/10.1111/j.2517-6161.1990.tb01775.x

Hosmer, D. W., Jr. (1973). On MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics Theory and Methods, 1,* 217–227.

Hsu, D., & Kakade, S. M. (2013). Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In R. Kleinberg (Ed.), *Proceedings of the 4th conference on Innovations in Theoretical Computer Science* (pp. 11–20). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/2422436.2422439

Huang, Q., Ge, R., Kakade, S., & Dahleh, M. (2015). Minimal realization problems for hidden Markov models. *IEEE Transactions on Signal Processing, 64,* 1896–1904. http://dx.doi.org/10.1109/TSP.2015.2510969

Huber, P. J. (2011). *Robust statistics*. Berlin, Germany: Springer Berlin Heidelberg.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11,* 193–206. http://dx.doi.org/10.1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Atlanta, GA: Sage.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. http://dx.doi.org/10.1177/1740774507079441

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: a meta-analytic test of their relative validity. *Journal of Applied Psychology, 89,* 755–768.

Kalai, A. T., Moitra, A., & Valiant, G. (2010). Efficiently learning mixtures of two Gaussians. In L. J. Schulman (Ed.), *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing* (pp. 553–562). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/1806689.1806765

Kalai, A. T., Moitra, A., & Valiant, G. (2012). Disentangling gaussians. *Communications of the ACM, 55,* 113–120. http://dx.doi.org/10.1145/2076450.2076474

Kim, T. H., & White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters, 1,* 56–73.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22,* 2693–2710. http://dx.doi.org/10.1002/sim.1482

Knuth, D. E. (1997). *The art of computer programming: Fundamental algorithms* (3rd ed.). Reading, MA: Addison Wesley Longman.

Kothari, P. K., Steinhardt, J., & Steurer, D. (2018). Robust moment estimation and improved clustering via sum of squares. In M. Henzinger (Ed.), *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (pp. 1035–1046). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/3188745.3188970

Levin, D. A., Peres, Y., Wilmer, E. L., Propp, J. G., & Wilson, D. B. (2017). *Markov chains and mixing times*. Providence, RI: American Mathematical Society. http://dx.doi.org/10.1090/mbk/107

Li, J., & Schmidt, L. (2017). Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. *Journal of Machine*

Learning Research: Workshop and Conference Proceedings, 65, 1302–1382.

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics, 74,* 785–794. http://dx.doi.org/10.1111/biom.12817

Lindsay, B. G. (1989). Moment matrices: Applications in mixtures. *Annals of Statistics, 17,* 722–740. http://dx.doi.org/10.1214/aos/1176347138

Lindsay, B. G., & Basak, P. (1993). Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association, 88,* 468–476. http://dx.doi.org/10.1080/01621459.1993.10476297

Liu, S., Huang, J. L., & Wang, M. (2014). Effectiveness of job search interventions: A meta-analytic review. *Psychological Bulletin, 140,* 1009–1041.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 84). New York, NY: M. Dekker.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: Wiley.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11,* 730–749. http://dx.doi.org/10.1177/1745691616662243

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. http://dx.doi.org/10.1037/0033-2909.105.1.156

Moitra, A., & Valiant, G. (2010, October). *Settling the polynomial learnability of mixtures of gaussians*. Annual IEEE Symposium on Foundations of Computer Science, Las Vegas, NV. http://dx.doi.org/10.1109/FOCS.2010.15

Muethel, M., Siebdrat, F., & Hoegl, M. (2012). When do we really need interpersonal trust in globally dispersed new product development teams? *R&D Management, 42,* 31–46.

Naaktgeboren, C. A., van Enst, W. A., Ochodo, E. A., de Groot, J. A., Hooft, L., Leeflang, M. M., . . . Reitsma, J. B. (2014). Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. *Journal of Clinical Epidemiology, 67,* 1200–1209. http://dx.doi.org/10.1016/j.jclinepi.2014.05.018

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: A reanalysis of 'power failure' in neuroscience using mixture modeling. *The Journal of Neuroscience, 37,* 8051–8061. http://dx.doi.org/10.1523/JNEUROSCI.3592-16.2017

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569. http://dx.doi.org/10.1080/10705510701575396

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies, 23,* 66–81. http://dx.doi.org/10.1177/1548051815614321

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A, 185,* 71–110. http://dx.doi.org/10.1098/rsta.1894.0003

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26,* 195–239. http://dx.doi.org/10.1137/1026034

Regev, O., & Vijayaraghavan, A. (2017, October). *On learning mixtures of well-separated gaussians*. Annual IEEE Symposium on Foundations of Computer Science, Berkeley, CA. http://dx.doi.org/10.1109/FOCS.2017.17

