Journal of Abnormal Psychology
in press.

https://doi.org/10.1037/abn0000421

Evaluating the Evidential Value of Empirically Supported Psychological Treatments (ESTs): A Meta-Scientific Review

*John Kitchener Sakaluk
University of Victoria

*Alexander J. Williams
University of Kansas

Robyn E. Kilshaw
University of Utah

Kathleen T. Rhyner
Canandaigua VA Medical Center

Empirically supported treatments (or therapies; ESTs) are the gold standard in therapeutic interventions for psychopathology. Based on a set of methodological and statistical criteria, the APA has assigned particular treatment-diagnosis combinations EST status and has further rated their empirical support as Strong, Modest, and/or Controversial. Emerging concerns about the replicability of research findings in clinical psychology highlight the need to critically examine the evidential value of EST research. We therefore conducted a meta-scientific review of the EST literature, using clinical trials reported in an existing online APA database of ESTs, and a set of novel evidential value metrics (i.e., rates of misreported statistics, statistical power, R-Index, and Bayes Factors). Our analyses indicated that power and replicability estimates were concerningly low across almost all ESTs, and individually, some ESTs scored poorly across multiple metrics, with Strong ESTs failing to continuously outperform their Modest counterparts. Lastly, we found evidence of improvements over time in statistical power within the EST literature, but not for the strength of evidence of EST efficacy. We describe the implications of our findings for practicing psychotherapists and offer recommendations for improving the evidential value of EST research moving forward. **General Scientific Summary:** This review suggests that although the underlying evidence for a small number of empirically supported therapies is consistently strong across a range of metrics, the evidence is mixed or consistently weak for many, including some classified by Division 12 of the APA as "Strong."

Data, analysis code, supplementary material: https://osf.io/73drs/

Keywords: Empirically supported treatments; evidential value; meta-science; replicability

Clinical efficacy underpins everything the psychotherapy industry promises and is ethically-bound to deliver. Questions about whether psychotherapy works—and if so, for whom and under what conditions—have guided research for the better part of a century (e.g., Eysenck, 1952). However, only since the 1970s have controlled trials like those used in medicine flourished in the field of psychology (e.g., Klerman et al., 1974). In an effort to better synthesize and disseminate the results of this efficacy research, an APA Division 12 Task Force (1995) created a continually-updated list of the therapies that have reached a certain level of research support for treating patients with specific diagnoses. These therapies came to be known as empirically-supported psychological treatments (or therapies) (ESTs; Kendall, 1998).

The level of evidential support for each EST has been further rated according to a set of methodological and statistical criteria (Chambless et al., 1998). In recommending updated criteria for the evaluation of ESTs, Tolin et al. (2015) argued that the EST

*Williams and Sakaluk share first authorship; their order was arbitrarily decided by a bitterly-contested coin toss. The authors sincerely thank Drs. Corker, Fried, Grubbs, Kirk, and Nuijten for their consultation on methodological and clinical matters throughout our review, and Kyle Dirck and Dr. Robert Williams for their helpful suggestions on earlier drafts of the manuscript. This research was supported by a SSHRC Insight Development Grant awarded to Sakaluk. Sakaluk is a statistical consultant at other journals and methodological/quantitative expert who has published on the replicability crisis before. Williams is psychological clinic director who has used ESTs in his own practice and encourages trainees to use them. Rhyner is a VA psychologist.

movement has "had substantial impact in psychology and related mental health disciplines," noting the dissemination of ESTs via Division 12's website as an "immediately tangible effect." Furthermore, many meta-analyses and literature reviews support the efficacy of specific ESTs (e.g, Kliem, Kröger, & Kosfelder, 2010) or groups of ESTs (e.g., Hollon, Thase & Markowitz, 2002), and some scholars have argued that the quality of evidence underlying ESTs has improved over time (Thoma et al., 2012).

Evaluating the Evidential Value of ESTs

Traditional criteria of evidential value. Chambless and colleagues articulated the original criteria that Division 12 used to identify ESTs and classify them by the strength of their underlying scientific evidence (for expanded criteria, see Chambless et al., 1998; Chambless & Hollon, 1998). Specifically, therapies that have repeatedly demonstrated statistically significant improvements over no-treatment, placebo, or another alternative treatment, are described by Division 12 as Strongly supported. Modestly supported therapies are those that have only once significantly outperformed no-treatment, placebo, or another alternative treatment; and therapies that have shown inconsistent improvement over no-treatment, placebo, or another alternative treatment are designated as having Controversial support (see https://www.div12.org/psychological-treatments/frequently-asked-questions/#support).

Although these criteria marked a desirable step towards greater rigor in the evaluation of psychological treatments, their reliance on statistical significance for determining EST-status presents a concern. As nearly a decade of the replicability crisis (see Nelson et al., 2018) has demonstrated, common misuse and misunderstanding of null-hypothesis significance testing (NHST) renders statistical significance a precarious and easily-misleading standard of evidence. Although conversations about the replicability crisis have only recently begun in the clinical realm (Tackett et al., 2017), there is growing awareness that many of the factors undermining the replicability of findings in other areas of psychology (Open Science Collaboration, 2015) have likely impacted the clinical science literature as well (see Coyne & Kok, 2014; Cuijpers & Cristea, 2016; Flint, Cuijpers, Horder, Koole, & Munafo, 2014).

