AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2018 American Psychological Association
0096-3445/18/$12.00

http://dx.doi.org/10.1037/xge0000504

When Averaging Goes Wrong: The Case for Mixture Model Estimation in Psychological Science

David Moreau and Michael C. Corballis
The University of Auckland

Recent failed attempts to replicate numerous findings in psychology have raised concerns about methodological practices in the behavioral sciences. More caution appears to be required when evaluating single studies, while systematic replications and meta-analyses are being encouraged. Here, we provide an additional element to this ongoing discussion, by proposing that typical assumptions of meta-analyses be substantiated. Specifically, we argue that when effects come from more than one underlying distributions, meta-analytic averages extracted from a series of studies can be deceptive, with potentially detrimental consequences. The underlying distribution properties, we propose, should be modeled, based on the variability in a given population of effect sizes. We describe how to test for the plurality of distribution modes adequately, how to use the resulting probabilistic assessments to refine evaluations of a body of evidence, and discuss why current models are insufficient in addressing these concerns. We also consider the advantages and limitations of this method, and demonstrate how systematic testing could lead to stronger inferences. Additional material with details regarding all the examples, algorithm, and code is provided online to facilitate replication and to allow broader use across the field of psychology.

Keywords: mixture modeling, expectation–maximization, meta-analysis, replication, effect size

Thirty-six percent[1]. The relatively low percentage of psychology studies successfully replicated in the recent initiative led by the Open Science Framework (Open Science Collaboration, 2015) has sparked an intense debate within academia (e.g., Anderson et al., 2016; Gilbert, King, Pettigrew, & Wilson, 2016; Lindsay, 2015), relayed shortly after in the media. On the one hand, some have argued that low replication is a sign that the foundations of psychological science are shaky, and that many published findings are probably false, echoing the popular claim about the medical literature (Ioannidis, 2005). In contrast, others have pointed out that the initiative itself shows the effectiveness of self-correcting mechanisms in science, and thus that its conclusions should be praised (Bohannon, 2015; but see also Ioannidis, 2012).

Among several measures intended to improve the strength of psychology as a field (see e.g., Bersoff, 1999; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012), recent discussions have emphasized the need for larger sample sizes (e.g., Schimmack, 2012). For a fixed effect size and Type I error rate, a larger sample size results in an increase in statistical power, the probability to detect an effect if it is present. In this context, meta-analyses represent an important step toward refining estimates of effect sizes and precision—more data points amount to more power and

David Moreau and Michael C. Corballis, School of Psychology and Centre for Brain Research, The University of Auckland.

This methodological contribution was inspired by the recent push toward greater replicability in psychological research. David Moreau is particularly sensible to this issue given his work on cognitive interventions, a field where discrepancies and inconsistencies are numerous. We believe that finer analyses can contribute to more robust and replicable findings, an observation that motivated the development of tools to help refine our understanding of psychological phenomena, with the potential to benefit a large body of research. In this context, we hope that our contribution can facilitate detecting and accounting for deviations from typical assumptions, both in single experiments and meta-analyses, so as to refine cumulative science. This article is part of a larger project dedicated to the improvement of methodological practices and analyses in psychology and neuroscience, within the Brain Dynamics Lab at The University of Auckland (David Moreau is the principal investigator). The tools we develop (publications and repositories) can be found on the lab website, *BrainDynamicsLab.com*. We are grateful to Kristina Wiebels for her comments and suggestions on an earlier version of this article and to Andrew Heathcote and Quentin Gronau for their insightful comments. Part of this work was presented at the Association for Psychological Science conference in San Francisco, California, in May 2018. All materials, codes, and data are available at https://github.com/davidmoreau/MixtureModel. David Moreau and Michael C. Corballis jointly generated the idea for the study. David Moreau programmed all simulations and methods, wrote the article, and created the online repository. Both authors critically edited the manuscript and approved the final submitted version.

Correspondence concerning this article should be addressed to David Moreau, School of Psychology and Centre for Brain Research, The University of Auckland, 23 Symonds Street, Office 227, Auckland 1010, New Zealand. E-mail: d.moreau@auckland.ac.nz

[1] The exact figure changes slightly based on the specific criterion used to determine SRs. When estimated from statistical significance, SR = .36; when based on subjective ratings, SR = .39. Finally, when inferred from effect sizes (overlapping 95% confidence intervals), SR = .47.

to more confidence about specific claims. Indeed, an important rationale for meta-analyses is that of increased power, compared with single studies—effects that might be too subtle to emerge from limited and noisy data can, when pooled together, reach significance.

This rationale is contingent upon meta-analyses leading to increases in precision around true effect sizes, that is, increases in accuracy. The latter is an assertion that, we argue, requires substantiation, given that it makes at least two implicit assumptions. First, it postulates that effect sizes come from a single distribution; if they do not, averages can be misleading, and possibly erroneous. Second, typical meta-analytic models assume that effect sizes are distributed continuously, at least in theory (e.g., Hedges & Vevea, 1998). The latter is often acknowledged in meta-analyses, as deviations from normality can be quantified, for example via model fit indices. In contrast, the former is supposedly controlled for by random-effect models, yet even these models typically assume a normal, continuous distribution for the underlying effects. As we discuss in the article, this assumption has important limitations, with the potential to blur entire fields of research.

In this article, we propose to directly test for the plurality of distributions of effect sizes, before averaging. The idea complements typical quality controls undertaken prior to including studies in a meta-analysis—based on well-defined criteria, the researcher decides to include or exclude studies in the final analysis. Similarly, we argue, we should test *mathematically* whether effect sizes can be combined together, or if they are thought to have come from different distributions. In the latter case, no elaborate model or random-effects analyses can make up for the latent heterogeneity in the data. We explain the rationale for modeling underlying distributions of effect sizes, using the example of brain training. We then propose the use of mixture models, to test for the plurality of distribution modes (i.e., mixture components), with the goal to provide a more accurate description of a series of findings. This data-driven approach allows for minimal assumptions about distribution parameters, instead using available observations to build adequate models. Finally, we discuss the advantages and limitations of the proposed method, and suggest that it could allow for more accurate descriptions and analyses of a body of evidence. To facilitate reproducibility and allow the reader to extend upon our analyses, we provide all the code and data online (https://github.com/davidmoreau/MixtureModel).

Why Model Distributions of Effect Sizes?

We illustrate herein the importance of modeling effect size distributions with the example of brain training. This field of study has gained traction in recent years, for good reasons, including the prospect of finding an activity that can elicit transfer to a wide range of cognitive tasks, with apparently very few side effects (for a comprehensive review of this field of research, see Simons et al., 2016). In short, the rationale is that a specific training regimen consisting of a single or multiple cognitive tasks can potentially lead to generalized improvement on a variety of similar (near transfer) or apparently unrelated tasks (far transfer). On the surface, this trend of research seems to contradict decades of evidence demonstrating the specificity of expertise (Chase & Simon, 1973; Ericsson, Krampe, & Tesch-Römer, 1993) and appears to be at odds with the highly heritable character of numerous cognitive abilities and traits (Benyamin et al., 2014; Davies et al., 2011).

Beyond its obvious implications, the attention that this field of research has received is also driven by inconsistencies across findings. For example, some laboratories have consistently found sizable effects of working memory training on fluid intelligence (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi, Buschkuehl, Shah, & Jonides, 2014), whereas others have systematically failed to replicate these findings (Harrison et al., 2013; Thompson et al., 2013), even in close replications of the original studies (Redick et al., 2013). These discrepancies also extend to meta-analyses, with working memory training appearing to be effective (Au et al., 2015; Karbach & Verhaeghen, 2014) or not (Dougherty, Hamovitz, & Tidwell, 2016; Melby-Lervåg & Hulme, 2013), depending on the particular research group assessing the literature. Although it has been suggested that failures to take into account individual differences may be partly responsible for such inconsistencies (Jaeggi et al., 2014; Moreau, 2014), only systematic idiosyncrasies could explain the tendency for particular laboratories to consistently find an effect or fail to do so.

