aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Shifting Minds: A Quantitative Reappraisal of Cognitive-Intervention Research

Perspectives on Psychological Science
2021, Vol. 16(1) 148–160

DOI: 10.1177/1745691620950696
www.psychologicalscience.org/PPS
SAGE

David Moreau
School of Psychology & Centre for Brain Research, University of Auckland

Abstract

Recent popular areas of research in psychology suggest that behavioral interventions can have profound effects on our cognitive abilities. In particular, the study of brain training, video gaming, mindset, and stereotype threat all include claims that low-cost, noninvasive manipulations of the environment can greatly affect individual performance. Here, I provide a quantitative reappraisal of this literature, focusing on recent meta-analytic findings. Specifically, I show that effect-size distributions in the four aforementioned areas are best modeled by multiple rather than single latent distributions, suggesting important discrepancies in the effect sizes reported. I further demonstrate that these multimodal characteristics are not typical within the broader field of psychology, using 107 meta-analyses published in three top-tier journals as a comparison. The effect-size distributions observed in cognitive-intervention research therefore appear to be uncommon, and their characteristics are largely unexplained by current theoretical frameworks of cognitive improvement. Before the source of these discrepancies is better understood, the current study calls for constructive skepticism in evaluating claims of cognitive improvement after behavioral interventions and for caution when this line of research influences large-scale policies.

Keywords

environment, cognitive improvements, intelligence, brain plasticity, genetics, meta-analysis, mixture modeling

Psychological science influences all spheres of society. We are constantly bombarded by psychological findings when reading popular news outlets, watching television shows, or browsing the Internet. Psychology informs the algorithms that determine the friends we make, the movies we watch, and the music we listen to. Research in the field also has profound implications—it influences school curricula (U.S. Department of Education, 2015), plays an active role in judiciary decisions (Suggs, 1979), and shapes societal policies, affecting the lives of millions (Halpern, 2014).

The widespread appeal of psychological science is especially palpable when research purports large changes with reasonably simple manipulations, embedded in a framework that appears to reduce individual differences or inequalities (Moreau, Macnamara, & Hambrick, 2019). A few minutes spent feet apart, hands on hips, and chin upward can help us land our dream job at the next interview (Carney, Cuddy, & Yap, 2010); consuming sugary lemonade restores our ability to exert self-control (Gailliot et al., 2007). Other manipulations are thought to elicit a range of remarkable (and potentially lasting) effects; for example, being exposed to structured events makes us more willing to pursue personal goals (Kay, Laurin, Fitzsimons, & Landau, 2014), whereas holding a pen between our teeth lets us see life in a more cheerful way (Strack, Martin, & Stepper, 1988). The implicit demands for sensational, headline-grabbing findings feed back into research programs, dictating the pace of science as well as the type of research being incentivized (e.g., Lilienfeld, 2017).

In recent years, the notion that behavior or abilities can be easily influenced via experimental manipulations has been called into question. Large, well-powered studies have failed to replicate all of the effects discussed above, from power posing (Garrison, Tang, & Schmeichel, 2016) to ego depletion (Hagger et al., 2016), structure seeking (Klein et al., 2018), and facial feedback (Acosta et al., 2016), casting serious doubt on the malleability of behavior, at the very least following these specific manipulations. One could argue, however, that the

Corresponding Author:
David Moreau, School of Psychology & Centre for Brain Research, University of Auckland
E-mail: d.moreau@auckland.ac.nz

ramifications of these research areas are limited and therefore that failures to replicate such findings have little impact on individuals and society.

This argument hardly holds for a number of related areas with well-established, direct applications, often at the institutional level. For example, several research programs have flourished on the basis of the notion that cognitive abilities can be substantially influenced by relatively short interventions. In particular, four areas of research have generated great interest—brain training, video gaming, mindset, and stereotype threat. These areas are all aimed at improving cognitive performance either by targeting cognitive abilities directly (brain training, video gaming) or by focusing on barriers thought to impede cognitive performance (mindset, stereotype threat).

Beyond their widespread popularity, the goal of each of these areas of research differs from that of other means of cognitive improvement that have come under scrutiny in recent years, such as bilingualism (Lehtonen et al., 2018), chess and music training (Sala & Gobet, 2017), or physical exercise (Diamond & Ling, 2019). Although the latter bring about a number of benefits (i.e., mastery of another language, the game of chess, or a musical instrument; keeping fit and healthy), irrespective of the evidence for cognitive gains, there are important opportunity costs associated with brain-training, video-gaming, mindset, and stereotype-threat interventions, and individual and collective decisions rest primarily on the reliability of scientific claims of improvement. For this reason, it is imperative to better understand the underlying mechanisms responsible for the observed mixed evidence. These four areas of research thus represent the focus of this article and are discussed in more detail hereafter.

Popular interventions for improving cognitive performance

Although the research paradigms of brain-training, video-gaming, mindset, and stereotype-threat research share many features, they also include important differences in the way they propose to address limitations in cognitive performance. Brain-training programs are designed to directly influence cognitive abilities via targeted regimens; forms of training are either unimodal, in that they focus on a single cognitive ability (e.g., working memory training; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008), or multimodal, including "brain exercises" targeting a range of abilities (for a review, see Simons et al., 2016). Brain-training proponents argue that practicing these games or exercises can have a profound impact beyond the trained tasks, leading to generalized cognitive improvements that affect multiple domains, most notably academic (Klingberg et al., 2005; Loosli, Buschkuehl, Perrig, & Jaeggi, 2012) and professional (Adler et al., 2015).

Whereas the purpose of brain-training regimens is to elicit cognitive improvements, video gaming is a leisure activity that has been harnessed for cognitive gain in recent years. Therefore, video-gaming regimens intended to train cognition capitalize on content that has the primary purpose of entertaining. Studies have reported enhanced performance in a range of abilities, including visual processing (Green & Bavelier, 2003, 2007), attention (Belchior et al., 2013), spatial ability (Goldstein et al., 1997; Okagaki & Frensch, 1994), and executive function (Basak, Boot, Voss, & Kramer, 2008; Green, Sugarman, Medford, Klobusicky, & Bavelier, 2012). Given the central role of these abilities in many aspects of everyday life, it has been hypothesized that video-gaming improvements can generalize to a variety of real-world domains (Bavelier, Green, Pouget, & Schrater, 2012).