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Sanjeev, A., & Kannan, R. (2001, July). Learning mixtures of arbitrary gaussians. In J. S. Vitter, P. G. Spirakis, & M. Yannakakis (Eds.), *Proceedings of the 33rd Annual ACM SIGACT Symposium on Theory of Computing* (pp. 247–257). New York, NY: Association for Computing Machinery. http://dx.doi.org/10.1145/380752.380808

Schlattmann, P., Verba, M., Dewey, M., & Walther, M. (2015). Mixture models in diagnostic meta-analyses—Clustering summary receiver operating characteristic curves accounted for heterogeneity and correlation. *Journal of Clinical Epidemiology, 68,* 61–72. http://dx.doi.org/10.1016/j.jclinepi.2014.08.013

Schulze, R. (2004). *Meta-analysis—A comparison of approaches*. Cambridge, MA: Hogrefe Publishing.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66,* 605–610. http://dx.doi.org/10.1093/biomet/66.3.605

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge. http://dx.doi.org/10.1201/9781315140919

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9,* 666–681. http://dx.doi.org/10.1177/1745691614553988

Staples, D. S., & Webster, J. (2008). Exploring the effects of trust, task interdependence and virtualness on knowledge sharing in teams. *Information Systems Journal, 18,* 617–640.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96–111. http://dx.doi.org/10.1037/0021-9010.87.1.96

Steinley, D. (2006). Profiling local optima in K-means clustering: Developing a diagnostic technique. *Psychological Methods, 11,* 178–192. http://dx.doi.org/10.1037/1082-989X.11.2.178

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology, 79,* 354–359. http://dx.doi.org/10.1037/0021-9010.79.3.354

Teicher, H. (1961). Identifiability of mixtures. *Annals of Mathematical Statistics, 32,* 244–248. http://dx.doi.org/10.1214/aoms/1177705155

Thompson, S. G. (1994). Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309,* 1351–1355. http://dx.doi.org/10.1136/bmj.309.6965.1351

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21,* 1559–1573. http://dx.doi.org/10.1002/sim.1187

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age.

Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling, 17,* 165–192. http://dx.doi.org/10.1080/10705511003659318

Vempala, S., & Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences, 68,* 841–860. http://dx.doi.org/10.1016/j.jcss.2003.11.008

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

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20,* 360–374. http://dx.doi.org/10.1037/met0000023

Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician, 51,* 59–64.

Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods, 14,* 24–31. http://dx.doi.org/10.1177/1094428110383988

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817–838. http://dx.doi.org/10.2307/1912934

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315–321. http://dx.doi.org/10.1037/0021-9010.75.3.315

Wilson, E. B. (1927). On the proof of Sheppard's corrections. *Proceedings of the National Academy of Sciences of the United States of America, 13,* 151–156. http://dx.doi.org/10.1073/pnas.13.3.151

Wu, Y., & Yang, P. (2018). Optimal estimation of Gaussian mixtures via denoised method of moments. *arXiv*. Retrieved from https://arxiv.org/abs/1807.07237

Appendix A

Detailed Design for Our Mixture Decomposition Algorithm

The technical problem of mixture decomposition can be defined as follows.

Technical Problem Definition [Mixture Decomposition]

Given In_{M}: $\{r(es_1), N_1, r(se_1)\}$, $\{r(es_2), N_2, r(se_2)\}$, . . . , $\{r(es_m), N_m, r(se_m)\}$, the objective of mixture decomposition is to find Out_{M}: $\{w_1, \mu_1, \sigma_1\}$, $\{w_2, \mu_2, \sigma_2\}$, . . . , $\{w_k, \mu_k, \sigma_k\}$ that minimizes a predefined *error function* $d(In_{\mathrm{M}}, Out_{\mathrm{M}})$ that measures the distance between the mixture distribution predicted by Out_{M} and the distribution actually observed in practice (i.e., In_{M}).

Somewhat surprisingly, while the existing work on mixture decomposition feature deep and complex theoretical results, the general workflow of the state-of-the-art algorithms for mixture decomposition are quite simple. A basic version of it is to just enumerate all possible value combinations of $\{w_1, \mu_1, \sigma_1\}$, . . . , $\{w_k, \mu_k, \sigma_k\}$, in order to find the one that minimizes the error function (Kalai et al., 2012). With such a simple workflow, solving the mixture decomposition problem centers on addressing two main issues, which we refer to as the *identifiability challenge* and the *computational challenge*, respectively. The identifiability challenge is to properly define the error function $d(In_{\mathrm{M}}, Out_{\mathrm{M}})$, so the value of Out_{M} that minimizes the error function indeed represents the mixture components to be recovered. The computational challenge, on the other hand, focuses on finding ways to reduce the computational time required by the simple enumeration. While addressing the computational challenge is an important part of the computer science research on this topic (e.g., Daskalakis & Kamath, 2014), it is unlikely to be a concern for our purpose due to the usually small input size in meta-analysis. Thus, in most part of this appendix, we will focus on addressing the identifiability challenge, specifically by discussing how to develop an error function $d(In_{\mathrm{M}}, Out_{\mathrm{M}})$ that fits the unique requirements of meta-analysis. Later in the appendix, we will borrow heavily from the computer science literature to identify potential solutions should the computational challenge arise. At the end of the appendix, we discuss the technical limitations of our algorithm and the potential directions for future research.