What mechanism could be responsible for findings that are internally reliable but externally inconsistent? Suppose that whether brain training elicits transfer or not is moderated by subtle details in the experimental setup, such as unreported design specificities, the particular population that was sampled from, or cues given by the experimenters. Brain training studies are rarely double-blind, and even if protocols are run by research assistants supposedly naive to the specific hypothesis being tested, previous literature published by a research group is typically known by all individuals involved in a project. In this context, it may not be that an effect exists or that it does not (i.e., brain training either “works” or “doesn’t work”), but rather that either conclusion can hold under specific circumstances, depending on experimental conditions.

This is a subtle but important consideration, because failures to account for confounded distributions can blur conclusions as evidence accumulates, as we illustrate in the next section. In this context, correctly modeling mixture distributions is critical to refine meta-analyses, and thus appears particularly valuable in the evaluation of replications. In numerous cases, modeling underlying distributions can substantially improve assessments of cumulative evidence—models accounting for multimodal distributions may provide finer estimates than those based on unimodal distributions. We further propose that many of the apparent discrepancies in the aggregation of studies in psychology and neuroscience have emerged from sampling two or more underlying distributions, rather than a single population (e.g., Nord, Valton, Wood, & Roiser, 2017).

Mixture Models to Refine Cumulative Evidence

One particularly appealing approach for dealing with this kind of problems lies in the comparison of a model that assumes a single source of effect sizes—and therefore a single, typically Gaussian, distribution—versus a model that allows for multiple sources of effect sizes, oftentimes best described with a multimodal distribution. In the latter case, one can then estimate the probability of each effect size coming from a given distribution. Once the underlying distribution of effect sizes—or of individual data points in the case of a single study—has been accurately modeled, the probability of belonging to a given distribution can be computed for each effect size.

This idea is better illustrated with a concrete example. Suppose we conduct an experiment to estimate the effect of an intervention on cognitive abilities. In a minimalist design, we randomly assign participants to either an experimental or an active control group. In addition, let us assume that the true effect we set to detect is $d = 0.5$, that is, the treatment is effective and leads to improvements of half a standard deviation on average. With a small sample size of 20 participants per condition, we would typically find that the true effect falls within our confidence interval, even though our estimate would be fairly imprecise. With a large sample size of 500 participants per condition, our estimate would be refined, as the error is reduced. In a meta-analysis of 25 studies with a small sample size of 20 participants per condition (making a total of 500 participants for each condition), our estimate would differ slightly from that of the large study and precision would be smaller, given that effects are modeled as random. Yet regardless of the specifics about the way observations are collected and aggregated—whether with a large single study or with a combination of smaller ones—larger sample sizes lead to increases in power, thus providing estimates that are both more accurate and more precise (Figure 1A).

In this example, the meta-analysis model is a random-effects model, because the simulated studies represent a random selection from a larger population of studies (for an excellent discussion of the difference between fixed- and random-effects models, see Hedges & Vevea, 1998). The random-effects model is based on restricted maximum-likelihood estimation[2], a method that provides an estimate for the amount of heterogeneity and takes that variability into account in the overall model (see e.g., Kalaian & Raudenbush, 1996). Although there is some variability in the precise method used in meta-analyses (Hedges & Vevea, 1998; Kelley & Kelley, 2012; Schmidt, 1992), this process matches how evidence is typically accumulated in the behavioral sciences, with each additional study thought to contribute to the refinement of our understanding about a particular research question.

In a number of cases, however, this may not be a realistic scenario. As recent discussions have suggested, many of the effects psychologists study in the lab are volatile—they can appear under the right circumstances and vanish under alternative conditions (Open Science Collaboration, 2015; for an illustration in the field of brain training, see Shipstead, Redick, & Engle, 2012). To increase accuracy with the accumulation of evidence, this factor should be taken into account (Johnson, 2013; Kamary, Mengersen, Robert, & Rousseau, 2014; Marin, Mengersen, & Robert, 2005). Continuing with our earlier example, imagine now that instead of studying a constant effect size of half a standard deviation, we set out to detect an effect size of $d = 1$ present in some instances (say, $P = .4$) but not in others ($P = .6$). What would the literature look like in this case? Some well-powered studies would detect an effect worth reporting by frequentist standards of statistical significance (i.e., $p < .05$). Others studies would not allow rejecting the null-hypothesis, despite adequate power. The literature would be inconsistent, and yield disparate results.

This is exactly what we observe if we simulate such a model (Figure 1B). When more than one distribution is allowed to contribute to a sample of effect sizes, precision and accuracy no longer covary. If distributions are segregated (i.e., if they are not mixed with one another in single studies), a small study may yield a substantially different effect size from a large study, obscuring inferences drawn from the data and undermining out-of-sample predictions. More prejudicial still, a meta-analysis of 25 small studies provides a false impression of resolution, with an overall effect size estimated to lie somewhere between the two true population effect sizes. The inherent problem is that the degree of certainty associated with a point estimate typically increases with greater power. In cases where only one distribution is allowed to contribute to a sample of effect size, this may be reasonable (Figure 1A). However, this assumption is no longer valid if multiple distributions contribute to a sample of effect sizes (Figure 1B). In the latter case, greater power leads to an increase in precision around an erroneous effect size, with potentially detrimental consequences on applied policies and decisions.[3]

So, effect sizes vary, and cannot always be trusted to represent estimates of a single underlying effect. Is this not what random-effects models in meta-analyses are designed to account for? Not exactly. Using random effects to model effect sizes coming from multiple sources is akin to scrambling eggs, only to later attempt to separate egg yolks from the whites (for a discussion of the assumptions and limitations of random-effects meta-analysis, see Higgins, Thompson, & Spiegelhalter, 2009). In contrast, by using mixture modeling–the identification and estimation of subpopulations within a distribution–we propose to deal with this issue before it becomes problematic, that is, before averaging (see Appendix A for details).

Random-effects models, in this context, have other important limitations; they do not take into account the inherent uncertainty in the between-study variance estimate (the profile likelihood approach, which uses nested iterations for converging, is an exception). Even post hoc corrections, such as the Hartung–Knapp–Sidik–Jonkman method (Hartung & Knapp, 2001; Sidik & Jonkman, 2002), do not allow disentangling underlying distributions of effect sizes. Figure 1B provides a visual example of these limitations—even though the meta-analysis depicted uses a random-effects model, the average estimate is wrong. The problem does not lie in the averaging method per se, but rather in its application.

Note that the problem of combining results from different experiments is not new: work has been developed previously with the goal to model uncertainty when the exact pooling specifications are unknown (Evans & Sedransk, 2001; Malec & Sedransk, 1992). This approach is especially popular in the field of individual differences, whereby differences across cognitive processes and variation within these processes are the central subject of investigation (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014; Dennis, Lee, & Kinnell, 2008). Similarly, others have implemented Bayesian methods to combine results from binomial experiments (Consonni & Veronese, 1995) or to refine meta-analytic models by first

[2] The restricted maximum-likelihood method is usually preferable to maximum likelihood models because the latter is biased when sample size is small. As N increases, results from both methods tend to converge (for an in-depth discussion, see Kelley & Kelley, 2012).

[3] By extension, ignoring a mixture structure *within* a study can potentially introduce bias when estimating effect sizes. Although perhaps less plausible in experimental psychology research, this problem can be partly remedied with the use of hierarchical mixture models, whereby individual and group parameters are estimated separately but within the same model.