Both brain-training and video-gaming regimens typically span weeks or months, yet some interventions are much shorter. In particular, interventions targeting beliefs about ability, rather than abilities themselves, either in the form of mindsets or stereotypes, have been reported to elicit improvements in a single or a couple of sessions. Mindset proponents argue that holding a malleable view of intelligence and other cognitive aptitudes (*growth* mindset) is associated with a range of positive outcomes, whereas holding a stable view of human aptitudes (*fixed* mindset) impedes learning, progress, and improvements in a variety of areas, including schools, businesses, and sports (Dweck, 2006; Yeager et al., 2019). In addition, beliefs about the malleability of cognitive aptitudes appear to *themselves* be malleable (Blackwell, Trzesniewski, & Dweck, 2007; Paunesku et al., 2015); short interventions that promote a malleable view of aptitudes have a positive impact on cognitive performance, whereas the converse is true when feedback promotes a fixed mindset (Yeager & Dweck, 2012).

Likewise, research on the topic of stereotype threat suggests that cognitive performance is remarkably susceptible to one's belief about their own group performance—if primed with a reminder that they belong to a particular group known to typically perform poorly on a test or task, individuals' performance will tend to decrease. In this context, subtle manipulations of individual beliefs about their own group or statements suggesting that their group typically performs better or as well as others can sometimes completely erase preexisting differences (e.g., Steele & Aronson, 1995).

Heterogeneity in cognitive-intervention research

All of these areas of research share a common rationale—relatively brief interventions, typically ranging from a single session to several months, can profoundly affect individual performance on a range of tests, including cognitive tasks. Yet beyond their common rationale, these four areas of research also share another feature: They have all been questioned by recent evidence, either with failed replications or large meta-analyses (Melby-Lervåg, Redick, & Hulme, 2016; Sala, Tatlidil, & Gobet, 2018; Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018; Stoet & Geary, 2012). Although these contradictory findings might not question the validity of the research areas themselves, they point out important limitations in our understanding of the hypothesized underlying mechanisms.

One point that has been relatively ignored but deserves further investigation is heterogeneity and its implications. Meta-analyses in cognitive-intervention research often include a wide range of effect sizes, from null to very large, yet few fall along that continuum. In this context, common statements based on average effect sizes can be misleading, as they fail to capture variability across effect sizes and may not be representative of prospective outcomes. This is especially problematic for intervention research, given that this line of work is inherently applied—if the main outcome observed at the meta-analytic level is not a plausible outcome at the level of an individual study, let alone an individual subject, correct interpretation is challenging, and informed decisions can be difficult.

Many measures have been developed to quantify heterogeneity in meta-analyses; for example, Cochran's Q, I^2, and T^2 all provide some estimate of the variability or inconsistencies across studies (Deeks, Higgins, & Altman, 2011; Higgins & Thompson, 2002), often compared with what could be expected by chance alone. As informative as these measures are, however, they do not provide any information about the latent distributions of effect sizes—all assume single underlying distributions.[1] Yet this assumption is one that requires substantiation, as central-tendency measures (mean, median, and mode) can be greatly misleading when multiple distributions contribute to a meta-analysis.

Recent methodological developments have made it possible to further characterize heterogeneity in meta-analyses beyond traditional measures. For example, mixture modeling has been proposed as a promising framework for identifying and modeling multiple subpopulations of effect sizes (Moreau & Corballis, 2019) when the central-limit theorem does not apply. The overall idea is straightforward: Mixture modeling enables probabilistic inferences about the presence of multiple subpopulations and allows estimating the corresponding "mixing weights"—that is, the respective proportions of studies that are thought to belong to each subpopulation. For example, the mixture of two distributions A and B with mixing weights of .6 and .4 implies that 60% of the effect sizes are thought to belong to subpopulation A and 40% to subpopulation B. One can then estimate the mean of each subpopulation, often providing a much more accurate estimate of the overall effect sizes in a research area.[2]

The importance of estimating mixture distributions is often illustrated with the example of height: Although the mean height in the United States is 168.8 cm (Fryar, Gu, & Ogden, 2012), this value relates to an overall distribution that conflates two rather distinct subpopulations: men and women. In this context, referring to separate measures of central tendency for the male and female subpopulations is much more informative (e.g., mean height for men is 175.9 cm and for women, 162.1 cm) and helps guide decisions across a range of domains. In psychology, failures to account for mixture distributions can even be more consequential, particularly in cases in which the overall mean is not a possible outcome for a single study or a single individual (Moreau & Corballis, 2019). Mixture modeling thus provides a flexible and powerful framework for exploring heterogeneity in meta-analyses, especially when multiple, hidden subpopulations are thought to contribute to an overall distribution of effect sizes.

Current study

The rationale for this study was threefold. First, recent work in our group has underlined that a number of research areas in psychology appear to overemphasize the role of the environment on human performance, including in the context of cognitive interventions (Moreau et al., 2019). In addition, brain training, video gaming, mindset, and stereotype threat are research areas that have all gained traction recently, and there are large-scale efforts to replicate early findings or to meta-analyze results (Forscher et al., 2019; Melby-Lervåg et al., 2016; Sala et al., 2018; Sisk et al., 2018; Stoet & Geary, 2012), which motivated inclusion in the current study. Second, recent failures to replicate early findings in cognitive-intervention research, together with substantial heterogeneity in effect sizes, suggest that current theoretical frameworks of cognitive improvements are weak, thus hindering fine predictions at the individual level. Such lack of predictive power suggests that a closer look at the literature could be beneficial. Finally, mixture modeling has been established as an important tool for providing further insight into existing data (Gronau, Duizer, Bakker, & Wagenmakers, 2017; Nord, Valton, Wood, & Roiser, 2017), including in

the context of meta-analyses (Moreau & Corballis, 2019), with the potential to uncover new, promising directions for future research.

In line with this rationale, I aimed to systematically and quantitatively investigate how cognitive interventions compare to other areas of research. Using a data-driven, preregistered method based on mixture modeling and including 111 meta-analyses, I tested the hypothesis that cognitive-intervention research presents specific characteristics that differ from studies in the broader field of psychology. Specifically, I examined whether effect sizes were better characterized by a single- or multicomponent distribution for each meta-analysis, following the predefined criteria outlined hereafter. The procedure was designed to determine whether individual studies within a meta-analysis are comparable, stemming from a common, coherent theoretical framework, or fundamentally discrepant so as to suggest hidden moderators that warrant further investigation.