(*Appendices continue*)

Identifiability Challenge

Note from the problem definition that the error function $d(In_M, Out_M)$ is supposed to measure the distance between the mixture distribution predicted by Out_M, which we denote as G_{out}, and the ground-truth effect-size distribution estimated from In_M, which we denote as G_{in}. With this requirement, the definition of the error function essentially boils down to defining a *synthesized data structure* Θ for G_{out} and G_{in}, so the error function can be expressed as the vector norm of the difference of Θ between the two distributions:

$$d(In_M, Out_M) = \|\Theta(G_{in}) - \Theta(G_{out}))\|. \quad (7)$$

Many different forms of Θ have been used in the modern implementation of mixture decomposition, including a vector formed by the first six statistical moments (i.e., $E[(X - E(X))^i]$ where E represents the expected value and $i = 1, \ldots, 6$) of the distribution (Kalai et al., 2010), the first three moments (Anandkumar, Hsu, & Kakade, 2012), the cumulative density function (Daskalakis & Kamath, 2014), and so forth. All these forms of Θ have been proved to have the *identifiability property*, that is, so as long as two distributions G_1 and G_2 are close on Θ (i.e., $\Theta(G_1) \approx \Theta(G_2)$), every mixture component of the two distributions must also be similar to each other.

Unfortunately, a unique challenge in meta-analysis, the presence of sampling error in $r(es_i)$, prevents us from directly using many of these definitions of Θ. Specifically, note that our input to the mixture decomposition algorithm is *not* a sample of the mixture distribution (as is commonly assumed in the literature of statistics or computer science; Dasgupta, 1999; Kalai et al., 2012). Instead, each reported effect size $r(es_i)$ can be considered as the *sum* of a sample from the mixture distribution (i.e., es_i) and a sample from its sampling-error distribution (i.e., se_i), which varies from one primary study to another. To properly define the error function, we must ensure an accurate estimation of $\Theta(G_{in})$ based on the input data In_M.

A seemingly simple idea to address the challenge is to use statistical moments (like Kalai et al., 2010) as the synthesized data structure Θ, and *estimate* $\Theta(G_{in})$, that is, the moments of the mixture distribution, based on the reported effect sizes and standard errors. Unfortunately, there are two obstacles facing the implementation of this idea: First, while numerous methods (e.g., Hedges & Vevea, 1998; Hunter & Schmidt, 2004; White, 1980) have been proposed for the correction of sampling error in estimating the first two statistical moments (i.e., mean and variance), these studies generally assume the underlying effect-size distribution to be Gaussian (Brannick, Yang, & Cafri, 2011), which contradicts our assumption of it being a mixture of multiple Gaussian distributions. Second and more importantly, it is unclear how these methods can be extended to *higher-order* moments. Note that, while there is no obvious need to estimate any moment of order higher than two under the traditional Gaussian assumption,[18] doing so is essential for the purpose of mixture decomposition because even very simple mixture distributions cannot be accurately decomposed based on only moments of the first two orders (Achlioptas & McSherry, 2005; Chang, 1996). While one could always use a Monte Carlo approach to conduct a brute-force search for an optimal estimate of any higher-order moment, doing so would reveal another, even more fundamental, issue with using statistical moments to decompose a mixture distribution in meta-analysis: Higher-order statistical moments are known to be inherently *unstable* when the sample size is small because, given the definition of the i-th moment (i.e., $E[(X - E(X))^i]$), the larger i is, the more influence an outlier likely has on the moment estimation (Kim & White, 2004; Kothari, Steinhardt, & Steurer, 2018). As a result, even modern implementations of the method of moments for mixture decomposition tend to require a larger sample size than what is normally available for a meta-analysis (Hardt & Price, 2015).

One way to allow for the correction of sampling error in $\Theta(G_{in})$ is to define Θ as the *histogram* of the input distribution, specifically an h-dimensional vector corresponding to the (estimated) probability for the mixture distribution to fall within h equal-width bins:

$$\Pr\{es \in (b, b + d]\}, \Pr\{es \in (b + d, b + 2d]\}, \ \ldots, \ \Pr\{es \in (b + (h - 1)d, b + hd]\}$$

where es is a sample from the mixture distribution, d is the width of each bin, and $(b, b + hd]$ is the range for the histogram. While the parameters of the histogram, specifically the values of h, b, and d, are important for the proper running of the algorithm, we relegate the detailed discussions of their setup to Appendix B. Assuming a proper setup of the parameters, we now discuss why this histogram structure is a proper definition of Θ, and how it enables the correction of sampling error in estimating $\Theta(G_{in})$ for the ultimate computation of $d(In_M, Out_M)$.