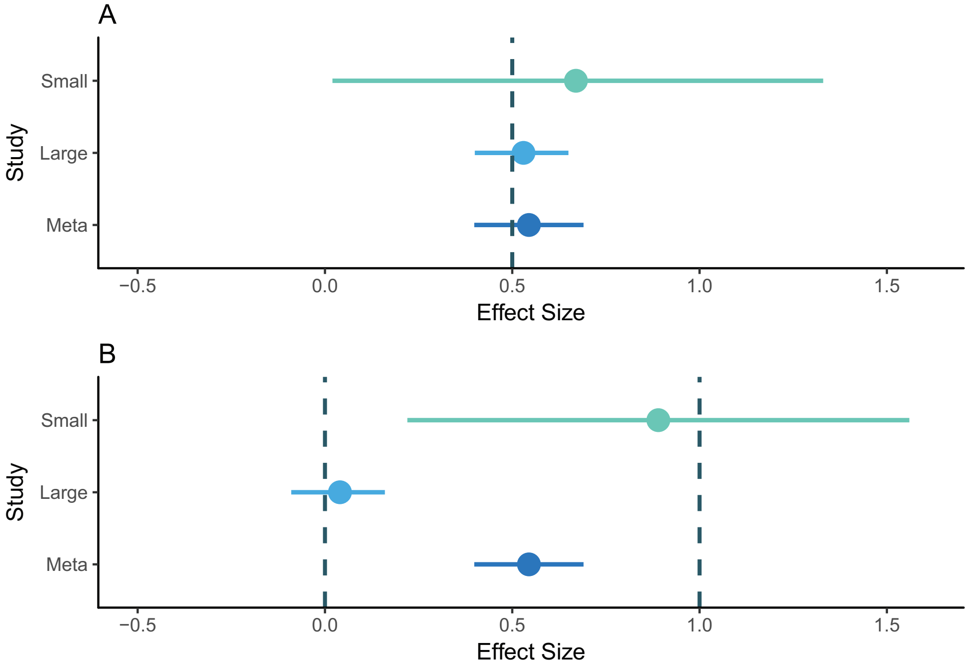

Figure 1. Underlying distributions of effect sizes affect accuracy and precision differently. Here, we present mean effect sizes from simulated independent *t* tests performed on two samples randomly drawn from two simulated populations, which differ by half a standard deviation ($d = 0.5$). "Small," "large," and "meta" studies depict simulations with $N = 20$, 500, and 20 (resampled 25 times), respectively, per cell. The vertical dotted line shows the true effect sizes. (A) In the typically assumed scenario of a single underlying distribution of effect sizes, increases in power lead to more precise estimates. Here, additional power allows point estimates closer to the true effect, $d = 0.5$, with all types of study showing a fair estimate of the true effect size. Precision and accuracy increase together. (B) When more than one underlying distribution is present, however, precision and accuracy may not vary together. In this scenario, an effect was either present ($d = 1$, $P = .4$) or not ($d = 0$, $P = .6$). In such case, studies with high statistical power provide more precise estimates of a specific, nonrepresentative effect size (large experiment) or a precise estimate of the wrong effect size (meta-analysis), leading to a decrease in overall accuracy. See the online article for the color version of this figure.

estimating the clustering structure of effect sizes (Moreno, Vázquez-Polo, & Negrín, 2017). Although these approaches are perhaps more elegant mathematically than the one we describe in this article, they require prior model specifications that render them less user-friendly and have prevented their systematic implementation. Here, we have based our estimation of mixture proportion on a widely implemented algorithm, with well-known properties, which could facilitate a wider use in the experimental psychology community. This idea is gaining traction beyond this specific problem—mixture modeling has recently been applied to distributions of *p* values in the context of psychological research, to help refine publication bias estimates (Gronau, Duizer, Bakker, & Wagenmakers, 2017).

Before moving on to potential remedies, let us ponder the plausibility of the scenario depicted in Figure 1B. One might assume that such dichotomy is uncommon, perhaps specific to the particular example we chose. Given the striking between-labs variability but within-labs consistency in experimental results, we suspect it is not (Au et al., 2015; Dougherty et al., 2016; Karbach & Verhaeghen, 2014; Melby-Lervåg & Hulme, 2013; Simons et al., 2016). At the behavioral level, moderating factors could include expectancy effects (Boot, Simons, Stothart, & Stutts, 2013) or experimenter bias, undocumented differences in protocol (Schultz, 1969), systematic recruiting methods, or targeted populations. For example, cerebral laterality and handedness have been shown to influence a wide range of cognitive abilities, ranging from line bisection judgments (Scarisbrick, Tweedy, & Kuslansky, 1987) to spatial reasoning (Somers, Shields, Boks, Kahn, & Sommer, 2015) and intelligence (Papadatou-Pastou & Tomprou, 2015). These effects can in some instances be indirect; for example, it has been argued that handedness may mediate the influence of other variables, such as stereotype threat, on reasoning task performance (Wright & Hardie, 2015).

Additional sources of variations can be found at different levels of investigation. At the neural level, discrepancies could be based on individual thresholds for long-term potentiation or depression/depotentiation, also critical factors of plasticity (Ridding & Ziemann, 2010), and greatly affected by additional factors such as age (Barnes, 2003). At the genetic level, the difference between the presence or absence of an effect could be variations in specific polymorphisms, such as *COMT* or *BDNF*, known to influence cortical plasticity (Witte et al., 2012), response to training interventions (Moreau, Kirk, & Waldie, 2017), and typically distributed unequally across subpopulations (González-Castro et al., 2013;

Moreau et al., 2017; Petryshen et al., 2010). Given the sensitivity to protocol specificities inherent to experimental designs in general, and of repeated-measures experiments in particular, this list is only an excerpt of the variables of potential influence. In addition, and because the focus of psychology experiments is often on participants' behavior, researchers might be oblivious to extraneous variables, with unintended but conspicuous consequences (e.g., Simmons, Nelson, & Simonsohn, 2011).

In theory, these disparities should be equated via random assignment, yet such assumption is highly dependent upon appropriate sample sizes, with typical designs being rarely satisfactory in this regard (e.g., Moreau, 2014). Besides, random assignment only controls for interindividual variability if the latter is truly random, yet in many instances it is not the case. If the source of error, or bias, is systematic, this assumption no longer holds.

Furthermore, outcomes do not have to be binary to give rise to problematic and potentially misleading aggregates. Whenever effect sizes come from a discrete distribution, averaging methods can yield imprecise estimates. From the central limit theorem, we know that the sampling distribution of the mean representing repeated draws from a multinomial distribution eventually approximates a normal distribution, but it does not follow that the arithmetic mean necessarily reflects a genuine possibility on a single study or for a single individual. If one throws a fair six-sided die a number of times and compute the tally at the end of the sequence, the mean will soon approximate 3.5, even though 3.5 is not a possible outcome on a given trial. In order to model the behavior of a single die, one needs to be mindful when averaging outcomes. Thus, only if we consider that effect sizes truly vary along a continuum can we disregard the aforementioned reflections.

This idea has important implications, which perhaps depend on the primary goal of meta-analysis research. If the aim is to provide estimates of mean effect sizes, so as to best reflect the average outcome of a set of studies, then estimates that are not representative of a possible outcome are not necessarily problematic. However, with the growing emphasis on prediction in psychology (see e.g., Yarkoni & Westfall, 2017), we might want to know the typical outcome of an experimental manipulation, a treatment, or an intervention. If the goal is to predict future outcomes, the ability to discard impossible outcomes is critical. In that sense, modeling effect size distributions can have important benefits. To be clear, we are not arguing that every distribution of effect sizes is plurimodal. We are not even arguing that it makes for the majority of cases—recent evidence from the Many Labs Project (e.g., Klein & Ratliff, 2014) or the Reproducibility Project: Psychology (Open Science Collaboration, 2015) may suggest that seemingly important variables do not always moderate effects (although we should point out that these large-scale projects may not be representative of the typical effects in psychology). But because mixture estimation is straightforward to implement and can provide additional insight into one's data, we believe that the plausibility of mixtures should be estimated, as these can genuinely inform conclusions based on cumulative evidence. In the next section, we explain how we propose to deal with such problems, to allow modeling multiple distributions, and finer estimates of a body of research.

Applications to Meta-Analyses

Continuing with our earlier brain training example, suppose that instead of assuming a single normal distribution, we directly test the plausibility of this assumption. Based on the expectation–maximization (EM) algorithm (see Appendix B for details), we can estimate the number of underlying distributions in the available sample of effect sizes (i.e., how many modes the distribution contains). Once we have determined the most plausible number of distributions, we can estimate posterior likelihoods for each effect size. These likelihoods allow quantifying the probability of coming from a given distribution for each data point. We can then infer the overall proportions of the entire set of data points that come from a given distribution. These estimates are known as mixing weights, or mixing coefficients.