Alternative metrics of evidential value. Guided by concerns about the consequences that the reliance, misunderstanding, and/or misuse of NHST may have on the replicability of psychological findings (Cumming, 2014), psychological methodologists have proposed a number of other metrics for assessing the strength of research results beyond breaching the $p < .05$ threshold. Although no one of the following metrics, individually, is a pipeline to the "true" evidential value of a particular set of effects, they each capture some unique feature of persuasive evidence. Cumulatively, they can provide a robust picture of the state of evidence underlying the EST literature. We believe that any therapy meriting the term *empirically supported* should fare well across these metrics.

First, the misreporting of inferential statistics (e.g., degrees of freedom, test statistic values, *p*-values) is a surprisingly common practice in psychological research, with nearly 50% of published articles using NHST containing at least one reporting error, and roughly 13% containing a "gross" reporting error that undermines conclusions of statistical significance (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Typically the product of simple human error, misreported statistics nevertheless erode confidence in the credibility of results. In the case of ESTs, readers should place greater confidence in therapies for which the supporting results were reported accurately.

Next, statistical power—which refers to the long-run probability of detecting an effect when one truly exists—has been a recurring concern in the replicability of scientific findings (Maxwell, 2004). Despite recommendations that studies be conducted with 80% power for the expected effect size, recent reviews have found that the average social science study possesses only a 44% chance of detecting an existing medium-sized true effect (Szucs & Ioannidis, 2017). Power can be computed post hoc using a study's design, sample size, and the magnitude of the observed effect. Although these post-hoc power estimates are generally unreliable estimates of true power in the singular, when aggregated across studies they become increasingly accurate approximations of true power (see Schimmack, 2012). Consequently, in the case of the EST literature, reviewers should have more confidence in therapies predicated on well-powered studies that could reliably detect reasonably-sized effects.

As an extension of power calculations, the Replicability-Index (R-Index; Schimmack, 2016) can be used to further evaluate how plausible a set of reported significant effects is, given the level of observed power for those effects. Inflated rates of significance beyond the power observed for a set of effects (e.g., 80% significant effects in a set of effects when median post-hoc power was .60) can suggest effects are implausible, and potentially the result of

selective reporting and/or questionable research practices (Schimmack, 2016). Readers should therefore have more confidence in ESTs supported by studies that have non-inflated rates of statistical significance given their statistical power.

In contrast to the dichotomous decision-making of NHST, Bayes Factors offer a means of evaluating competing hypotheses on a continuous metric of evidential strength. Explicitly, a Bayes Factor is the (weighted average) likelihood ratio of two different hypotheses for a given set of data. For example, in a study comparing the efficacy of an EST to that of a control condition, a Bayes Factor of 10 suggests that the data are 10 times more probable under the alternative hypothesis (i.e., EST efficacy) than they are under a null hypothesis of no effect (a Bayes Factor of 0.10, meanwhile, would indicate the data were 10 times more probable under the null of no effect). Jeffreys (1961) provided descriptive thresholds for the strength of evidence in a given Bayes Factor, including *anecdotal* evidence ($1 < BF < 3$), *moderate* evidence ($3 < BF < 10$), *strong* evidence ($10 < BF < 30$), and *very strong* evidence ($BF > 30$). Bayes Factors have recently been extended to the meta-analytic context, where they can be used to synthesize the relative strength of evidence for an effect across studies (e.g., Moden et al., 2018). In the case of the EST literature, readers should have greater confidence in therapies with data that are much more likely under the hypothesis of therapeutic efficacy than under the null of no therapeutic effect.

The Current Meta-Scientific Review

In the current meta-scientific review, we assess the evidential value of the literature underlying 79 ESTs identified by Division 12 of the APA. Instead of relying exclusively on traditional statistical significance, however, our review adopts a broader conceptualization of what metrics constitute evidence for (or against) a therapy. Specifically, we use rates of misreporting, estimates of statistical power, R-Index values, and Bayes Factors to answer the following questions:

1. *What is the evidential value underlying ESTs collectively and individually?*
2. *What is the evidential value for ESTs classified as possessing Strong empirical support, relative to those classified as having only Modest support?*
3. *Have the standards of evidential support for ESTs improved over time?*

Methods

Sample of Effects

Effects were drawn from articles listed by Division 12 as "Key References" or "Clinical Trials" for each of the ESTs listed on the Division 12 website (https://www.div12.org/treatments/). In this way, we proceeded with a sample of literature that was defined by others, removing one area in which our subjectivity might otherwise bias our appraisal of the literature.

Each article typically reported a myriad of statistical tests that were not related to the efficacy of the EST(s) under consideration (e.g., demographic comparisons at baseline, covariate tests, subgroup analyses). We therefore determined criteria for choosing which effect(s) we considered relevant to our meta-scientific review (see preregistration, https://osf.io/jz7ty/). In summary, we attempted to code inferential and descriptive statistics (for pairwise comparisons between treatment/control groups) described by authors as "primary," or, failing that, those appearing relevant to treatment efficacy for the particular diagnosis.