[18] Because the i-th ($i > 2$) moment of any Gaussian distribution $N(\mu, \sigma^2)$ is either 0 (when i is odd) or a function of i and σ when i is even (specifically, $(i - 1)!! \cdot \sigma^i$, where $(i - 1)!!$ is the product of all odd integers between 1 and $i - 1$).

(*Appendices continue*)

First, the identifiability of the histogram data structure, that is, two distributions close on the histogram must also be close on their mixture compositions, can be readily established from the identifiability property of the probability density function (Teicher, 1961), as histogram is simply a discretized version of the probability density function. Second, and more importantly, $\Theta(G_{\text{in}})$ can be easily estimated from the reported effect sizes $r(es_1), \ldots, r(es_n)$ and their standard errors $r(se_1), \ldots, r(se_n)$. Specifically, we can estimate the i-th element of $\Theta(G_{\text{in}})$, that is, the probability for the effect size to belong to range $(b + (i-1)d, b + id]$, simply as the average probability for each effect size es_j to fall in the range.

$$\Theta_i(G_{\text{in}}) \approx \frac{1}{2m} \cdot \sum_{j=1}^{m} \left(erf\left(\frac{b + i \cdot d - r(es_j)}{\sqrt{2} r(se_j)} \right) - erf\left(\frac{b + (i-1) \cdot d - r(es_j)}{\sqrt{2} r(se_j)} \right) \right), \quad (8)$$

where $erf(\cdot)$ is the Gaussian error function.

A desirable property of the histogram structure is its *robustness* to outliers (in the sense of robust statistics; Huber, 2011; Kothari et al., 2018). That is, the value of the histogram and, therefore, the error function, never changes drastically with the insertion or deletion of one or a few primary studies. To understand why, note that the total change an outlier study can incur on *all* of $\Theta_1(G_{\text{in}}), \ldots, \Theta_h(G_{\text{in}})$ is $1/m$, no matter how extreme the reported effect size or standard error is. Leveraging this property, we define the error function $d(In_{\text{M}}, Out_{\text{M}})$ is the ℓ_1-norm[19] of the difference between $\Theta(G_{\text{in}})$ and $\Theta(G_{\text{out}})$. This ensures the robustness of the histogram data structure after sampling-error correction, and stands in sharp contrast with the usage of statistical moments as Θ, in which case the impact of one outlier primary study on the error function can be unbounded.

Computational Challenge

As discussed earlier in the section, given the definitions of Θ and the error function, a simple approach to finding Out_{M} is to enumerate all possible value combinations of w_i, μ_i, σ_i, and finding the mixture composition that minimizes the error function. While potentially inefficient in practice, it is a popular method of choice in theoretical studies of the problem (Kalai et al., 2010). For example, if a meta-analysis uses the Pearson correlation coefficient as the effect size, then a reasonable strategy would be to consider 21 candidate values for μ_i: from -1 to 1 with a step of 0.1, 10 candidate values for σ_i: from 0.1 to 1 with a step of 0.1, and nine candidate values for w_i: from 0.1 to 0.9 with a step of 0.1. This way, the total candidate set would contain $21 \times 10 \times 9 = 1{,}890$ single-component mixtures, $(1{,}890 \times 1{,}889)/2 = 1{,}785{,}105$ two-component mixtures, and $(1{,}890 \times 1{,}889 \times 1{,}888)/(3 \times 2) = 1{,}123{,}426{,}080$ three-component mixtures. While enumerating this set of over one billion candidates is doable with today's computing infrastructure, to achieve better precision than 0.1, one may need to improve upon the simple-enumeration method, in order to reduce the computational time of finding Out_{M}. For the sake of simplicity, we focus on the case of two-component mixture when introducing our method to address this computational challenge, and defer discussions of the three-component case (and the method to determine the number of components) to Appendix D.

There are two main ideas for addressing the computational challenge. One is to prune the candidate values for w_i, μ_i, and σ_i for every component. For example, if no observed effect size is below -0.5, then we can safely exclude from consideration all candidates with $\mu_i < -0.5$, because obviously the optimal mixture composition will not include such a component. We discuss in Appendix C the pruning strategies for w_i, μ_i, and σ_i based on the input In_{M}.