In the simple but representative meta-analysis initially described in Figure 1, this method allows for a more precise account of effect sizes, with a distinction between studies that tend to suggest the absence of an effect, and those that suggest a moderate effect. Specifically, the mixing weights allow differentiating between two underlying distributions, one with $\lambda = .64$ and the other with $\lambda = .36$ (Figure 2A). Roughly speaking, this indicates that the overall population contains two distributions, one including 64% of the data points, while the other contains 36% of them. Each cluster of studies can then be aggregated separately, to provide precise estimates of effect size when an effect is present versus when it is not, or when an effect is ecologically meaningful versus when it is of limited influence. In our earlier example, this approach allows modeling two normal distributions, centered at $d = 0.06$ and $d = 1.22$, respectively (Figure 2B), making for a more accurate model of the underlying data (Figure 2C).

Interestingly, one can then look for reported or unreported differences in samples, designs, or analyses, so as to gain a better understanding of the specific circumstances that cause or prevent the effect under study. This is a clear improvement over averaging methods that do not take distribution properties into account, as the approach provides a more accurate depiction of reality—in this case, two separate effect sizes, with very different implications. It also highlights the potentially pernicious effect of typical meta-analytic techniques when conducted on inconsistent data. Importantly, the same approach can be applied to a single replication; rather than modeling mixture distributions from the effect sizes generated by all studies, the algorithm can be run on vectors of individual data points, to infer how likely they are to come from different distributions.

Let us gauge the reliability of this method on a real dataset. Consistent with our earlier example, we used a recent meta-analysis in the field of brain training (Melby-Lervåg, Redick, & Hulme, 2016). Specifically, the study investigated the evidence for far-transfer following working memory training, that is, transfer to abilities that were not directly targeted in the training regimen. We applied the procedure presented on simulated data (and further detailed in the online material and R code) to the vector of effect sizes (Hedge's *g*) across studies. Results showed evidence for a mixture distribution with multiple components (Figure 3A). Further analyses indicated that a two-component solution was sufficient, after penalizing for model complexity (i.e., total number of components). The corresponding log-likelihoods, estimated for each data points, are shown in Figure 3B. The mixing weights

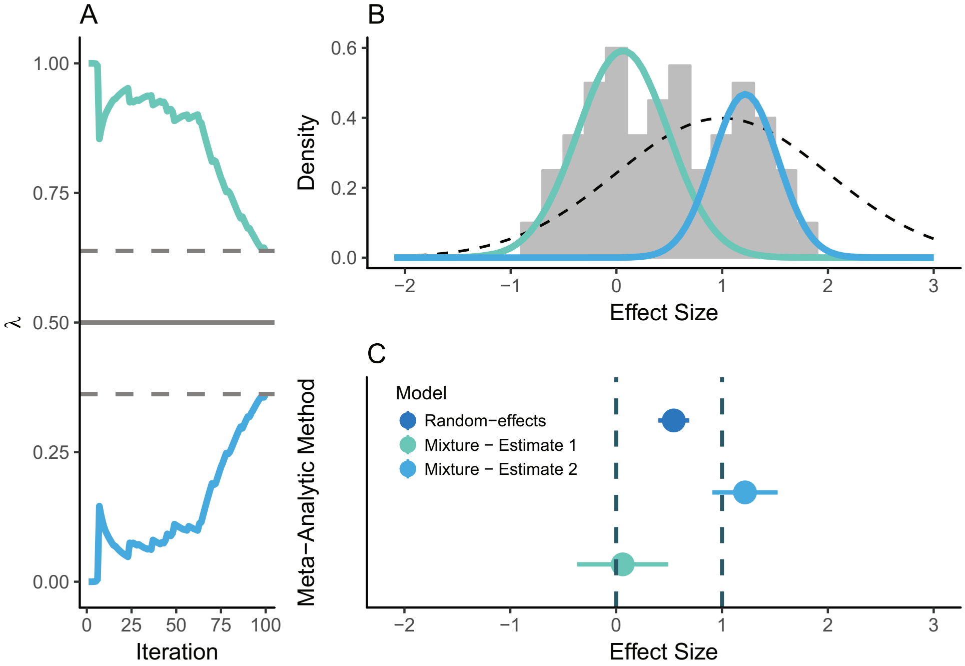

Figure 2. Mixing weights estimation and probability densities for simulated mixture. (A) Posterior likelihoods, given by the expectation–maximization algorithm (see https://github.com/davidmoreau/MixtureModel), show a random walk that eventually converges to the true mixing weights ($\lambda = .64$ and $\lambda = .36$, respectively, horizontal dashed lines). The mixing weights help determine the shape of the underlying distribution of effect sizes, based on incomplete information (i.e., a given sample of effect sizes). In this example, effect sizes (Cohen's *d*) are simulated from a mixture distribution of two single Gaussian distributions $N(\mu, \sigma^2)$ with $\mu = 0$ or $\mu = 1$, with probability $P = .6$ and $P = .4$, respectively), and $\sigma^2 = 0.25$. (B) The expectation–maximization algorithm evaluates model fit (log-likelihoods) for various numbers of Gaussian components by considering up to *k* components, and incrementally increasing the size of the mixture (number of components) as long as the difference between *k* and $k + 1$ is significant (at the 5% level in the implementation we report in the article, but this parameter can be adjusted). The *p*-value tests the difference between H_0 (*k* components) and H_1 ($k + 1$) until it no longer reaches significance. Here, the method allows determining the probability densities of effect sizes assuming a single distribution (dotted line) versus two distinct distributions (solid lines). Even with a limited number of observations, true values are retrieved fairly accurately, as shown by the peaks of the two solid density distributions. (C) These estimates of effect size can then be compared to the ones extracted from a random-effects meta-analysis. Recall that in this example (shown in Figure 1B) a random-effects model gives a precise estimate of the wrong effect size; in contrast, estimating the mixture components first leads to two distinct estimates of effect sizes, including their respective true effect size. See the online article for the color version of this figure.

extracted for the two components were $\lambda = .83$ and $\lambda = .17$, respectively.

Posterior estimates suggested that two underlying distribution contributed to the overall distribution of effect sizes (Figure 3C). That is, brain training appears to have very limited influence on cognitive performance most of the time, but to be quite effective in other, more restricted instances. With the approach we presented in this article, we can extract estimates for the density of both distributions, allowing a more fine-grained analysis of the literature. We can then identify precisely the effect sizes that emerged from each distribution, thus providing fine-grained estimates of the discrepancies (Figure 3D). This result is interesting in and of itself, because it suggests that there are important differences that need to be further explored, whether related to protocols, demographics, or individual responses to training. One can then look for moderators or extraneous variables that were either initially coded but not analyzed, or, more plausibly, that were not thought to be of interest. This allows moving the discussion forward—rather than concluding that brain training has a limited effect, we can see that it appears to be ineffective most of the time, but possibly quite potent in rare instances.