Coding Strategy

Coders (all authors except Sakaluk) were trained—and the coding system refined—by applying the coding system to two trial ESTs. All ESTs were then coded by two of the three coders. Because articles reported various numbers of effects (only some of which fell within our criteria), the first coder proposed the number of relevant effects, with the other coder confirming (or challenging) that number. Disagreements about particular effects were discussed until a consensus was reached; the qualities of each relevant effect were then coded independently.

Analytic Strategy

Computation of metrics. We used Schönbrodt's (2018) *p-checker* app (http://shinyapps.org/apps/p-checker/) to facilitate identification of misreported statistical results (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016)[1], estimation of median observed power, and computation of R-Index values (Schimmack, 2016). We calculated an additional index of statistical power—what estimated effect size could be reliably detected (80% power) given the sample sizes of the treatment and control conditions—using the *pwr* package (Champley, 2018) for R.

We also used the *BayesFactor* package (Morey & Rouder, 2018) to calculate Bayes Factors (BF_{10}) of the evidence for EST efficacy relative to that of control conditions using two different analytic strategies. As

with all Bayesian methods, computation of BF_{10} requires the specification of a prior distribution for each competing hypothesis (i.e., a probability distribution for the null and alternative hypotheses, given the data; see Etz, Haaf, Rouder, & Vandekerckhove, 2018). In situations such as our analyses, one common approach is to use an uninformative prior such as the Cauchy distribution (typically referred to as the Jefferys-Zellner-Siow, or JZS, prior), which does not give preference to any particular range of probability, and instead allows the data to dominate the resulting likelihood.

First, we calculated the range of individual BF_{10} from *t* statistics (or descriptive statistics) in all studies within a given EST. We then conducted Bayesian Meta-Analysis (see Moden et al., 2018) to synthesize BF_{10} across effects and to calculate posterior estimates of effect sizes for each EST[2]. As most studies reported more than one relevant effect size, and accessible software for multilevel Bayesian Meta-Analysis is not yet mainstream, we fit two different meta-analytic models: the first using the smallest effect from each study (a *pessimistic* meta-analytic estimate), and the second using the largest effect from each study (an *optimistic* meta-analytic estimate). All BF_{10} were calculated as directional (i.e., one-way) tests of treatment efficacy using JZS priors, and we note whether results meeting a particular threshold of evidence are contingent on the width of prior. Finally, as per the Division 12 criteria for establishing an EST as Strong, the meta-analysis of Strong EST effects (vs. Modest and Controversial) was limited to comparisons against active controls, placebo pill or treatment, or another bona fide therapy.[3]

Appraisal of metrics. In order to appraise the evidential value of the EST literature based on our selected metrics, we had to determine for each what we would consider strong, modest, or controversial evidence (https://osf.io/2nwrs/). We consulted with four methodological and clinical experts who generally felt our thresholds for interpretation were reasonable (yet arbitrary); this should neither be taken as evidence that they stand by other features of our process, nor that our synthesis is correct. Rather, our consultants have simply—yet importantly—suggested that our approach to one of the most subjective aspects of our synthesis was not *entirely* ridiculous.

Sample of Effects

We have endeavored to make as many aspects of our research process as transparent as possible. All features of our research process are open (https://osf.io/73drs), including our data set, initial selection criteria (preregistered https://osf.io/jz7ty/), interpretation thresholds, and clarifications/changes made to this protocol. In the course of preregistering, we also detailed each of our individual predictions for the synthesis and any potential conflicts of interest.

Despite our efforts to reduce bias in our synthesis, readers may be skeptical of our process and disagree with particular coding and/or analytic decisions we made. Given the quality of statistical reporting in the EST literature, our authorship team believes such skepticism is warranted; it is likely not all will agree with our coding decisions. We have therefore taken the additional step of making our entire dialogue and troubleshooting efforts transparent (see OSF).

Results

Sample of Effects and Reliability

Dual-coding of our sample of EST-relevant papers yielded a total of n = 3463 effects from 453 articles, although not all of these effects were usable, as omitted statistical information (e.g., sample sizes per group, degrees of freedom, test statistic value, etc.,) prohibited the inclusion of particular effects when calculating certain metrics.[4] We evaluated the reliability of our coding strategy in the following way: for a subset of 10 ESTs, power estimates and individual BF_{10} range were computed using each coder's sample of effects; the resulting estimates were then compared for consistency. Reliability of power estimates was excellent (r = .99) with coders producing identical results for eight ESTs and minor discrepancies for the other two. Reliability of coding for individual BF_{10} was similar, with only two appreciable discrepancies, and a high level of consistency between minimum BF_{10} (r = .99) and maximum BF_{10} (r = .99) calculated for each EST.

Observations of Low Reproducibility and High Analytic Flexibility

Although not among our original aims, one of the first findings—and potentially the most important—from our meta-scientific synthesis was readily apparent to the naked eye: the vast majority of statistical analyses in the EST literature were not reported with sufficient detail to make inferential tests verifiable (i.e., checking for reporting errors) or re-analyzable (i.e., calculating the other metrics of evidential value). The lack of complete reporting details for inferential

statistics (and even descriptive statistics, including group means and standard deviations) is puzzling considering the APA's statistical reporting standards (American Psychological Association, 2010).