The second idea, indeed a commonly used one in the recent literature of mixture decomposition (e.g., Bandi et al., 2019; Daskalakis & Kamath, 2014), is to *only* enumerate the values of w_1, μ_1, σ_1, and then *derive* for a given w_1, μ_1, σ_1 the corresponding w_2, μ_2, σ_2 that minimizes the error function. Obviously, if such a derivation is possible, one can sharply reduce the number of enumerations without affecting the accuracy of Out_{M}. We follow the method developed by Daskalakis and Kamath (2014) for the derivation. Given $\Theta(G_{\text{in}})$ and w_1, μ_1, σ_1, the derivation starts with computing an estimated histogram for the second component, by "deducting" the first component (as defined by w_1, μ_1, σ_1) from $\Theta(G_{\text{in}})$. Specifically, for any $i = 1, \ldots, h$, there is

$$\Theta_i(C_2) \approx \frac{\Theta_i(G_{\text{in}})}{1 - w_1} - \frac{w_1}{2(1 - w_1)} \left(erf\left(\frac{b + id - \mu_1}{\sqrt{2}\sigma_1} \right) - erf\left(\frac{b + (i-1)d - \mu_1}{\sqrt{2}\sigma_1} \right) \right), \quad (9)$$

where $erf(\cdot)$ is the Gaussian error function, and b, d, and h are parameters for the histogram data structure as discussed before.

The next step is to derive the component parameters μ_2 and σ_2 from the estimated histogram $\Theta(C_2)$. The key requirement here is to ensure that the parameter estimates are *robust* to small changes of $\Theta(C_2)$, especially at the extreme ends (e.g., $\Theta_i(C_2)$ where $i \approx 1$ or h). For example, while a natural idea for estimating μ_2 is to directly estimate the mean of C_2 from $\Theta(C_2)$, doing so could violate the robustness requirement, given the well-known sensitivity of mean to extreme values (Huber, 2011). To this end, Daskalakis and Kamath (2014) introduced two well-known robust statistics for estimating μ_2 and σ_2: the estimated *median* of C_2 for μ_2, and the estimated interquartile range (IQR) divided by a constant $2\sqrt{2}erf^{-1}(1/2)$ (where erf^{-1} is the inverse of the Gaussian error function) for σ_2. Specifically, we compute

[19] i.e., the sum of the absolute value of every element in the input vector.

(*Appendices continue*)

$$\mu_2 = b + \left[\min\left(j \,|\, \sum_{i=1}^{j} \Theta_i(C_2) \geq \frac{1}{2}\right) - \frac{1}{2}\right] \cdot d, \tag{10}$$

$$\sigma_2 = \frac{d}{2\sqrt{2}erf^{-1}(1/2)} \cdot \left[\min\left(j \,|\, \sum_{i=1}^{j} \Theta_i(C_2) \geq \frac{3}{4}\right) - \min\left(j \,|\, \sum_{i=1}^{j} \Theta_i(C_2) \geq \frac{1}{4}\right)\right]. \tag{11}$$

before applying the standard Sheppard's correction (Wilson, 1927) to correct for the downward bias caused by the binning of data in the histogram. The robustness of these estimations to outliers directly follows from the robustness of median and IQR (Huber, 2011). Intuitively, it is easy to see that no matter how extreme an outlier effect size is, it cannot change the median or IQR estimate by more than the histogram bin-width d.

The pseudocode in Appendix F shows the detailed design of our mixture decomposition algorithm with the optimization for addressing the computational challenge.

Limitation and Future Technical Development

Our mixture decomposition algorithm has two interrelated limitations, on the small number of mixture components and the potentially high computational complexity, respectively. As discussed earlier in the article, these limitations might not be critical for the specific application of moderator estimation in meta-analysis. Nonetheless, they may become critical for applying the algorithm in a broader set of applications. To understand the boundary conditions of using the algorithm, it is important to examine the feasibility of overcoming these limitations and the potential methods to do so. To this end, we summarize here a series of important recent results in theoretical computer science (e.g., Regev & Vijayaraghavan, 2017) that established the mathematical bounds pertaining to the use of the algorithm, specifically the tradeoff among the following six factors: (a) the number of mixture components k; (b) the sample size m; (c) the dimensionality of the mixture distribution d ($d = 1$ in our case); (d) the degree of separation among mixture components, for example, the minimum distance s between two adjacent means of mixture components; (e) the maximum standard deviation of a mixture component σ; and (f) the maximum tolerable error in recovering the mean of each mixture component δ.

The table in Appendix E summarizes the main findings from this recent set of work, which include both algorithm designs and infeasibility results. For example, the last row in the table depicts an infeasibility result showing that the sample size cannot be less than exponential to the number of mixture components k if the separation between components is arbitrarily small.