Importantly, the problems we identify herein extend beyond the specific example we chose as an illustration. Based on a sample of 705 between-study estimates reported in meta-analyses published in the journal *Psychological Bulletin* between 1990 and 2013 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017), we found that heterogeneity (τ) between studies—within each meta-analysis—spread widely ($\sigma = 0.15$; see additional analyses in the online R code). In addition, the distribution was rather skewed, indicating that although a large number of meta-analyses show consistency between studies, a nontrivial number of them show important internal variability ($N = 27$ with $\tau > .5$). Regardless of the specific example, mixture model estimation provides a structured framework to probe subtle differences within a given population. In

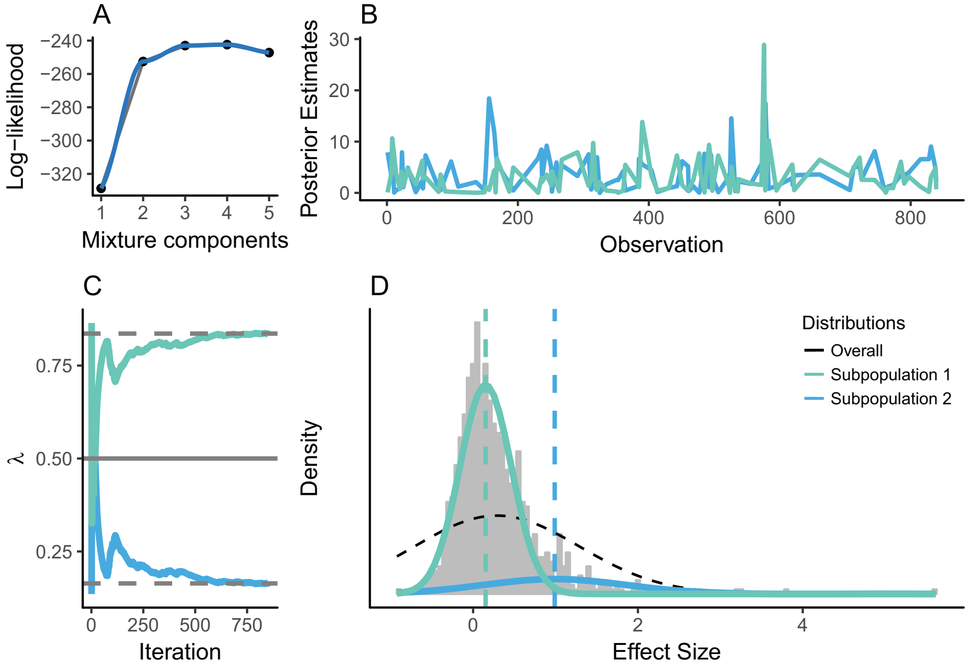

Figure 3. Mixture estimation for the brain training meta-analysis. (A) The expectation–maximization algorithm indicated that two components should be retained. The curve shows locally weighted smoothing. (B) Estimated log-likelihoods for the two components selected, over the entire set of observations. For readability, only a random, nonconsecutive subset (N = 100) of the whole vector of effect sizes (N = 854) is shown. (C) Estimated mixing weights (λ) based on posterior likelihoods from the vector of effect sizes. Here, the mixing weights (λ = .83 and λ = .17, solid lines, respectively) suggest that the data mostly comes from the first distribution, whereas the second is less well represented. (D) Probability densities of effect sizes. Densities are retrieved from the mixing weights estimated in C (solid lines). The mixture represented by the two solid lines is a better fit to the data than the single Gaussian (dotted line). The corresponding estimated means are shown with the vertical dashed lines. See the online article for the color version of this figure.

psychology, many researchers are interested in these questions, but oftentimes analyses are performed after sources of discrepancies have been mixed and processed. Mixture model estimation enables moving away from dichotomous thinking based on averages, in favor of a more constructive focus on the source of discrepancies.

Robustness of the Method

The approach we propose is effective in a number of cases and with a wide range of parameters. However, it also has limitations that need to be acknowledged. In some instances, assessments of either the number of components of a mixture or the probability of the mixture distribution may become unreliable. Limitations are inherent to any statistical tool, yet it is important to understand the range of parameters under which the method performs well, so as to anticipate potential pitfalls. We address these limitations in detail in the online R code (https://github.com/davidmoreau/MixtureModel), but present some of the main limitations hereafter.

One typical concern pertains to false alarms—specifically, how likely is a researcher to wrongly assume a mixture distribution when the best fit is a single distribution? To evaluate robustness formally, we generated effect size estimates from two underlying models. Model 1 was a single Gaussian, with a unimodal distribution. Model 2 was a mixture distributions generated from two Gaussian distributions. We aimed to determine how well the EM algorithm performs—how often do we correctly identify from which underlying distribution the data come?

Our analyses showed that false alarms remain low overall for plausible distribution parameters (see Table 1 for details). Generally, the algorithm shows improved performance as the distance between distributions increases. This makes intuitive sense: As the distribution modes spread out, overlap between these distributions increases, making it more difficult to identify the respective source of each data point. Related to this idea, we observed improved performance as σ^2, the variance of each distribution, decreases. This translates into less overlap between distributions, which for our purpose is akin to a greater distance between modes. In addition, we observed better performance overall for unbalanced λ values, that is, when distribution densities are unequal. This suggests that the method might be better suited to picking up subtleties in distribution properties that random-effects modeling alone. Specifically, we observed better performance overall but a rapid decrease in model performance as σ^2 increases for unbalanced λ values, whereas this decrease was less steep for more balanced λ values. The correlation between δ (residuals) and $l(\theta)$ (model fit) was $r = -.53$ (BF = 4.27, assuming a bivariate normal distribution and a uniform prior on ρ). The method performs well in a large

Table 1
Reliability Analyses for the Normal Mixture Model Estimates

σ^2 (Variance)	λ (Latent)	$\hat{\lambda}$ (Estimated)	δ (Residuals)	l (θ) (Fit)
.1	.1	.10	.00	−50.9
	.2	.14	.06	−50.4
	.3	.19	.11	−83.6
	.4	.41	.01	−70.6
	.5	.47	.03	−85.2
.2	.1	.10	.00	−72.1
	.2	.30	.10	−93.2
	.3	.47	.17	−98.8
	.4	.29	.11	−91.2
	.5	.26	.24	−103.2
.3	.1	.29	.19	−93.2
	.2	.30	.10	−95.3
	.3	.28	.02	−112.2
	.4	.18	.22	−99.6
	.5	.14	.36	−104.4
.4	.1	.20	.10	−110.9
	.2	.12	.08	−106.7
	.3	.17	.13	−110.9
	.4	.19	.21	−108.9
	.5	.14	.36	−106.8

Note. The table shows the results of a Monte Carlo simulation of mixture distributions (N = 100,000 observations per row) each generated from two normal Gaussian N (μ, σ^2) with fixed means (μ = 0; 1), but incremental variance (σ^2 = .1; .2, .3; .4) and mixing weights (λ = .1, .9, .2, .8; .3, .7; .4, .6; .5, .5). For each iteration, we estimated the mixing weights $\hat{\lambda}$ generated via the expectation–maximization algorithm. We provide indices of reliability with δ (residuals of the model) and log-likelihoods $l(\theta)$ (model fit).

number of cases, with a fairly wide range of parameters. Only multiple distributions that overlap extensively are difficult to tease apart. Direct estimates of convergence (log-likelihoods) are reported in the R code available with this article. In all cases presented, convergence checks indicate adequate performance, with only minor fluctuations in subsequent log-likelihoods (see Table 1).

Note that the algorithm we present in this article makes an estimate regarding the number of components to be retained, that is, how many distributions are conflated. Ultimately however, the decision to retain a simpler, unimodal model is to be determined by the researcher based on the question at hand; no single algorithm can replace educated assumptions about the plausibility of a distribution. Even so, the degree of belief, rather than mere dichotomous thinking, can be modeled, so as to reflect confidence in a model. When in doubt, that is, when the evidence for mixture is weak, a sensible decision is to default to a unimodal distribution (e.g., $N[\mu, \sigma^2]$). Importantly, when a mixture distribution is assumed, the influence of mixing weights on the model decreases as λ approaches 0 (see R code). This combination of algorithm-based mixture estimation and informed decision makes for a process that is fairly robust to overfitting.

Evaluating the robustness of mixture model estimation is also essential for a wide implementation. In the model we presented, we make particular assumptions about the shape of the underlying distribution—specifically, we assumed a normal distribution. We tested the extent to which deviations from normality impair the performance of the algorithm, and found that the method is fairly robust to violations of this assumption (for details, see R code). We also tested performance with non-Gaussian distributions such as t or Cauchy distributions. Although combinations of the former appeared to be well approximated by Gaussian mixtures, even as degrees of freedom approached one, the underlying parameters of a combination of the latter are often not well retrieved if one assumes normality. Importantly, nonparametric or semiparametric alternatives exist and can easily be implemented for all cases where the assumption of normality is violated. In the brain training example we discuss in this article, comparisons between values retrieved via parametric and semiparametric showed that densities from both methods were in agreement (see Figure 3 and online material). More generally, other alternatives exist—for example, one could assign a prior distribution to each parameter, for a fully Bayesian implementation of mixture modeling (e.g., Marin et al., 2005). The approach we presented here is by no means the only one designed for this kind of problems, but it is a method that has shown to perform well in various circumstances (Dempster, Laird, & Rubin, 1977).