Method

The present study was preregistered on OSF on April 6, 2019. The preregistration describes the research rationale, inclusion and exclusion criteria, modalities for data extraction, analyses, and diagnostics. All scripts and data have been made available at https://osf.io/ce9vr/.

Inclusion criteria and data extraction

Meta-analyses in the field of brain training, video gaming, mindset, and stereotype threat (cognitive-intervention studies) and meta-analyses in the broader field of psychology (control studies) were included. Inclusion criteria and data extraction were adapted for each group; the specific procedure is described hereafter.

Cognitive-intervention meta-analyses. Meta-analytic data were obtained for each of the four fields of research considered on the basis of the following criteria: (a) The article had to have been published in the past 5 years; (b) data from the article had to be either available openly on a repository (e.g., Open Science Framework, GitHub) or available on request (for further details, see Table S1 in the Supplemental Material available online); and (c) where multiple publications were eligible, articles that were more general (e.g., adult vs. older adult populations) and that were more comprehensive (as defined by the total number of effect sizes included in the final meta-analytic sample) were selected. In accordance with the above criteria, data were extracted from the following publications: Melby-Lervåg et al. (2016; brain training); Sala et al. (2018; video gaming); Sisk et al. (2018; mindset); and Lamont, Swift, and Abrams (2015; stereotype threat).

Control meta-analyses. A larger search was performed for meta-analyses across subfields to establish an overall base rate for multicomponent distributions in psychology. Specifically, three of the major outlets publishing meta-analyses in psychology (*Psychological Science, Perspectives on Psychological Science*, and *Psychological Bulletin)* were surveyed on October 23, 2019, for studies published in the past 5 years (2015–2019). Meta-analyses were selected on the basis of criteria (a) and (b) discussed above, and criterion (c) discussed above was adapted to include all relevant publications. This search resulted in 247 references across the American Psychological Association's PsycArticles database and the Association for Psychological Science's journal website (https://journals.sagepub.com/aps). These references included a number of articles that were not relevant to the current study, such as editorials ($n = 2$), commentaries ($n = 9$), replies ($n = 4$), reviews ($n = 5$), corrections ($n = 1$), replications ($n = 1$), repeats ($n = 18$), articles that contained the relevant keywords but were not original meta-analyses ($n = 51$), articles for which effect sizes could not be obtained ($n = 38$) or that included serious formatting issues ($n = 8$), and articles already included in the sample of cognitive-intervention studies ($n = 3$). A total of 107 studies met all inclusion criteria (for a full list that includes intervention and control studies, see Table S1 in the Supplemental Material). A breakdown of the inclusion process is described in a Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols flow diagram (Fig. 1), together with saved search records (online repository at https://osf.io/ce9vr/) for reproducibility.

Analyses

A method previously described in Moreau and Corballis (2019) was adapted for the purpose of this article. The method is based on mixture modeling, a framework that allows probabilistic assessments of the presence of subpopulations within an overall population, together with their estimation. In the context of meta-analyses, this approach has shown promise in the detection of different clusters (i.e., components) in a distribution of effect sizes (Moreau & Corballis, 2019). The framework follows from Gaussian mixture models (for a primer, see Appendix A in Moreau & Corballis, 2019).

The implementation detailed in Moreau and Corballis (2019) and in this article uses the expectation-maximization (EM) algorithm to identify underlying distributions of effect sizes and to compute the respective probabilities that each effect size would belong to a given distribution (for an accessible primer, see Do & Batzoglou, 2008). The EM algorithm was run until convergence (the point of equilibrium in the algorithm), defined as changes in log-likelihood values smaller than 10^{-2} (ε).