There are three important observations from the table. First, the limitation on the number of mixture components in our algorithm is the unique consequence of our decision to require the least amount of assumptions from a researcher when using our algorithm. This can be observed from the last row of the existing algorithm section in the table. Because we do not require researchers to make any a priori assumptions about the degree of separation between different mixture components, essentially allowing any separation $s > 0$, the sample size and computational time required by our algorithm becomes sensitive (i.e., exponential) to k, the number of mixture components. This is the reason why all simulation or experimental results in the literature (e.g., Bandi et al., 2019; Li & Schmidt, 2017) included at most three components when the sample size is below 1,000 and no separation assumption is made about the mixture components. It is also the reason why we offered the caveat earlier in the article that, given the number of primary studies in a meta-analysis rarely exceeds 1,000, our algorithmic design for moderator estimation in meta-analysis only considers $k = 2$ or 3. Note from the table that it has been proven infeasible to ease this limitation on k without making additional assumptions about the data distribution, as shown in the last row of Infeasibility Result.

The second and third observations from the table pertain to ways of easing the limitation if researchers are comfortable imposing certain assumptions on the data. For example, if the mixture components are assumed to have a clear separation (e.g., $s > \sigma\sqrt{\log k}$), Regev and Vijayaraghavan (2017) developed an algorithm that guarantees an estimation error of at most δ when the sample size is polynomial to k and $1/\delta$. The algorithm is also efficient, with computational complexity polynomial to k and $1/\delta$. Because k is now moved from being an exponent to a polynomial factor, one could potentially support a large number of components with a small sample size. While the separation assumption might not hold in the context of meta-analysis (as discussed earlier in the paper), researchers may be able to leverage this algorithm in a primary study where such separation assumptions are reasonable (e.g., when the data points are expected to form separable clusters).

The third and final observation points to an intriguing finding in the third row of the existing algorithms section in the table, that is, when the data points are high-dimensional (e.g., $d > k$). Under this condition, a series of recent work demonstrated the feasibility of exploiting the high dimensionality to launch spectral methods that are capable of decomposing mixture components *arbitrarily close* to each other with a sample size (and computational complexity) polynomial to k, d, and $1/\delta$ (Anandkumar et al., 2014; Belkin & Sinha, 2015; Goyal et al., 2014; Hsu & Kakade, 2013; Huang et al., 2015). In other words, even when the separation assumption is not valid, one can still allow a large number of mixture components if the dimensionality of data exceeds the number of mixture components. For the specific application studied in the paper (i.e., moderator estimation in meta-analysis), it is rare to have multiple dependent variables forming a high-dimensional mixture distribution. Nonetheless, in either a meta-analysis or in a primary study, the high-dimensional algorithms could be useful as an exploratory tool for researchers to gain a holistic understanding of the data distribution.

(*Appendices continue*)

Appendix B

Histogram Setup

In this appendix, we discuss the design of the histogram parameters b, d, and h. First of all, we note that the robustness characteristics discussed earlier in the article, such as the maximum total change of $1/m$ that an outlier study can incur on the histogram structure, hold regardless of the values of b, d, and h. Nonetheless, their values may affect the computational overhead of the mixture decomposition algorithm and the precision of the outputs it generates. Balancing between the two goals is therefore essential in selecting the histogram parameters. The selection of an optimal bin width d has been extensively studied in statistics as the "bandwidth selection" problem (Wand, 1997), with a famous heuristic for Gaussian distribution being $d = 3.49\sigma n^{-1/3}$ (Scott, 1979), where n is the sample size and σ is the standard deviation. For our purpose, we need the histogram to have enough "resolution" for all components of the mixture. Therefore, the bin width we choose should not exceed $3.49 \cdot \min_i(\sigma_i) \cdot n^{-1/3}$, where $\sigma_1, \ldots, \sigma_k$ are the standard deviations of the k mixture components. The challenge here is that σ_i is not available at the time when we have to determine h. Fortunately, we can derive an approximate lower bound for $\min_i(\sigma_i)$ by leveraging a well-known result in order statistics: the standard deviation σ of a Gaussian distribution can be approximated by $\sqrt{\pi} \cdot \lambda_2$ (Hosking, 1990), where λ_2 is the second L-moment of the distribution, i.e., $(E(X_{2:2}) - E(X_{1:2}))/2$, where $X_{1:2}$ and $X_{2:2}$ are the smaller and larger values of a size-2 simple random sample taken from the Gaussian distribution, respectively, and $E(\cdot)$ represents the expected value taken over the randomness of the size-2 sample. Equipped with this result, our procedure for estimating $\min_i(\sigma_i)$ can be stated as follows: First, we order the m reported effect sizes from small to large as $\theta_1, \ldots, \theta_m$, and construct $m - w_{\text{inf}} \cdot m + 1$ *sliding windows* of effect sizes, where w_{inf} is the minimum possible weight of a component (e.g., the default value 0.1 in our method). Each sliding window has size $s = w_{\text{inf}} \cdot m$:

$$\{\theta_1, \ldots, \theta_s\}, \{\theta_2, \ldots, \theta_{s+1}\}, \ldots, \{\theta_{m-s+1}, \ldots, \theta_m\}.$$