Concluding Remarks

Through modeling and simulations, we have illustrated a potentially pervasive problem in the context of replication and meta-analysis: different underlying distributions conflated together in meta-analytic models. In a simple but potentially detrimental manifestation of this problem, some results in the literature may come from null effects (H_0), whereas others come from true positive effects (H_1). In more complex scenarios, this rationale can be extended to accommodate multimodal distributions, yet the inherent problems remain the same—typical models are sometimes too coarse to accurately aggregate a series of studies, and thus are at risk of misrepresenting overall trends and findings. Rather than solely relying on random-effect models—which can account for some but not all of the problems associated with mixture distributions—we have argued that meta-analytic methods should evaluate these discrepancies directly, by estimating the plausibility of mixture models and the corresponding mixing weights. This process allows for the identification of multimodality in distributions and refined estimates of effect sizes, paving the way for more accurate predictions and informed decisions based on all the evidence accumulated on a specific research question. In closing, we would like to leave the reader with a few related thoughts.

First, the discussion we hope to stir up in the context of meta-analysis in psychology has direct implications for replications. Although encouraging replication is a definite step forward, the degree to which combinations of studies are informative is contingent upon the use of appropriate aggregating methods. Specifically, and despite recent promising directions (e.g., Dreber et al., 2015; Earp & Trafimow, 2015), disagreements remain regarding how to best determine whether a replication is supportive or unsupportive of the original study (Simonsohn, 2015; Verhagen & Wagenmakers, 2014). There is currently no clear consensus concerning valid criteria for successful replications (SRs). Should it be statistical significance? Neighboring effect sizes? Subjective ratings? Or similar implications between studies regarding applied policies, such as health recommendations? For excellent discussions on this topic, see Brandt et al. (2014) and Makel, Plucker, and Hegarty (2012).

To make interpretations even more difficult, original studies and replications often yield a wide range of effect size estimates, varying from one extreme to another. In this context, inferences about true effect sizes can be particularly arduous. Part of the discrepancies in the interpretation of replications lies in false dichotomous thinking, with replication studies being framed as "successful" or "unsuccessful." Well-powered replications are thought to bring a sense of resolution, a final verdict on a research question. This is hardly ever the case. Besides being ill posed, this position does not allow for constructive or solvable answers. Rather, questions addressing the extent to which a replication study provides independent evidence for the presence of an effect, or for a given theory, are useful and help the field move forward. All evidence is of importance, and ironically, converting a fine-grained, continuous estimate into a binary characterization is akin to lowering statistical power—precisely what researchers are trying to avoid.

Promising solutions have been proposed to better model predictions in a replication context (Brandt et al., 2014; Simonsohn, 2015; Spence & Stanley, 2016; Verhagen & Wagenmakers, 2014), within a dynamic that allows finer assessments of the literature. Many of these solutions have emphasized the role of power analyses to design more robust replications (Lakens & Evers, 2014; Perugini, Gallucci, & Costantini, 2014). Some researchers have suggested heuristics to help guide researchers planning a replication, such as taking 2.5 times the sample size of the original study (Simonsohn, 2015), or simply maximizing power (Brandt et al., 2014). This current emphasis on power analysis is not coincidental—one of its appeals is its simplicity in providing a specific sample size based on effect size averages. In the case of one single distribution of effect size, increasing power is indeed an adequate way to increase accuracy.

As informative as these recommendations might be, they do not address a fundamental source of discrepancy in successive results—sampling from different underlying populations. If multiple distributions are allowed to contribute to a given effect, whether it is within a given study, across studies, or both, typical methods of aggregating data can yield spurious results (e.g., Speelman & McGann, 2013). Power is therefore part of the answer, but it is not sufficient in and of itself. Traditional methods for improving the strength of an experimental design typically do not take into account discrepancies in underlying distributions, or only partially (e.g., random effects models, Hedges & Vevea, 1998), and in this regard do not provide adequate safeguards against conceptual inaccuracies. These limitations need to be acknowledged, especially when results appear to be disparate.

With this in mind, we have also emphasized that inferring plurimodality in effect size distributions is not without pitfalls. When misapplied, it can lead to biased or erroneous assessments of a body of work, and provide misleading recommendations concerning the power of a replication. Although advances in computing and statistics allow refining probabilistic estimates via stochastic simulations (e.g., Markov chain Monte Carlo) complemented with advanced methods to cope with uncertainty (see e.g., Alfaro, Zoller, & Lutzoni, 2003; Liu & Chen, 1998; Rubinstein & Kroese, 2011), these techniques only refine assessments if probabilities can be accurately estimated in the first place. Perhaps most importantly, for the posterior probabilities to be most accurate, all studies should be published (van Assen, van Aert, Nuijten, & Wicherts, 2014). This gives additional weight to recent initiatives that either encourage preregistration of all studies (e.g., Open Science Framework, AsPredicted) or at least promote the use of data repositories. In practice, this allows determining the power of a study based on an accurate estimate for the probability of a planned test to detect an effect, rather than via distribution-agnostic power analyses.

Finally, one broader lesson here is that in some instances (and contrary to common claims), more research is not needed, because the amount of data available exceeds what is necessary to make accurate estimates. Arguably, this is especially true in the behavioral sciences, because numerous experiments are not theory driven, at least not in the sense of a unified theory to explain brain-behavior interactions. Rather than additional data, what is often lacking are better methods to aggregate prior studies, to increase the quality of meta-analyses.[4] In this dynamic, testing for mixture distributions has the potential to greatly refine estimates of effect sizes and improve interpretations of overall bodies of research—critical aspects for stronger, more accurate science.

[4] This point itself is contingent on adequate quality of the studies available (e.g., preregistration, full disclosure of all analyses and dependent variables, etc.).

References

Alfaro, M. E., Zoller, S., & Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution, 20,* 255–266. http://dx.doi.org/10.1093/molbev/msg028

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037. http://dx.doi.org/10.1126/science.aad9163

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review, 22,* 366–377.

Barnes, C. A. (2003). Long-term potentiation and the ageing brain. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 358,* 765–772. http://dx.doi.org/10.1098/rstb.2002.1244

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology, 59,* 132–150. http://dx.doi.org/10.1016/j.jmp.2013.12.002

Ben-Israel, A. (1966). A Newton–Raphson method for the solution of systems of equations. *Journal of Mathematical Analysis and Applications, 15,* 243–252. http://dx.doi.org/10.1016/0022-247X(66)90115-6

Benyamin, B., Pourcain, B., Davis, O. S., Davies, G., Hansell, N. K., Brion, M.-J., . . . the Wellcome Trust Case Control Consortium 2 (WTCCC2). (2014). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry, 19,* 253–258. http://dx.doi.org/10.1038/mp.2012.184

Bersoff, D. M. (1999). Motivated reasoning and unethical behavior. *Personality and Social Psychology Bulletin, 25,* 28–39. http://dx.doi.org/10.1177/0146167299025001003

Bohannon, J. (2015). Many psychology papers fail replication test. *Science, 349,* 910–911. http://dx.doi.org/10.1126/science.349.6251.910

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8,* 445–454. http://dx.doi.org/10.1177/1745691613491271

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217–224. http://dx.doi.org/10.1016/j.jesp.2013.10.005

Chase, W., & Simon, H. (1973). The mind's eye in chess. In W. Chase (Ed.), *Visual information processing* (pp. 215–281). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-170150-5.50011-1

Consonni, G., & Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association, 90,* 935–944. http://dx.doi.org/10.1080/01621459.1995.10476593

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., . . . Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry, 16,* 996–1005. http://dx.doi.org/10.1038/mp.2011.85

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,* 1–38.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361–376. http://dx.doi.org/10.1016/j.jml.2008.06.007

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology, 26,* 897–899. http://dx.doi.org/10.1038/nbt1406

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2016). Reevaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review, 23,* 306–316.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., . . . Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 15343–15347. http://dx.doi.org/10.1073/pnas.1516179112