Relatedly, it was often unclear which statistical test(s) for which measure(s) constituted the "focal test(s)" for a given study; a primary reason why we were not able to utilize other metrics of evidential value (e.g., *p*-curve[5], test of insufficient variance) that rely on the selection of one—and only one—effect from a given sample. Indeed, even for very commonly-repeated designs (e.g., randomized controlled trials with pretest, posttest, and follow-up assessments), studies in our sample varied widely with respect to which test was deemed key to evaluating the evidence of the EST in question (e.g., the main effect of *condition* versus the *time*×*condition* interaction; see *Self-System Therapy for Depression* and *Cognitive Processing Therapy for PTSD* as examples).

What is the Evidential Value of ESTs Collectively and Individually?

Summary of trends across ESTs. Descriptive statistics for each metric across all ESTs are listed in Table 1. Rates of both gross and minor errors varied widely, although for most ESTs, results were reported accurately. Statistical power and R-Index, meanwhile, appeared to be the most consistently concerning metrics, as EST research was typically underpowered with a co-occurring inflation in the reporting of statistically significant effects. Finally, Bayes Factors across ESTs suggest that the strength of evidence in favor of treatment efficacy is highly contingent on whether one places greater stock in selecting the more pessimistic or optimistic effects from this literature.[6]

Summary of individual ESTs. Metrics of evidential value for each EST are listed in Table 2. The most common pattern for individual ESTs was mixed, with ESTs scoring well on one or two metrics and less well on others. A small number of ESTs (e.g., both *Cognitive Processing Therapy* and *Prolonged Exposure for PTSD*) scored consistently well across all or most metrics, whereas a larger number of ESTs—including a number classified as Strong (e.g., *Behavioral Activation for Depression*, *Cognitive Remediation for Schizophrenia*, *Dialectical Behavior Therapy for Borderline Personality Disorder*)—performed relatively poorly across most or all of our metrics of evidential value. Low reporting quality in articles for other ESTs (e.g., *Cognitive Behavioral Therapy for Insomnia*, *Family-Based Treatment for Bulimia Nervosa*) made it impossible to calculate many metrics.

What is the Evidential Value for Div. 12 Classifications of Strength of Evidence?

Metrics of evidential value across Division 12's categorizations of strength of supporting evidence are presented in Table 3. ESTs classified as possessing Modest or Strong evidence both had comparably low rates of misreported statistics. ESTs classified as Strong, however, were better powered and had stronger replicability estimates, although studies for both Modest and Strong ESTs were typically poised to detect effects of comparable magnitude. Interestingly, BF_{10} did not indicate that the strength of evidence in favor of Strong ESTs was consistently greater than the strength of evidence for Modest ESTs; although meta-analytic BF_{10} suggested Strong ESTs were better supported using optimistic estimates from each study, Modest ESTs were decisively better supported when using pessimistic estimates from each study. ESTs classified as Controversial had very few usable effects, but what metrics we could calculate suggested that their evidential value is relatively low.

Are Standards of Evidential Value Improving?

Finally, we evaluated to what extent, if any, standards of evidence underlying research on ESTs have improved over time. We evaluated this possibility using two different metrics tapping into relatively distinct conceptualizations of evidential value: statistical power and meta-analyzed (optimistic) BF_{10}. Results for the analysis of statistical power (Figure 1) suggest that standards of statistical power have improved over time. Trends in BF_{10} offer a much less clear picture of evidential value over time. Though BF_{10} generally increased from the 1970s through the 1980s, very modest BF_{10} periodically reappeared, including relatively recently. Therefore, although methodological quality of EST research appears to be gradually improving over time, there is no compelling pattern to suggest the same is true of evidence for treatment efficacy *per se*.

General Discussion

Our meta-scientific evaluation found inconsistent support for the evidential value underlying the literature on ESTs. Although some ESTs performed well across all metrics of replicability (e.g., *Exposure*

Therapy for Specific Phobias), the support for other ESTs is decidedly mixed or weak (e.g., *Family Psychoeducation for Schizophrenia*). In summary, inadequate reporting of inferential statistics, primary hypotheses, and focal analyses is a prominent issue across all ESTs. Furthermore, studies supporting ESTs were equipped to reliably detect only implausibly large effects (see Simmons, 2014), consistently fell short of recommendations for 80% power, and yielded replicability estimates that fell below what is currently normative for clinical research (see Schimmack, 2017).

Individual and meta-analyzed Bayes Factors suggested that under the most pessimistic selections of effects, statistical evidence often favored a null of no treatment effect; under optimistic selections, statistical evidence of treatment efficacy was quite strong. Perhaps most interestingly, our review suggests that there is not always a clear distinction between Division 12 classifications of Strongly and Modestly supported ESTs, as meta-analyzed pessimistic selections for Strong ESTs (which were in the range of Moderate support, according to the cutoffs proposed by Jeffrey's, 1961) were actually several orders of magnitude weaker than the corresponding effects for Modest ESTs.