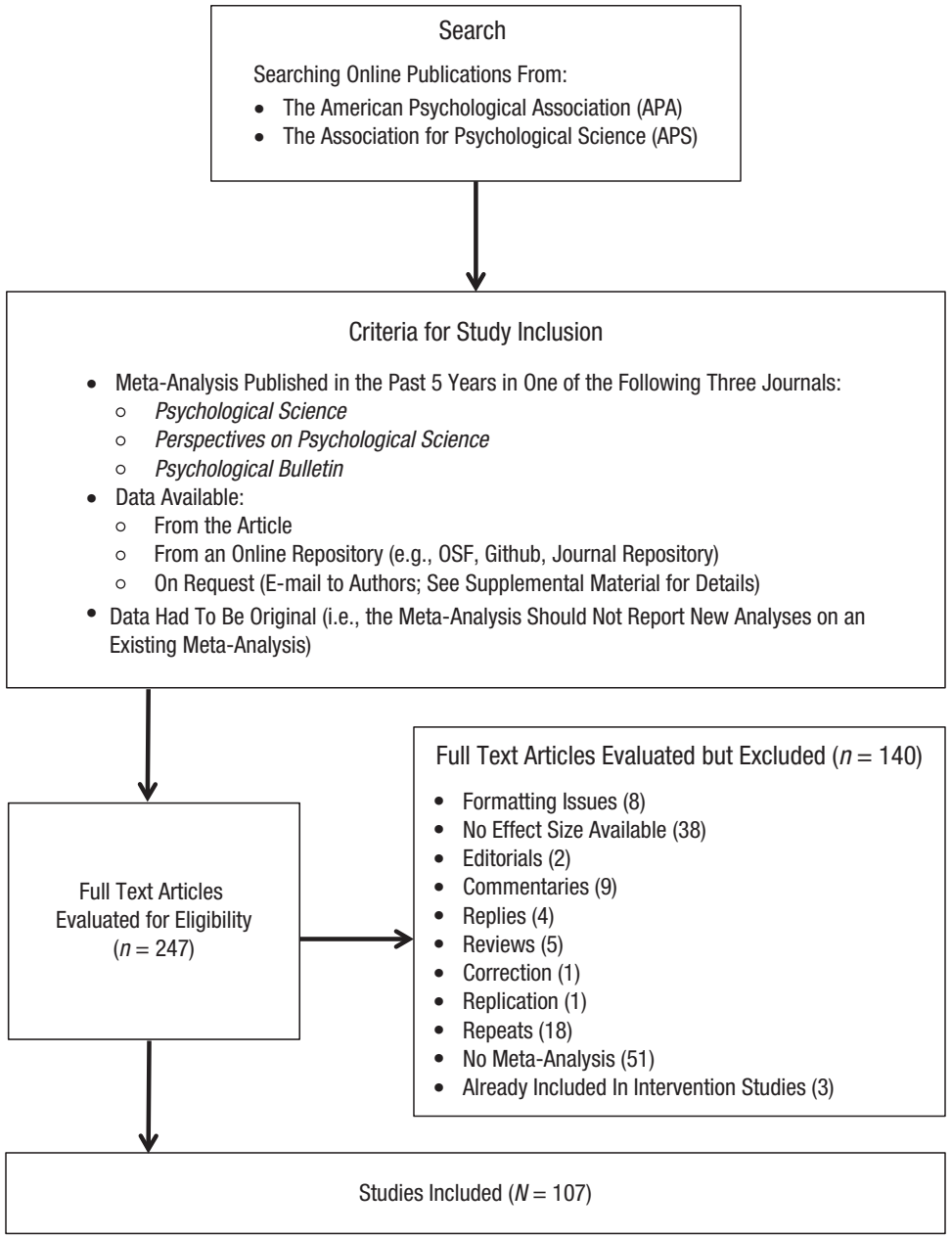

Fig. 1. Flow diagram for the control studies. These meta-analyses were compared with the cognitive-intervention meta-analyses (brain training, video gaming, mindset, and stereotype threat).

For each meta-analysis, the following metrics were estimated: (a) an index of confidence in a multicomponent model (log-likelihoods; higher values indicate better model fit), (b) posterior estimates, (c) the estimated mixing weights (λ, or proportion of effect sizes belonging to a given subpopulation), and (d) the corresponding (estimated) density distributions of effect sizes.

Robustness checks

Multicomponent solutions inferred from log-likelihoods were confirmed with the Bayesian information criterion (BIC). The latter penalizes the model according to the total number of parameters and therefore favors simpler models (everything else being equal). Because log-likelihoods do not penalize model complexity, solutions based purely on these estimates might be detrimental to model parsimony. The BIC allows balancing model fit (indexed by log-likelihoods) against the total number of model parameters—the larger the BIC, the stronger the evidence for the model and number of clusters.[3] Finally, robustness was further assessed with semiparametric and nonparametric versions of the EM algorithm to check for consistency across a range of

Table 1. Mixture-Model Fit for All Cognitive-Intervention Meta-Analyses

Research area	*N* (ES)	Log-likelihood	BIC	λ	μ
Brain training	854	−321.23	−1,130.37	.84/.16	0.16/1.01
Video gaming	359	−160.13	−591.59	.19/.81	−0.06/0.10
Mindset	43	−17.88	−64.46	.40/.60	0.08/0.18
Stereotype threat	82	−31.28	−147.58	.58/.42	−0.06/0.62

Note: The total number of effect sizes in the meta-analysis, parametric mixture-model fits for a two-component solution (log-likelihood and BIC), mixing weights (λ), and distribution means (μ) are shown for each research area. BIC = Bayesian information criterion; ES = effect size.

assumptions, or lack thereof—for details, see the code for the R software environment (Version 3.6.0; R Core Team, 2019).

Results

Two sets of complementary findings are reported hereafter. In the first section, I present empirical results based on the data extracted from the meta-analyses (intervention and control studies). Because probabilistic inference of latent distributions is inevitably prone to error, in the second section I report the results of simulations intended to gauge the validity and reliability of the method used in this article.

Meta-analyses

In the current implementation, multicomponent solutions were restricted to two components (bimodal distributions) to favor parsimony because estimating the precise number of clusters has been shown to increase uncertainty (e.g., McLachlan & Peel, 2000) and was not the primary purpose of this study. Results indicated a multicomponent mixture solution for all cognitive-intervention meta-analyses (see Table 1 and Fig. 2) based on log-likelihood and the BIC. Both methods were consistent, providing strong evidence for the presence of multiple subpopulations in all cases. When moderator analyses existed in the original meta-analyses, moderators were checked against the clusters identified from the mixture modeling algorithm to determine whether they matched. However, latent distributions tapped variation that could not be adequately explained by known moderators. Results held with fewer (semiparametric mixture) or no parametric assumptions, indicating that the observed results did not hinge on specific assumptions about underlying distributions; on the contrary, results were reproduced with a wide range of reasonable assumptions.

Control meta-analyses showed a different pattern. Of the effect-size distributions included in the 107 control studies, only 38% showed evidence for a multicomponent solution (see Fig. 3). Moreover, the number of meta-analyses exhibiting multiple subpopulations was much lower when assumptions of normality were relaxed, either with semiparametric or nonparametric implementations (for details, see R code). Together, these results indicate that the pattern observed for the four domains of interest does not generalize to psychological findings more broadly; rather, it appears to be specific to a few subfields of study. Exploratory analyses per subfield within psychology (cognitive, social, clinical) showed that the number of meta-analyses best characterized by multiple components was higher for clinical psychology (44%) and lower for social psychology (34%), cognitive psychology falling in between (38%). Because the four areas of research that are the focus of this article fall either under the cognitive or under the social category, it could be argued that these two subsets represent more adequate controls than psychology as a whole. Regardless, evidence was equally robust when using meta-analyses in the field of cognitive psychology as comparison and slightly stronger when contrasting with social-psychology studies.