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6,* 621. http://dx.doi.org/10.3389/fpsyg.2015.00621

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Evans, R., & Sedransk, J. (2001). Combining data from experiments that may be similar. *Biometrika, 88,* 643–656. http://dx.doi.org/10.1093/biomet/88.3.643

Friendly, M. (2017). *HistData: Data sets from the history of statistics and data visualization*. Retrieved from https://CRAN.R-project.org/package=HistData

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037. http://dx.doi.org/10.1126/science.aad7243

González-Castro, T. B., Tovilla-Zárate, C., Juárez-Rojop, I., Pool García, S., Genis, A., Nicolini, H., & López Narváez, L. (2013). Distribution of the Val108/158Met polymorphism of the COMT gene in healthy Mexican population. *Gene, 526,* 454–458. http://dx.doi.org/10.1016/j.gene.2013.05.068

Gronau, Q. F., Duizer, M., Bakker, M., & Wagenmakers, E.-J. (2017). Bayesian mixture modeling of significant p values: A meta-analytic method to estimate the degree of contamination from H_0. *Journal of Experimental Psychology: General, 146,* 1223–1233. http://dx.doi.org/10.1037/xge0000324

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24,* 2409–2419. http://dx.doi.org/10.1177/0956797613492984

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine, 20,* 1771–1782. http://dx.doi.org/10.1002/sim.791

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. http://dx.doi.org/10.1037/1082-989X.3.4.486

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

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician, 58,* 30–37. http://dx.doi.org/10.1198/0003130042836

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. http://dx.doi.org/10.1177/1745691612464056

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 6829–6833. http://dx.doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 10081–10086. http://dx.doi.org/10.1073/pnas.1103228108

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition, 42,* 464–480. http://dx.doi.org/10.3758/s13421-013-0364-z

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 19313–19317. http://dx.doi.org/10.1073/pnas.1313476110

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227–235. http://dx.doi.org/10.1037/1082-989X.1.3.227

Kamary, K., Mengersen, K., Robert, C. P., & Rousseau, J. (2014). Testing hypotheses via a mixture estimation model [preprint]. arXiv. Retrieved from https://arxiv.org/pdf/1412.2044.pdf page 11

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science, 25,* 2027–2037. http://dx.doi.org/10.1177/0956797614548725

Kelley, G. A., & Kelley, K. S. (2012). Statistical models for meta-analysis: A brief tutorial. *World Journal of Methodology, 2,* 27–32. http://dx.doi.org/10.5662/wjm.v2.i4.27

Klein, R., & Ratliff, K. (2014). Data from investigating variation in replicability: A "many labs" replication project. *Journalism, 2,* 13–16.

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9,* 278–292. http://dx.doi.org/10.1177/1745691614528520

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science, 26,* 1827–1832. http://dx.doi.org/10.1177/0956797615616374

Liu, J. S., & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association, 93,* 1032–1044. http://dx.doi.org/10.1080/01621459.1998.10473765

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7,* 537–542. http://dx.doi.org/10.1177/1745691612460688

Malec, D., & Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika, 79,* 593–601. http://dx.doi.org/10.1093/biomet/79.3.593

Marin, J.-M., Mengersen, K., & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In D. K. Dey & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 25, pp. 459–507). New York, NY: Elsevier. http://dx.doi.org/10.1016/S0169-7161(05)25016-2

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49,* 270–291. http://dx.doi.org/10.1037/a0028228

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer." *Perspectives on Psychological Science, 11,* 512–534. http://dx.doi.org/10.1177/1745691616635612

Moreau, D. (2014). Making sense of discrepancies in working memory training experiments: A Monte Carlo simulation. *Frontiers in Systems Neuroscience, 8,* 161. http://dx.doi.org/10.3389/fnsys.2014.00161

Moreau, D., Kirk, I. J., & Waldie, K. E. (2017). High-intensity training enhances executive function in children in a randomized, placebo-controlled trial. *eLife, 6,* e25062. http://dx.doi.org/10.7554/eLife.25062

Moreno, E., Vázquez-Polo, F. J., & Negrín, M. A. (2017). Bayesian meta-analysis: The role of the between-sample heterogeneity. *Statistical Methods in Medical Research*. Advance online publication. http://dx.doi.org/10.1177/0962280217709837

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

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217–243. http://dx.doi.org/10.1080/1047840X.2012.692215

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. http://dx.doi.org/10.1177/1745691612459058

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

Papadatou-Pastou, M., & Tomprou, D. M. (2015, September). Intelligence and handedness: Meta-analyses of studies on intellectually disabled, typically developing, and gifted individuals. *Neuroscience & Biobehavioral Reviews, 56,* 151–165. http://dx.doi.org/10.1016/j.neubiorev.2015.06.017

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319–332. http://dx.doi.org/10.1177/1745691614528519

Petryshen, T. L., Sabeti, P. C., Aldinger, K. A., Fry, B., Fan, J. B., Schaffner, S. F., . . . Sklar, P. (2010). Population genetic study of the brain-derived neurotrophic factor (BDNF) gene. *Molecular Psychiatry, 15,* 810–815. http://dx.doi.org/10.1038/mp.2009.24

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., . . . Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General, 142,* 359–379. http://dx.doi.org/10.1037/a0029082

Ridding, M. C., & Ziemann, U. (2010). Determinants of the induction of cortical plasticity by non-invasive brain stimulation in healthy subjects. *The Journal of Physiology, 588,* 2291–2304. http://dx.doi.org/10.1113/jphysiol.2010.190314

Rubinstein, R. Y., & Kroese, D. P. (2011). *Simulation and the Monte Carlo method*. Hoboken, NJ: Wiley.

Scarisbrick, D. J., Tweedy, J. R., & Kuslansky, G. (1987). Hand preference and performance effects on line bisection. *Neuropsychologia, 25,* 695–699. http://dx.doi.org/10.1016/0028-3932(87)90061-3

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551–566. http://dx.doi.org/10.1037/a0029487

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181. http://dx.doi.org/10.1037/0003-066X.47.10.1173

Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin, 72,* 214–228. http://dx.doi.org/10.1037/h0027880

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138,* 628–654. http://dx.doi.org/10.1037/a0027473

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21,* 3153–3159. http://dx.doi.org/10.1002/sim.1262

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17,* 103–186. http://dx.doi.org/10.1177/1529100616661983

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559–569. http://dx.doi.org/10.1177/0956797614567341

Somers, M., Shields, L. S., Boks, M. P., Kahn, R. S., & Sommer, I. E. (2015). Cognitive benefits of right-handedness: A meta-analysis. *Neuroscience and Biobehavioral Reviews, 51,* 48–63. http://dx.doi.org/10.1016/j.neubiorev.2015.01.003

Speelman, C. P., & McGann, M. (2013). How mean is the mean? *Frontiers in Psychology, 4,* 451.

Spence, J. R., & Stanley, D. J. (2016). Prediction interval: What to expect when you're expecting . . . A replication. *PLoS ONE, 11,* e0162874. http://dx.doi.org/10.1371/journal.pone.0162874

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., . . . Gabrieli, J. D. E. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS ONE, 8,* e63614. http://dx.doi.org/10.1371/journal.pone.0063614

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE, 9,* e84896. http://dx.doi.org/10.1371/journal.pone.0084896

van Erp, S., Verhagen, J., Grasman, R., & Wagenmakers, E. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990 to 2013. *Journal of Open Psychology Data, 5,* 4. http://dx.doi.org/10.5334/jopd.33

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143,* 1457–1475. http://dx.doi.org/10.1037/a0036731

Witte, A. V., Kürten, J., Jansen, S., Schirmacher, A., Brand, E., Sommer, J., & Flöel, A. (2012). Interaction of BDNF and COMT polymorphisms on paired-associative stimulation-induced cortical plasticity. *The Journal of Neuroscience, 32,* 4553–4561. http://dx.doi.org/10.1523/JNEUROSCI.6010-11.2012

Wright, L., & Hardie, S. M. (2015). Left-handers look before they leap: Handedness influences reactivity to novel Tower of Hanoi tasks. *Frontiers in Psychology, 6,* 58. http://dx.doi.org/10.3389/fpsyg.2015.00058

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 58,* 1–23. http://dx.doi.org/10.1177/1745691617693393

Appendix A

The General Mixture Model Framework

Mixture models are used to make inferences about properties of subpopulations contained within a larger population. Suppose we plot the distribution of heights across a population of individuals (data retrieved from Friendly, 2017). If we fit a density distribution to the histogram, we can see the general trend of the distribution (Figure A1A). In this example, it is rather obvious that the underlying distribution is bimodal—that is, it contains two different distributions, with two modes. Here, the explanation is straightforward: The distribution includes heights from men and women, conflated together. When plotted separately, with distinct model fits, the reason for bimodality appears clear (Figure A1B).