Nevertheless, not all results were pessimistic: statistical misreporting was rare, especially compared to the rates observed by Nuijten et al. (2016), and when present to a more concerning degree, was clustered within particular ESTs. And although EST studies have remained underpowered, we find some evidence that statistical power has improved over time (though evidence of efficacy has not). These findings have multiple implications for practitioners and researchers alike.

Suggestions for EST Practitioners

Prasad & Cifu (2011) coined the term *medical reversals* to describe situations in which medical practices are put into use—often without evidence of efficacy—only to fall out of use due to subsequent evidence that they are ineffective, offer no advantage over less costly alternatives, or are even iatrogenic. Without evidence for EST superiority, the mental health fields may need *psychotherapy reversals* if the monetary and opportunity costs of EST training, dissemination, and use exceed those associated with other bona fide psychotherapies.

In light of mixed support for the evidential value of many ESTs, as well as similarities between Modest and Strong ESTs across some metrics (e.g., BF_{10}), readers may be inclined to conclude that most ESTs are roughly similar in their efficacy. We think such a conclusion would be unwarranted. Capitalizing on the transitive nature of BF_{10} (see Morey & Rouder, 2011), our Bayesian meta-analysis makes it possible for interested parties to quantify the relative support for competing ESTs simply by dividing the two respective BF_{10} involved. Using the optimistic estimates for two ESTs for PTSD as examples, our analyses suggest that the efficacy of *Cognitive Processing Therapy* is 38.89E+24 times more likely than that of *Eye Movement Desensitization and Reprogramming* (11.00E+26 divided by 28.82). Comparisons like these could be one useful mechanism for therapists to select among competing therapies, although in cases where competing therapies do not exceed a particular threshold of strength of evidence (e.g., $BF_{10} > 3$), this relative comparison may be less informative.

Transitive BF_{10} comparisons and other metrics of evidential value will not always render a clear answer for which EST has a stronger basis of evidence. Consequently, our analyses suggest that in many cases, we do not have a *dodo bird* (Rosenzweig, 1936), but rather a *don't know bird*. Based on the available evidence, we *don't know* if there are differences in the level of empirical support for ESTs, and we *don't know* if ESTs offer benefit beyond that of other bona fide psychotherapies in treating patients with specific diagnoses.

To maximize the potential benefit of treatment, mental health professionals should use ongoing, frequent, research-based assessments (e.g., https://www.div12.org/assessment-repository/). If a patient is not making progress, the therapist and patient should consider switching therapies. While this has always been sound advice, it takes on added importance when the therapist and patient cannot assume a given EST is predicated on strong evidence.

Suggestions for Future EST Research and Assessment of Evidential Value

Our review illuminates numerous pathways for improving future EST research and appraisals thereof. Developing and enforcing standards of reporting descriptive and inferential statistics would be a simple change; however, it would yield immediate dividends for the quality of EST literature by reducing omissions of key statistical information. Without these pieces of information, readers cannot "fully understand the analyses conducted" (APA, 2010, p. 116), including whether they were appropriately powered or even

whether they were correctly reported. A recent meta-scientific review (Cristea, Florian, Nutu, & Gentili, 2018) suggests that in order for these standards to be successful, they will need to be enforced by policy of consequence.

Increased preregistration is also needed to help demarcate which EST-related findings are exploratory and which are confirmatory. It was often the case that "key" effect(s) within a given study were not identified, and when they were, it was unclear why an effect for a particular symptom- or function-related measure was deemed "key" while another seemingly related measure was not. Though exploratory research into ESTs could be beneficial for discovering new pathways to therapeutic efficacy, preregistered studies with clearly demarcated measures deemed key *a priori* should be required before evidence for a given EST is considered strong. Professional societies could aid this effort by providing standardized guidelines for what type of statistical comparisons (e.g., a *time*×*condition* interaction) within common designs are—and are not—crucial to establishing therapeutic efficacy.

Finally, though the trend towards increased statistical power in EST research is a positive development, there must be greater continued effort to increase the evidential value—broadly construed—of the EST literature. As most studies remain underpowered for establishing clinical efficacy, clinical researchers should strive for yet-higher-powered studies, especially if they intend to detect smaller effects of either superiority relative to stronger controls or equivalence to other ESTs and/or medications. Supplementing traditional significance testing with Bayesian inference strategies could illuminate when comparisons between ESTs and controls are simply uninformative. Moving forward, however, EST research may need to eschew the model of small trials. A combined workflow of larger multi-lab registered reports (Chambers, 2013; Uhlmann et al., 2018) coupled with thorough analytic review (Sakaluk, Williams, & Biernat, 2014) would yield the highest degree of confirmatory, accurate evidence for the efficacy of ESTs.

Limitations

Our meta-scientific review is primarily limited by its scope. Although we avoided introducing bias in article inclusion criteria by relying on the Key References/Clinical Trials identified by Division 12 for each treatment, this may have been at the expense of a more comprehensive review. Further, the number of useable effects to calculate each metric of evidential value varied across articles and ESTs; as a result, some of our estimates were based on a very limited number of effects. In these cases, readers should be cautious about placing too much value in any one index of evidential value.