Simulations

A set of simulations was run to confirm that the algorithm performed adequately in the specific context of the current study (for R code and details, see https://osf.io/ce9vr/). For each of the 107 control meta-analyses, the number of effect sizes, the mean, and standard deviation were extracted from the observed data to generate random Gaussian distributions with matching numbers of observations and parameters (mean and standard deviation) between empirical and simulated data. This procedure yielded 107 simulated distributions with the same parameters as the empirical data, each with a clear latent Gaussian distribution. Given built-in univariate assumptions (i.e., all distributions were normal), it was expected that the mixture model would favor single-component solutions most of the time, except for the occasional misclassification. Results showed that single-component solutions were preferred

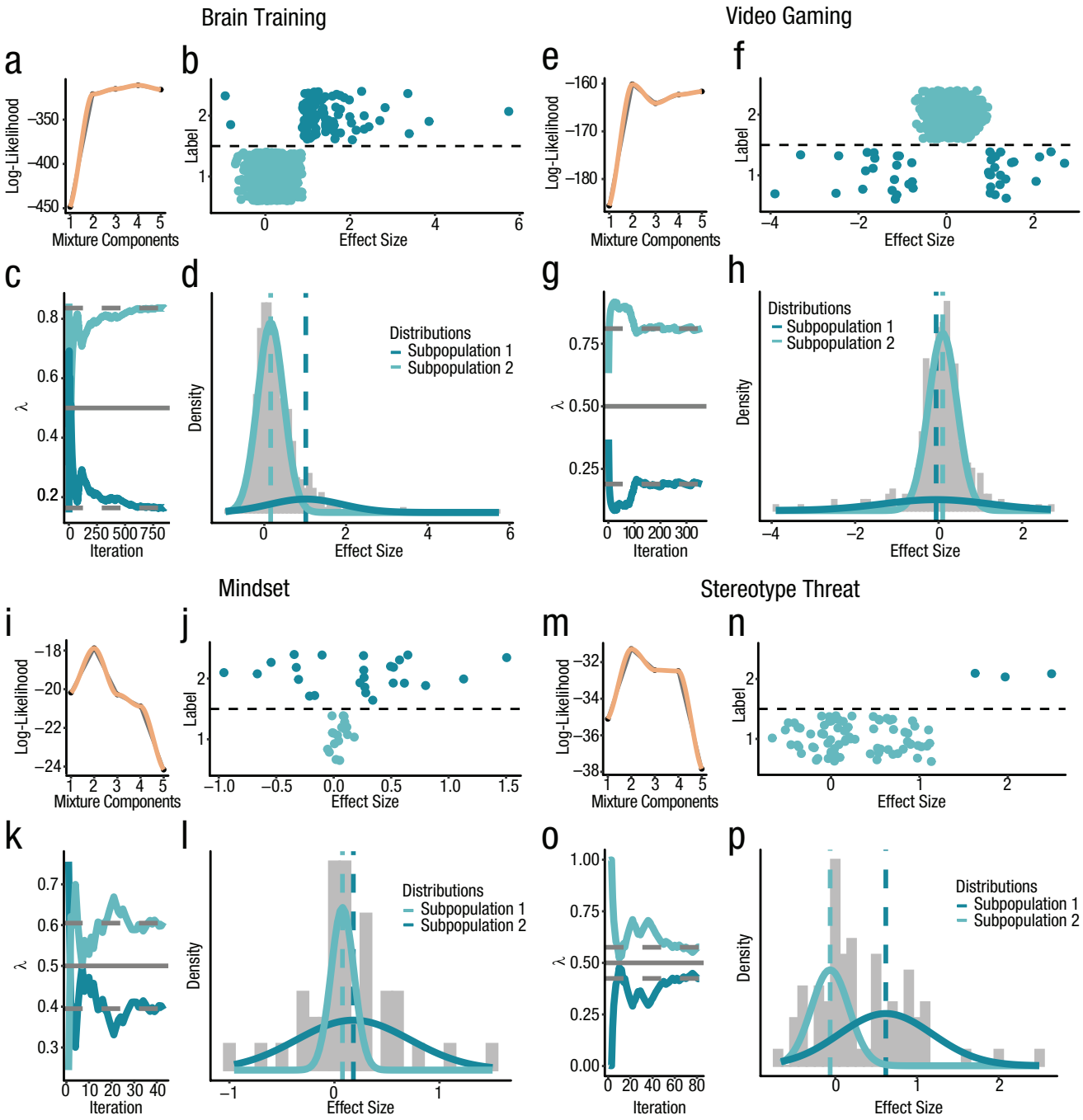

Fig. 2. Mixture estimation for the brain-training, video-gaming, mindset, and stereotype-threat meta-analyses. Log-likelihoods (a, e, i, m) are shown as a function of the number of components (i.e., clusters) in the model. Higher values indicate better model fits. A multicomponent solution is favored in all cases. Two-component clustering labels (b, f, j, n) are shown for all effect sizes within each of the four domains. In all four cases, the data suggest at least two underlying distributions, which are indicated in light teal and dark teal. Estimated mixing weights (λ) based on posterior likelihoods (c, g, k, o) from meta-analytic effect sizes are also shown. The mixing weights present the relative proportions of Subpopulations 1 and 2 for each meta-analysis. Histograms (d, h, l, p) show the meta-analytic effect sizes and estimated probability-density distributions inferred from the mixing weights (λ). The estimated means (μ) for each density distribution are indicated by vertical dashed lines (one for each subpopulation).

98% of the time (2% misclassification; i.e., false positives), indicating that the algorithm performed extremely well when the underlying distribution was a single Gaussian.

Likewise, the algorithm was evaluated when latent bimodal distributions were simulated from control study parameters. In this bimodal scenario, multicomponent solutions were preferred 76% of the time (24%

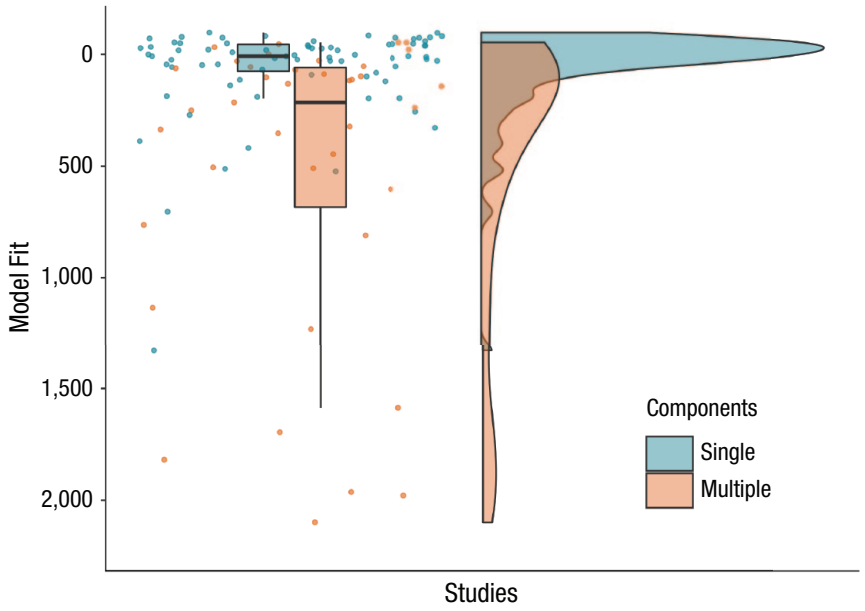

Fig. 3. Mixture-model fit for control meta-analyses. The plot shows individual data points, box plots, and density distributions for each meta-analysis within the broad field of psychology broken down by the number of estimated components (single vs. multiple). Meta-analyses that were classified as unimodal appear in green; those classified as multimodal are shown in light orange.