In this example, we know which observations are measurements from men and which are measurements from women. However, this is not always the case—oftentimes, subpopulations are not known, or at least not identified. Yet, we can estimate the shapes and modes of the two distributions, together with the probability of coming from either of these distributions for each data point. Importantly, the number of subpopulations is not limited to two, such that the overall framework can accommodate multiple sub-

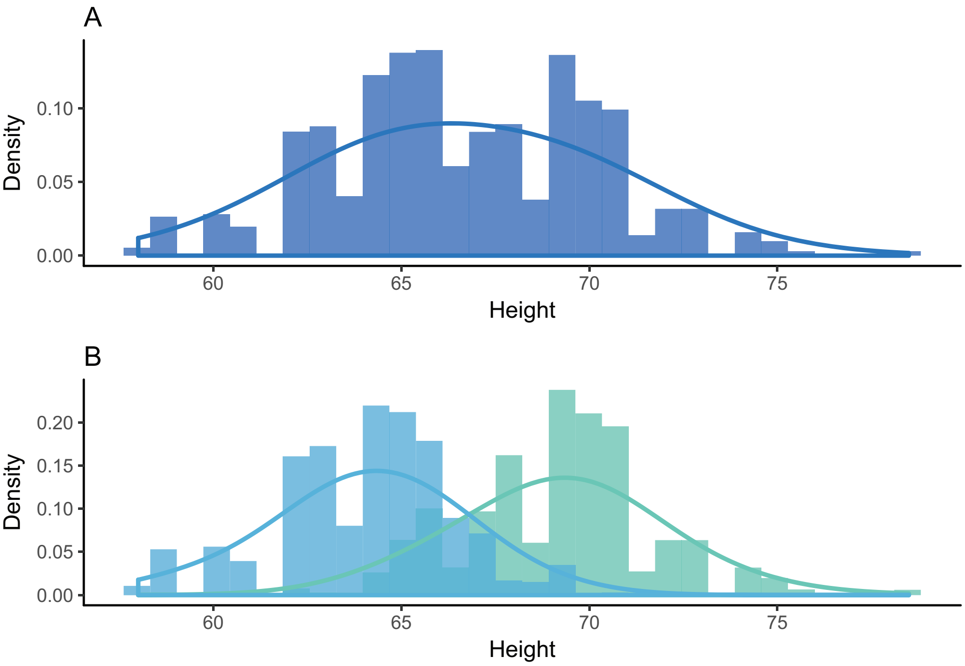

Figure A1. Distribution of heights in a population of men and women. (A) Overall, the distribution appears to be bimodal–the density line does not fit the data well. (B) When men and women are separated, the underlying trend is clearer: men are taller than women on average, although there is some overlap between the two subpopulations. Separate density lines are better fit to the data. See the online article for the color version of this figure.

(*Appendices continue*)

populations. In this context, the general mixture model framework is a way to mathematically represent subpopulations contained within a population. According to the framework, we can formally estimate densities of each value x_{i} in a vector as follows:

$$g_\theta(x_i) = \sum_{j=1}^{m} \lambda_j \phi_j(x_i), x_i \in R^r \tag{1}$$

where

$$\theta = (\lambda, \phi) = (\lambda_1, \ldots, \lambda_m, \phi 1, \ldots, \phi) \tag{2}$$

denotes the vector of parameters, with $\lambda_m \geq 0$, and ϕ_j drawn from a family of continuous multivariate distributions. In the special case we present in the article, that is, assuming normality of all distributions, the model parameter reduces to

$$\theta = (\lambda, (\mu_1, \sigma_1^2), \ldots, (\mu_m, \sigma_m^2)) \tag{3}$$

This is because the model parameter of a normal distribution is defined by a mean and variance. However, the same model (Equation 1) can accommodate distributions that are non-normal. So, how exactly can we estimate the probabilities of coming from a given distribution (i.e., subpopulation) for each data point? One of the possible solutions to this problem lies with the EM algorithm. We discuss it in more details in Appendix B.

Appendix B

Estimating Posterior Likelihoods With the Expectation–Maximization (EM) Algorithm

In this article, we use the EM algorithm to identify underlying distributions of effect sizes, and to compute the respective probabilities for each effect size to belong to a given distribution (for an accessible primer, see Do & Batzoglou, 2008). There are other ways to identify the nature of a distribution, such as the more general Majorize-Minimize/Minorize-Maximize algorithm (Hunter & Lange, 2004) or the widely used Newton–Raphson method (Ben-Israel, 1966); here, we chose to use the EM algorithm because it is fairly widespread and can accommodate numerous types of underlying distributions. This versatility allows adaptations to a wide range of problems, with varying distribution properties.

Specifically, the EM algorithm provides a way to find maximum likelihood solutions for models whose variables are unobserved, or latent. In this context, the EM algorithm maximizes the operator:

$$Q(\theta \mid \theta^{(t)}) = E[\log h_\theta(C) \mid x, \theta^{(t)}] \tag{4}$$

where $\theta^{(t)}$ is the value at iteration t, and the expectation step concerns $k_\theta(c \mid x)$ for the value $\theta^{(t)}$ of the parameter. The first step (E) is to compute Equation 4; the second step (M) is to set

$$\theta^{(t+1)} = argmax_{\theta \in \Phi} Q(\theta \mid \theta^{(t)}) \tag{5}$$

with Φ denoting the parameter space of θ. Based on the EM iterative algorithm, if a unimodal distribution is more likely, or if there is insufficient evidence for a mixture distribution, the assumption of a single underlying distribution is retained. In this case, no further estimate is required. If there is evidence for a mixture distribution (as defined by the likelihoods), we generate posterior probabilities of belonging to a given distribution, for each data point. In the specific case we have described herein, this represents the probability of belonging to distribution A or B, where $\theta = (\lambda, (\mu_A, \sigma_A^2), (\mu_B, \sigma_B^2))$. Note that at that stage the greater likelihood of a bimodal model (i.e., mixture) compared to a unimodal model, given the data, has already been established. Finally, convergence, the point of equilibrium in the EM algorithm, is given by the log-likelihood:

$$lnP(X \mid \mu, \sigma, \alpha = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \alpha_k N(x_n \mid \mu_k, \sigma_k^2 \tag{6}$$

Mixture model estimation typically uses the log-likelihood, rather than the likelihood. This is to avoid dealing with extremely small values, which can introduce unnecessary computational problems. The larger the log-likelihood, the better the model parameters fit the data. The final result of the EM algorithm gives a probability distribution of the estimated latent variables, in conjunction with point estimates for θ (either a maximum likelihood estimate or a posterior mode). Note that in a fully Bayesian version of the approach, one can estimate a probability distribution over both θ (treated as a latent variable) and the estimated latent variables.

In cases where the assumption of normality is violated, nonparametric or semiparametric alternatives that do not assume that parameter values of ϕ_j in the original algorithm are drawn from a family of continuous multivariate distributions can be implemented. Assuming that all observations in x_i are conditionally independent, the model can be rewritten as

$$g_\theta(x_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}) \tag{7}$$

where f_{jk} denotes a univariate density function that is not assumed to come from a family of densities defined by a finite-dimensional parameter vector, and densities are estimated via nonparametric density methods. This more general definition of a mixture model allows applications to a larger set of problems.

Received April 2, 2018
Revision received July 18, 2018
Accepted August 4, 2018 ■