While we do not know which θ_i belongs to which component, we do know that the value of λ_2 of either component *cannot* fall below the minimum λ_2 of the $m - w_{\text{inf}} \cdot m + 1$ sliding windows, for the simple reason that these moving windows already contain the most "tightly squeezed" subsets of the reported effect sizes. Thanks to this property, we can now estimate λ_2 for all sliding windows, find the smallest value $\min(\lambda_2)$, and then drive a lower bound on $\min_i(\sigma_i)$ as $\sqrt{\pi} \cdot \min(\lambda_2)$. After that, we set the bin width for the histogram as $d = 3.49 \cdot \sqrt{\pi} \cdot \min(\lambda_2) \cdot m^{-1/3}$. One subtle issue in estimating λ_2 is how to deal with the precision level in reporting effect sizes - e.g., while both $X_{1:2}$ and $X_{2:2}$ may be reported as 0.19, they could still differ on values beyond the second decimal point. To this end, we compute the difference between $X_{1:2}$ and $X_{2:2}$ as half of their precision level, i.e., 0.005 in the above example.

Given the bin width d, the other two parameters for the histogram setup, the lower limit b and the number of bins h, can be easily derived as $b = \theta_1$ and $h = \lceil(\theta_m - \theta_1)/d\rceil$, where $\lceil\cdot\rceil$ represents the ceiling function.

Appendix C

Candidate Values for First Component

In this appendix, we discuss the strategies for pruning w_i, μ_i, and σ_i, in order to minimize the number of value combinations for w_i, μ_i, σ_i to enumerate in the mixture decomposition algorithm. With regard to the weight parameter w_1, we can simply enumerate all possible values from a predetermined minimum weight w_{inf} to $1 - k \cdot w_{\text{inf}}$, with a predetermined precision level ε as interval. Given that many meta-analyses contain a 100 or fewer primary studies, a reasonable set of setting is $w_{\text{inf}} = 0.1$ and $\varepsilon = 0.05$ or 0.1, as any mixture component with weight lower than 0.1 is unlikely to have sufficient representation in the input sample to enable a reliable estimate.

For μ_1, we adopt a pruning idea developed by Daskalakis and Kamath (2014), which only includes the input effect sizes $r(es_1), \ldots, r(es_m)$ as the candidate values for μ_1. As proven by Daskalakis and Kamath (2014), when $m > 20\sqrt{2}\,\sigma/(3w_1\varepsilon)$, where σ is the standard deviation of $r(es_i)$, there is a higher than 99% probability that at least one of the reported effect sizes $r(es_i)$ is within ε of μ_1. Consider a setting of $\sigma = 0.2$, with $w_1 = 0.5$, we have 99% probability that one of $r(es_i)$ is within 0.05 of μ_1 so long as $m > 75.42$. Thus, this pruning idea is unlikely to substantially increase the error of the mixture decomposition output.

For σ_1, however, we found that an effective pruning idea developed by Daskalakis and Kamath (2014) cannot be applied because of a unique constraint in the inputs to meta-analysis. Specifically, the idea is to find the minimum distance between samples and use this observed distance to derive a likely range for σ_1. Intuitively, if no two samples are close to each other, then it is highly unlikely for either component to have a small variance. Unfortunately, implementing this idea in meta-analysis faces an obstacle: Most input effect sizes are only reported to a precision of 0.01, meaning that the minimum distance between samples is almost always 0. As such, we resort to the basic approach for enumerating σ_1 with a step of ε.

(*Appendices continue*)

Appendix D

Three-Component Case

As discussed earlier in the article, our handling of the three-component case is similar to the two-component case with one exception: to disentangle three components, we need to enumerate all possible parameter combinations for both C_1 and C_2, which could infer significant computational overhead if each component has a large number of possible parameter combinations. To alleviate the computational load, one method is to fix one of C_1 and C_2 to a component in the output of the two-component decomposition. Another method is to leverage the same idea for reducing computational complexity in the two-component case: Like how $\Theta(C_2)$ can be derived by deducting C_1 from $\Theta(G_{\text{in}})$ in the two-component case, $\Theta(C_3)$ can be derived by deducting C_1 and C_2 from $\Theta(G_{\text{in}})$. The only additional step occurs when the algorithm is called upon to determine the number of components (i.e., in Line 15).

Determining the "right" number of components has long been an important problem in latent mixture analysis for individual-level data. The fundamental challenge is to decide, as the number of component increases, when the increase in model fit no longer justifies the corresponding decrease of model parsimony. Besides considering the theoretical meaning of the solutions (Foti, Bray, Thompson, & Allgood, 2012), researchers could also resort to standard statistical procedures, which mostly focus on inspecting how various log-likelihood based fit statistics vary with the number of mixture components (Nylund et al., 2007).