misclassification; i.e., false negatives); that is, the right decision was made three out of four times. Note that the asymmetry between accuracy for single versus multicomponent characterizations (98% vs. 76%) is by design: The algorithm implemented in this study was biased toward univariate assumptions *unless* there was enough evidence for multiple underlying components. The results reported were robust to departures from typical distributional assumptions (i.e., Gaussian), and semiparametric and nonparametric implementations yielded similar performance. The R code provided with this article allows active exploration with varying parameters and assumptions.

Discussion

The current study aimed to provide a quantitative reappraisal of the literature in the domains of brain training, video gaming, mindset, and stereotype threat using four recent meta-analyses and a total of 1,338 effect sizes. The question is whether the distributions of effect sizes in cognitive-intervention research show multimodal properties. That is, were effect sizes clustered, or did they tend to follow a single distribution? In all four domains, there was evidence for multicomponent solutions, suggesting that more than one population of effect sizes contributed to each distribution. This finding contrasts with comparative analyses in the broader field of psychology, as evidenced by 107 meta-analyses published between 2015 and 2019 in three major outlets. In this control subset, only about a third of meta-analyses showed reasonable evidence for multimodality in their samples of effect sizes. There was only minor variation across cognitive, social, and clinical subfields, without evident implications for the current study. Together, these findings suggest that cognitive-intervention research shows characteristics that are not typical of the field of psychology as a whole or of relevant subfields within psychology.

What does multimodality mean in the context of cognitive interventions?

This study, the first to systematically model and characterize latent distributions of effect sizes in the context of cognitive-intervention research, provides novel information that supports recent advances in our understanding of cognitive malleability (Moreau et al., 2019; Sala & Gobet, 2017). This quantitative reappraisal has a number of implications for our understanding of cognitive improvement via interventions; most importantly, it suggests that even when inferred from well-conducted, comprehensive meta-analyses, claims based on central-tendency measures such as mean effect size can be misleading and may not provide a solid basis for decisions or policies. In the context of intervention research,

this is especially problematic, as it typically leads to conclusions that are not representative of expected outcomes. For example, generic claims about small but nonnull effects for a given intervention, if based on mixtures of distributions, may convey little information with respect to potential applications. At the very least, this possibility should be factored into the decision process when seeking to implement large-scale interventions.

The approach presented in this article complements traditional measures of heterogeneity typically reported in meta-analyses. Although these measures are helpful in documenting the spread of effect sizes across the mean or median, they do not provide any information about the hypothesized source and characteristics of this variation beyond those explained by moderator analyses. In many instances, however, we might not recognize that a particular variable matters; there might be specific characteristics of samples, investigators, or protocols that are not typically being documented because we may not realize that they represent important modulators of the observed effect or phenomenon. As a case in point, the cognitive-intervention literature typically includes a number of initial large effects later complemented by failures to replicate or by much more nuanced results. This pattern cannot be explained by random variation around single latent effect sizes; rather, it suggests the presence of moderators that have not been documented thus far. Ignoring heterogeneity in meta-analyses can be detrimental to the processes of inference and estimation, which has implications for power analyses, the precision of estimated effects, and replication research (Kenny & Judd, 2019). Methods such as mixture modeling are helpful for characterizing these variations, facilitating insight at the latent, rather than the observed, level, and enabling finer predictions that are testable and falsifiable.

Quantifying and characterizing heterogeneity across studies also has implications for the development of theoretical frameworks of cognitive improvement. A theoretical account that predicts interventions will work in specific settings, but not in others, is fundamentally different from one that can make more consistent predictions (Moreau & Corballis, 2019). As mentioned, such discrepancies in findings are often indicative of hidden moderators yet to be identified. For example, that brain training elicits improvement in some studies but not others, provided this finding is reliable, may be indicative of unknown moderating factors that could prove to be critical. In this context, data-driven techniques such as mixture modeling focused on the characteristics of effect-size distributions can facilitate targeted searches for moderators and help refine epistemological models of cognition.

Note that moderators do not have to relate to the intervention itself to be of influence—they could be embedded within the scientific process more generally. For example, multiple distributions of effect sizes could arise from well-known problems with current publishing practices, such as publication bias (Franco, Malhotra, & Simonovits, 2014) or perverse incentives (Stephan, 2012). However, for these issues to be the reason for the multimodality observed in cognitive-intervention research, they would need to exert a specific influence within these research areas that is mostly uncommon in other contexts, given the contrasting pattern observed in the broader field of psychology. Although a possibility—for example, publication bias could be exacerbated in cognitive-intervention research given pressure toward extreme, newsworthy findings that have applied potential—this hypothesis was beyond the current study.

Limitations of the current study

Before moving on to the broader implications of these findings, it should be pointed out that mixture-based assessments of multimodality remain probabilistic; that is, it cannot be definitively ascertained that the effect-size distributions in all four areas of research arose from multiple subpopulations or that the distributions that did not show this pattern in the control studies are indeed unimodal. A number of recent contributions have shown that mixture modeling does not always provide reliable assessments, either with respect to the number of components in a multimodal solution or to the mixing weights themselves (McLachlan & Peel, 2000). More specifically, mixture models can be affected by instabilities, most notably singularities of the likelihood function, resulting in all or most of the variance from a single component being concentrated on a single data point. This problem often leads to infinite likelihoods, effectively preventing convergence of the algorithm (Bishop, 2006; Caudill & Acharya, 1998; Murphy, 2012; Snoussi & Mohammad-Djafari, 2002).