Given these shortcomings, our review should be received with a healthy dose of skepticism. Still, these limitations cut both ways. It would seem difficult to justify decrying the validity of our findings due to the limited scope of our review or the sparseness of useable effects, while placing the utmost confidence in the existing assessments of ESTs—they both rely on the very same limited body of evidence.

Conclusion

The movement to identify ESTs has been one of the most important in the history of psychotherapy; however, existing classifications of ESTs must be re-evaluated in light of the replication crisis. This meta-scientific review produced mixed evidential support for many ESTs, with few demonstrating consistently strong research support and some exhibiting consistently weak support. When strong evidential support is lacking for all ESTs of a given diagnosis (e.g., *Anorexia Nervosa*), psychologists working with such patients should consider increasing their use of research-based assessments to track therapeutic benefit or the lack thereof. Future controlled-trial research on psychotherapies would benefit greatly from preregistration, standardized reporting of results, and increased sample sizes facilitated by collaboration across multiple labs. Researchers should keep in mind the needs of their colleagues to verify, replicate, and implement research findings if we hope to deliver on psychotherapy's mission to improve mental health.

Endnotes

1. Any effects flagged as grossly misreported by p-checker were manually confirmed by the first author. In some cases, authors indicated using a valid one-tailed test, in which case, results were not classified as misreported. In some instances, however, authors reported a one-tailed testing strategy that was not valid (e.g., for an ANOVA, which is already a one-tailed test), in which case, results were classified as a gross reporting error.

2. We only later resolved to also include meta-analytic posterior estimates of EST effect sizes, in addition to Bayes Factors, and so we did not propose

cutoffs (or vet potential cutoffs with our consultants). We therefore consider them a descriptive and exploratory, yet important metric in our review.

3. We use "bona fide psychotherapy" to mean any treatment that Division 12 list as an EST for a diagnosis other than the one under review. For example, interpersonal therapy is an EST for depression but not obsessive-compulsive disorder (OCD). If a study that was a Key Reference for OCD treatment used interpersonal therapy as a control condition, we coded interpersonal therapy as a bona fide psychotherapy for OCD.

4. 915 effects (26%) were available to submit to *p-checker* (misreporting, post-hoc power, R-Index), 2479 (72%) were available to use with the *pwr* package; and 1384 (40%) were available to use with the *BayesFactor* package.

5. Although a formal *p*-curve analysis was not possible given the availability of multiple p-values from the same sample, we have provided simple histograms of all exact p-values and all significant p-values in our OSF repository. Inexact significant *p*-values of $p < .05$ were much more commonly reported ($n = 257$) than were $p < .01$ ($n = 115$) or $p < .001$ ($n = 131$)

6. Given the presence of a select number of extreme Bayes Factors, we base our interpretation of typical levels of evidential value on the median summaries (vs. the mean).

References

Association (2010). *Publication Manual of the American Psychological Association* (6th Edition). Washington, DC: APA.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at *Cortex*. *Cortex*, *49*, 609-610. https://doi.org/10.1016/j.cortex.2012.12.016

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., ... & Johnson, S. B. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist*, *51*, 3-16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18. https://doi.org/10.1037/0022-006X.66.1.7

Champely, S. (2018). pwr: Basic functions for power analysis (ver. 1.2-2.). Retrieved from https://CRAN.R-project.org/package=pwr.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*, 145-153. https://doi.org/10.1037/h0045186

Coyne, J. C., & Kok, R. N. (2014). Salvaging psychotherapy research: A manifesto. *Journal of Evidence-Based Psychotherapies*, *14*, 105-124.

Cristea, I., Florian, N., Nutu, D., & Gentili, C. (2018). Open science practices in clinical psychology journals: an audit study. https://doi.org/10.31219/osf.io/t2k56

Cuijpers, P., & Cristea, I. A. (2016). How to prove that your therapy is effective, even when it is not: A guideline. *Epidemiology and Psychiatric Sciences*, *25*, 428–435. https://doi.org/10.1017/S2045796015000864

Cumming, G. (2014). The new statistics: Why and How. *Psychological Science, 25,* 7-29. Doi:10.1177/0956797613504966

Eysenck, H. J. (1952). The effects of psychotherapy: an evaluation. *Journal of Consulting Psychology, 16*(5), 319-324. http://dx.doi.org/10.1037/h0063633

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1,* 281-295. doi: 10.1177/2515245918773087

Flint, J., Cuijpers, P., Horder, J., Koole, S. L., & Munafò, M. R. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological Medicine*, *45*, 439–446. https://doi.org/10.1017/S0033291714001421

Hollon, S. D., Thase, M. E., & Markowitz, J. C. (2002). Treatment and prevention of depression. *Psychological Science in the Public Interest*, *3*, 39-77. https://doi.org/10.1111/1529-1006.00008

Jeffreys, H. (1961). *The theory of probability*. New York, NY: Oxford University Press.

Kendall, P. C. (1998). Empirically supported psychological therapies. *Journal of Consulting and Clinical Psychology*, *66*, 3–6. https://doi.org/10.1037/0022-006X.66.1.3

Klerman, G. L., Dimascio, A., Weissman, M., Prusoff, B., & Paykel, E. S. (1974). Treatment of depression by drugs and psychotherapy. *American Journal of Psychiatry*, *131(2)*, 186-191.