This challenge is admittedly not as critical in our method, because the mixture composition is not the final output but only an intermediate result. Nonetheless, if the mixture decomposition algorithm settles on a "wrong" number of components, the statistical power of moderator estimation could be reduced, either because the algorithm mistakenly merges the components corresponding to different moderator levels (i.e., when the number of components is less than ideal), or because the error of mixture decomposition is unnecessarily enlarged (i.e., when the number of components is more than ideal). To address this challenge, we introduce a recently developed procedure for decomposing mixture distributions formed by overlapping components (Bandi et al., 2019). First, we find from *Candidate_Set* the optimal candidate set, i.e., the one that minimizes $d(In_{\text{M}}, Out_{\text{M}})$, with one, two, and three components, respectively. Afterwards, we compute the following log-likelihood function for each of these three mixture compositions (i.e., with $k = 1, 2, 3$, respectively):

$$\mathcal{L} = \sum_{i=1}^{m} \log\left(\sum_{j=1}^{k} \frac{w_j}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}}\right).$$

As specified by Bandi et al. (2019), the three-component candidate is selected if its value of $\mathcal{L}$ is at least $(1 + c)$ times over the two-component solution and at least $(1 + c)^2$ times over the one-component solution, where c is a constant with a recommended value of 0.01. Otherwise, the two-component solution is selected if its value of $\mathcal{L}$ is at least $(1 + c)$ times over the one-component solution. Failing both, the single-component solution is selected.

Appendix E

Summary of Boundary Conditions for Mixture Decomposition

Condition	Sample size m	Running time	Reference
Existing Algorithms			
$s > \sigma$, $k = 2$	$\leq$ poly(k, d, 1/δ)	$\leq$ poly(k, d, 1/δ)	Balakrishnan, Wainwright, & Yu, 2017
$s > \sigma\sqrt{\log k}$	$\leq$ poly(k, d, 1/δ)	$\leq$ poly(k, d, 1/δ)	Regev & Vijayaraghavan, 2017
$s > 0$, $d > k^c$ ($c > 1$)	$\leq$ poly(k, d, 1/δ)	$\leq$ poly(k, d, 1/δ)	Huang, Ge, Kakade, & Dahleh, 2015
$s > 0$	**$\leq e^k$**	**$\leq$ poly(k, d, 1/δ)$^{k^2}$**	**Moitra & Valiant, 2010**
Infeasibility Results			
$s \leq \sigma\sqrt{\log k}$	$> k^c$, where $c > 1$	N/A	Regev & Vijayaraghavan, 2017; Anderson, Belkin, Goyal, Rademacher, & Voss, 2014
$s \leq c$, (c is a constant)	$\geq e^k$	N/A	Moitra & Valiant, 2010

Note. The bolded row represents the assumption made in our article and the corresponding requirements on sample size and running time. Poly means a polynomial function of the input variables. All the bounds in the table are asymptotic, as we did not include the Big-O notation (specifically $O(\cdot)$ for the upper bounds and $\Omega(\cdot)$ for the lower bounds; Knuth, 1997, Section 1.2.11 for the sake of simplicity).

(*Appendices continue*)

Appendix F

Pseudocode for the Mixture Decomposition Algorithm

Algorithm for mixture decomposition

1: Compute $\Theta(G_{in})$ according to the histogram parameters b, d, and h specified in Appendix B
2: *Candidate_Set* $\leftarrow$ {{1, μ, σ, 0, 0, 0}}, where μ, σ are the estimated mean and standard deviation of effect sizes after artifactual corrections, respectively. Note that this candidate is corresponding to a single Gaussian distribution with mean μ and standard deviation σ.
3: For each possible value combination w_1, μ_1, σ_1 (as defined in Appendix C) as C_1
4: Compute $\Theta(C_2)$ by deducting C_1 from $\Theta(G_{in})$
5: Derive μ_2 and σ_2 from $\Theta(C_2)$
6: Insert {w_1, μ_1, σ_1, $(1 - w_1)$, μ_2, σ_2} to *Candidate_Set*
7: End For
8: For each value combination w_1, μ_1, σ_1 (as defined in Appendix C) as C_1
9: For each value combination w_2, μ_2, σ_2 with $w_2 < 1 - w_1$ as C_2
10: Compute $\Theta(C_3)$ by deducting C_1 and C_2 from $\Theta(G_{in})$
11: Derive μ_3 and σ_3 from $\Theta(C_3)$
12: Insert {w_1, μ_1, σ_1, w_2, μ_2, σ_2, $(1 - w_1 - w_2)$, μ_3, σ_3} to *Candidate_Set*
13: End For
14: End For
15: Find the mixture composition in *Candidate_Set* that minimizes $d(In_M, Out_M)$, according to the procedure in Appendix D.

Received January 18, 2020
Revision received September 3, 2020
Accepted September 4, 2020 ■