In the current study, these potential limitations were mitigated in a number of ways. First, models were selected on the basis of log-likelihoods and confirmed via the BIC. This allowed combining methods that penalize complexity in a different way—either built in within model selection (BIC) or as an additional post-estimation step (log-likelihoods). Second, many of the issues in mixture modeling relate to the specific number of components (i.e., how many latent distributions are there?), which can vary substantially depending on methodology, parameters, and assumptions. In an effort to alleviate most of these concerns, the current study was not concerned with the specific number of components; rather, the goal was to characterize distributions as either unimodal or multimodal, irrespective of the specific number of estimated components. This method is much more robust to misclassifications

(Hunter & Lange, 2004; Moreau & Corballis, 2019), thus alleviating most of the typical concerns with mixture estimations.

Analyses of meta-analytic data were complemented by simulations with data-informed parameters to further quantify the performance of the algorithm in the specific context of the current study. If mixture modeling—or clustering techniques more generally—are in certain conditions hypersensitive to discrepancies in effect sizes within an area of research, we would expect to observe a high rate of multicomponent distributions in other areas of research; this was not the case in the current study. In addition, for miscalibration to be problematic in the current case, it would need to translate into the mislabeling of unimodal distributions as multimodal, yet results showed adequate reliability when using simulated distributions including the same parameters as those of each meta-analysis. These additional analyses showed that the method was well suited to the problem at hand; performance was acceptable overall and for each classification type.

Toward more accurate assessments of cognitive-intervention research

These potential limitations notwithstanding, the current findings provide further insight into the heterogeneity across the cognitive-intervention literature and suggest that until additional research can provide a better understanding of the source of these discrepancies, caution is warranted. However, these discrepancies should not be taken as evidence that each of these four areas of research is invalid or that the findings are questionable. A multitude of factors could contribute to the distributional properties observed in cognitive-intervention meta-analyses and more generally to the differences between this literature and other areas of research in psychology. For example, it could be argued that these interventions do not work for everyone and that their efficacy depends on individual traits and characteristics, which in turn drive the observed patterns. This is improbable here, however, as it represents a poor account of the pattern observed at the *group* level—for latent individual traits to underlie the observed effects, they would need to be systematic, and in this case resources should be dedicated to identifying hidden moderators. Other factors that may prevent generalizability should also be acknowledged; these factors may include population characteristics, or specificities of intervention protocols or of their implementations. The onus is arguably on proponents of cognitive interventions to identify moderators, not on the rest of the scientific community to speculate as to what these moderators might be.

Moreover, skepticism in the evaluation of cognitive-intervention research may seem at odds with mainstream findings in neuroscience praising the lifelong plasticity of the human brain. This apparent inconsistency is not well grounded, however, for a number of reasons. First, it is surprising that simple interventions that largely mimic natural feedback can have such a profound impact on individuals' abilities. If praising a growth mindset is all it takes to "unlock" children's full potential, how can years of feedback from teachers and parents be superseded by a couple of online praise sessions? Although there might be explanations involving complex factors yet to be identified, plausible answers are still lacking. Second, if cognitive abilities are largely dynamic and volatile, one would expect improvements to be transient and abilities to quickly revert back to individual baseline after the intervention (Taya, Sun, Babiloni, Thakor, & Bezerianos, 2015). This is typically not the case in the literature primarily considered herein, yet the underlying processes that would allow sustained improvements have not been detailed at the mechanistic level. Finally, the adaptive nature of neural systems (Anacker et al., 2018; Moreno-Jiménez et al., 2019; but see also Sorrells et al., 2018) does not necessarily mean that changes can be observed at the behavioral level. Neural changes are largely irrelevant in this discussion—the fact that behavioral improvements are plausible given current knowledge in neuroscience is not the question; even failed cognitive interventions can be related to neural changes (Román et al., 2016). Rather, the focus should be on whether *meaningful* behavioral improvements occur as a result of interventions.

Regardless of the current theoretical stance, one might assume that emphasizing malleability over fixed traits cannot be harmful, as it merely allows capitalizing on individual potential for change, with very little, if any, downside. It has been argued previously that this line of reasoning is questionable (Moreau et al., 2019)—beyond individual harm, overstating the effect of the environment, and especially of short-term interventions on achievement, impedes evidence-based changes in policies. For interventions to have a profound, meaningful impact, one needs to carefully consider the delicate balance between our natural willingness to have psychological findings translate to real-world applications and the necessary caution when those implementations come at the expense of evidence-based policies.

Concluding Remarks

Given the appeal of large changes with little resources or investment, cognitive interventions have generated a great deal of excitement among individuals and institutions

seeking personal or collective growth. However, the large, unexplained heterogeneity in cognitive-intervention meta-analyses hinders individual predictions and prevents sound institutional policies. In a mixture-modeling analysis of 111 meta-analyses in the field of psychology, I showed that cognitive interventions, including brain-training, video-gaming, mindset, and stereotype-threat research all appear to feature multiple subpopulations of effect sizes. This pattern was not common in meta-analyses within the broader field of psychology, suggesting that cognitive-intervention research exhibits particular characteristics that warrant further investigation. Although the specific nature of these characteristics remains undefined, the current findings further elucidate the specificities of cognitive interventions—a first step toward uncovering individual determinants of cognitive improvement.

Transparency

Action Editor: Aina Puce
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
D. Moreau is supported by a Marsden grant from the Royal Society of New Zealand, funding from the Neurological Foundation of New Zealand, and a University of Auckland Early Career Research Excellence Award.

ORCID iD

David Moreau https://orcid.org/0000-0002-1957-1941

Acknowledgments

The preregistration for this study describes the research rationale, inclusion and exclusion criteria, modalities for data extraction, analyses, and diagnostics. All scripts and data are available at https://osf.io/ce9vr.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/1745691620950696

Notes

1. Or, perhaps more accurately, all of these measures are usually interpreted under the assumption that a single latent distribution contributed to the overall distribution.
2. In practice, the estimation of means and mixing weights is combined with model comparison to ensure that the best characterization possible is selected on the basis of the data at hand. This point is detailed further in the Method section.
3. Note that this heuristic is the reverse of typical Bayesian information criterion (BIC) calculations in most contexts (e.g., regression analyses). In such cases, the BIC should be minimized; that is, a lower BIC indicates a better model.

References

Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Beek, T., Benning, S. D., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.

Adler, A. B., Bliese, P. D., Pickering, M. A., Hammermeister, J., Williams, J., Harada, C., . . . Ohlson, C. (2015). Mental skills training with basic combat training soldiers: A group-randomized trial. *The Journal of Applied Psychology*, *100*, 1752–1764.