Kliem, S., Kröger, C., & Kosfelder, J. (2010). Dialectical behavior therapy for borderline personality disorder: A meta-analysis using mixed-effects modeling. *Journal of Consulting and Clinical Psychology*, *78*, 936-951. https://doi.org/10.1037/a0021015

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163. DOI: 10.1037/1082-989X.9.2.147

Monden, R., Roest, A. M., van Ravenzwaaij, D., Wagenmakers, E. J., Morey, R., Wardenaar, K. J., & de Jonge, P. (2018). The comparative evidence basis for the efficacy of second-generation antidepressants in the treatment of depression in the US: A Bayesian meta-analysis of Food and Drug Administration reviews. *Journal of Affective Disorders*, *235*, 393-398. https://doi.org/10.1016/j.jad.2018.04.040

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406-419. https://doi.org/10.1037/a0024377

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs (ver. 0.9.12-4.2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*,*69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

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

Prasad, V., & Cifu, A. (2011). Medical reversal: why we must raise the bar before adopting new technologies. *The Yale Journal of Biology and Medicine*,*84*, 471-478.

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6(3)*, 412-415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102-113. https://doi.org/10.3758/s13423-017-1420-7

Sakaluk, J. K., Williams, A. J., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, *9*, 652-660.https://doi.org/10.1177/1745691614549257

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

Schimmack, U. (2016). The replicability-index: Quantifying statistical research integrity. Retrieved from https://wordpress.com/post/replication-index.wordpress.com/920

Schimmack, U. (March, 2017). *2016 replicability rankings of 103 psychology journals*. Retrieved from https://replicationindex.wordpress.com/2017/03/01/3950/.

Schönbrodt, F. D. (2018). p-checker: One-for-all p-value analyzer. Retrieved from http://shinyapps.org/apps/p-checker/.

Simmons, J. (April, 2014). *[18] MTurk vs. the lab: Either way we need big samples.* Retrieved from http://datacolada.org/18.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, 15, E2000797.

Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psychological procedures: Report and recommendations. *The Clinical Psychologist*, 48, 3-23.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., … Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science.*Perspectives on Psychological Science*,*12*, 742–756. https://doi.org/10.1177/1745691617690042

Thoma, N. C., McKay, D., Gerber, A. J., Milrod, B. L., Edwards, A. R., & Kocsis, J. H. (2012). A quality-based review of randomized controlled trials of cognitive-behavioral therapy for depression: An assessment and meta-regression. *American Journal of Psychiatry*, *169*, 22–30. https://doi.org/10.1176/appi.ajp.2011.11030433

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice*,*22*, 317-338. https://doi.org/10.1111/cpsp.12122

Uhlmann, E. L., Chartier, C. R., Ebersole, C. R., Errington, T. M., Kidwell, M., Lai, C. K., … Nosek, B. A. (2018, August 13). Scientific utopia: III. Crowdsourcing science. https://doi.org/10.31234/osf.io/vg649

Table 1

Descriptive Statistics for Each Index of Evidential Value

	% Misreported (Gross)	% Misreported (Minor)	Median Post Hoc Power	R-Index	80% Power to detect Median *d* (1-tailed)	80% Power to detect Median *d* (2-tailed)	BF_{10} Range (Min)	BF_{10} Range (Max)	Pessimistic Meta BF_{10}	Optimistic Meta BF_{10}
M	2.07	8.17	.58	.52	0.69	0.79	2.03E+30	4.75E+29	2.16E+30	1.05E+35
Mdn	0	0	.58	.50	0.69	0.78	0.12	71.24	0.45	140.62
SD	5.15	16.30	.21	.24	0.24	0.28	1.66E+31	3.75E+30	1.71E+31	8.41E+35
Min	0	0	.06	0	0.19	0.21	0.0009	0.1	0.0002	0.07
Max	22	100	1	1	1.16	1.49	1.36E+32	3.09E+32	1.36E+32	6.73E+36