Anacker, C., Luna, V. M., Stevens, G. S., Millette, A., Shores, R., Jimenez, J. C., . . . Hen, R. (2018). Hippocampal neurogenesis confers stress resilience by inhibiting the ventral dentate gyrus. *Nature*, *559*, 98–102.

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychology and Aging*, *23*, 765–777.

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: Learning to learn and action video games. *Annual Review of Neuroscience*, *35*, 391–416.

Belchior, P., Marsiske, M., Sisco, S. M., Yam, A., Bavelier, D., Ball, K., & Mann, W. C. (2013). Video game training to improve selective visual attention in older adults. *Computers in Human Behavior*, *29*, 1318–1324.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246–263.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*, 1363–1368.

Caudill, S. B., & Acharya, R. N. (1998). Maximum likelihood estimation of a mixture of normal regressions: Starting values and singularities. *Communications in Statistics - Simulation and Computation*, *27*, 667–674.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2011). Identifying and measuring heterogeneity. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). The Cochrane Collaboration. https://handbook-5-1.cochrane.org/chapter_9/9_5_2_identifying_and_measuring_heterogeneity.htm

Diamond, A., & Ling, D. S. (2019). Aerobic-exercise and resistance-training interventions have been among the least effective ways to improve executive functions of any method tried thus far. *Developmental Cognitive Neuroscience*, *37*, Article 100572. doi:10.1016/j.dcn.2018.05.001

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, *26*, 897–899.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Penguin Random House.

Forscher, P. S., Taylor, V. J., Cavagnaro, D., Lewis, N. A., Jr., Buchanan, E., Moshontz, H., . . . Chartier, C. (2019). *A multi-site examination of stereotype threat in Black college students across varying operationalizations*. doi:10.31234/osf.io/6hju9

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505.

Fryar, C. D., Gu, Q., & Ogden, C. L. (2012). Anthropometric reference data for children and adults: United States, 2007–2010. National Center for Health Statistics. *Vital and Health Statistics*, *11* (252). Retrieved from https://www.cdc.gov/nchs/data/series/sr_11/sr11_252.pdf

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., . . . Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*, 325–336.

Garrison, K. E., Tang, D., & Schmeichel, B. J. (2016). Embodying power: A preregistered replication and extension of the power pose effect. *Social Psychological and Personality Science*, *7*, 623–630.

Goldstein, J., Cajko, L., Oosterbroek, M., Michielsen, M., Van Houten, O., & Salverda, F. (1997). Video games and the elderly. *Social Behavior and Personality*, *25*, 345–352.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*, 534–537.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18*, 88–94.

Green, C. S., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior*, *28*, 984–994.

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.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573.

Halpern, D. (2014). Presidential column: Applying psychology to public policy. *APS Observer*, *27*(1). Retrieved from https://www.psychologicalscience.org/observer/applying-psychology-to-public-policy

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.

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

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, USA*, *105*, 6829–6833.

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated action. *Journal of Experimental Psychology: General*, *143*, 486–491.

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

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Nosek, B. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. doi:10.1177/2515245918810225

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., . . . Westerberg, H. (2005). Computerized training of working memory in children with ADHD—A randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 177–186.

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. *Psychology and Aging*, *30*, 180–193.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, *144*, 394–425.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*, 660–664.

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, *18*, 62–78.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley. doi:10.1002/0471721182

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": Evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*, 512–534.

Moreau, D., & Corballis, M. C. (2019). When averaging goes wrong: The case for mixture model estimation in psychological science. *Journal of Experimental Psychology: General*, *148*, 1615–1627.

Moreau, D., Macnamara, B. N., & Hambrick, D. Z. (2019). Overstating the role of environmental factors in success: A cautionary note. *Current Directions in Psychological Science*, *28*, 28–33.

Moreno-Jiménez, E. P., Flor-García, M., Terreros-Roncal, J., Rábano, A., Cafini, F., Pallas-Bazarra, N., . . . Llorens-Martín, M. (2019). Adult hippocampal neurogenesis is abundant in neurologically healthy subjects and drops sharply in patients with Alzheimer's disease. *Nature Medicine*, *25*, 554–560.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

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.

Okagaki, L., & Frensch, P. A. (1994). Effects of video game playing on measures of spatial performance: Gender effects in late adolescence. *Journal of Applied Developmental Psychology*, *15*, 33–58.

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

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Román, F. J., Lewis, L. B., Chen, C.-H., Karama, S., Burgaleta, M., Martínez, K., . . . Colom, R. (2016). Gray matter responsiveness to adaptive working memory training: A surface-based morphometry study. *Brain Structure & Function*, *221*, 4369–4382.

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, *26*, 515–520.

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin*, *144*, 111–139.

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.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, *29*, 549–571.

Snoussi, H., & Mohammad-Djafari, A. (2002). Penalized maximum likelihood for multivariate Gaussian mixture. *AIP Conference Proceedings*, *617*(1), Article 36. doi:10.1063/1.1477037

Sorrells, S. F., Paredes, M. F., Cebrian-Silla, A., Sandoval, K., Qi, D., Kelley, K. W., . . . Alvarez-Buylla, A. (2018). Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. *Nature*, *555*, 377–381.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797–811.

Stephan, P. (2012). Research efficiency: Perverse incentives. *Nature*, *484*, 29–31.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*, 93–102.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777.

Suggs, D. L. (1979). The use of psychological research by the judiciary: Do the courts adequately assess the validity of the research? *Law and Human Behavior*, *3*, 135–148.

Taya, F., Sun, Y., Babiloni, F., Thakor, N., & Bezerianos, A. (2015). Brain enhancement through cognitive training: A new insight from brain connectome. *Frontiers in Systems Neuroscience*, *9*, Article 44. doi:10.3389/fnsys.2015.00044

U.S. Department of Education. (2015, October 14). U.S. Department of Education announces first ever skills for success grants and initiative to support learning mindsets and skills [Press release]. Retrieved from https://www.ed.gov/news/press-releases/us-department-education-announces-first-ever-skills-success-grants-and-initiative-support-learning-mindsets-and-skills

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*, 302–314.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*, 364–369. doi:10.1038/s41586-019-1466-y