Table 2

Evidential Value Metrics for Each Empirically Supported Treatment

EST	Div. 12 Classification	# of Usable Effects	% Gross Errors	% Minor Errors	Median Post Hoc Power	R-Index	80% Power to detect Median *d* (1-tailed/2-tailed)	BF_{10} Range	Pessimistic Meta BF_{10} (Posterior 95% CR: LL_d/UL_d)	Optimistic Meta BF_{10} (Posterior 95% CR: LL_d/UL_d)
Acceptance and Commitment Therapy for Obsessive-Compulsive Disorder	Modest	1/2/2/2	0%***	0%***	.61**	.21*	0.62/0.69*	3.45^{x}**/64.69***	3.45**^{x} (0.10/1.04)	64.69*** (0.35/1.30)
Acceptance and Commitment Therapy for Chronic Pain	Strong	10/13/4/4	0%^{z}***	0%***	.63**	.56*	0.58/0.66*	.05*/4.69**	.06* (0.02/0.20)	.07* (0.002/0.24)
Acceptance and Commitment Therapy for Depression	Modest	19/50/41/16	11%*	5%*	.51*	.45*	0.85/0.97*	.16*/2.44E+12***	31.39*** (0.11/0.46)	15.51E+17*** (0.73/1.06)
Acceptance and Commitment Therapy for Mixed Anxiety Disorders	Modest	21/91/25/2[a]	0%***	0%***	.53*	.39*	0.55/0.62*	.002*/119.89***	79.58*** (0.32/1.14)	79.58*** (0.32/1.14)
Acceptance and Commitment Therapy for Psychosis	Modest	7/19/16/4	14%*	0%*	.50*	.57*	0.95/1.08*	.02*/19.62***	.07* (0.002/0.25)	189.64*** (0.48/1.47)
Applied Relaxation for Panic Disorder	Modest	6/17/6/2	0%^{z}***	0%***	.36*	.72**	0.85/0.96*	1.63*/3.03**^{x}	1.63* (0.08/1.74)	3.03**^{x} (0.19/2.01)
Assertive Community Treatment (ACT) for Schizophrenia	Strong	6/25/10/2	0%***	0%***	.79**	.58*	0.43/0.49**	.38*/3.21**^{x}	.38* (0.01/0.54)	3.21**^{x} (0.08/0.77)
Behavioral Activation for Depression	Strong	24/40/24/4	0%***	0%***	.15*	.00*	0.60/0.68*	.08*/4.03**^{x}	.23* (0.007/0.44)	.82* (0.02/0.61)
Behavioral and Cognitive Behavioral Therapy for Chronic Low Back Pain	Strong	12/27/27/6	0%***	0%***	.54*	.42*	0.74/0.84*	.12*/44.88***	.08* (0.002/0.23)	6248.62*** (0.44/1.07)

Behavioral Couples Therapy for Alcohol Use Disorders	Strong	10/25/14/6	0%***	0%***	.67**	.34*	1.10/1.26*	.17*/37.84***	5.03** (0.09/0.73)	428.09*** (0.33/0.98)
Behavioral Treatment for Obesity	Strong	0/2/2/2	--	--	--	--	0.19/0.21***	1.36E+32***/13.40E+29	1.36E+32*** (0.84/1.14)	70.90E+29*** (4.85/5.15)
Biofeedback-Based Treatments for Insomnia	Modest	15/27/24/6	0%*	53%*	.51*	.49*	1.16/1.32*	.005*/24.90***	.24* (0.006/0.54)	2352.77*** (0.78/1.87)
Cognitive Adaptation Training (CAT) for Schizophrenia	Modest	7/10/8/4	0%***	0%***	.77**	.83***	0.93/1.06*	3.06**[x]/73.47***	16.07** (0.26/1.24)	16004.36*** (0.87/1.87)
Cognitive and Behavioral Therapies for Generalized Anxiety Disorder	Strong	42/169/142/12	0%***	2%***	.95***	.90***	0.93/1.06*	.13*/9927.95***	.11* (0.002/0.27)	19.41E+18*** (1.33/1.91)
Cognitive Behavioral Analysis System of Psychotherapy for Depression	Strong	1/3/3/2	0%*	100%*	.89***	.78**	0.55/0.63*	.04*/45.37***	7.57** (0.15/1.09)	45.37*** (0.30/1.17)
Cognitive Behavioral Therapy (CBT) for Schizophrenia	Strong	12/28/_/2	8*%[y]	25%*	.72**	.77**	0.65/0.74*	--	.45* (0.02/0.91)	18.74*** (0.34/1.55)
Cognitive Behavioral Therapy for adult ADHD	Strong	14/39/23/6	0%**	7%**	.78**	.57*	0.69/0.78*	.72*/26.33***	36.06*** (0.18/0.70)	27948.16*** (0.43/0.93)
Cognitive Behavioral Therapy for Anorexia Nervosa	Modest/ Controversial	3/10/4/3	0%*	33%*	.06*	.12*	0.86/0.98*	.10*/29.06E+7***	NA	43.05*** (0.18/0.70)
Cognitive Behavioral Therapy for Binge Eating Disorder	Strong	20/43/18/4	0%***	0%***	.66**	.77**	0.56/0.63*	.08*/76.84***	.08* (0.002/0.26)	26.04*** (0.22/0.98)
Cognitive Behavioral Therapy for Bulimia Nervosa	Strong	19/51/22/6	0%**	16%**	.52*	.51*	0.67/.76*	.09*/135.52***	.07* (0.001/0.22)	6.72** (0.14/0.76)

Cognitive Behavioral Therapy for Chronic Headache	Strong	12/14/4/0	0%**	8.33%**	.54*	.25*	0.85/0.96*	.40*/2.15*	--	--
Cognitive Behavioral Therapy for Insomnia	Strong	0/3/0/0	--	--	--	--	0.72/0.82*	--	--	--
Cognitive Behavioral Therapy for Obsessive Compulsive Disorder	Strong	6/18/6/5	0%***	0%***	.97***	1.00***	0.64/0.72*	.44*/2059130.48***	58739.7*** (0.47/0.99))	38.42E+10*** (0.81/1.33)
Cognitive Behavioral Therapy for Panic Disorder	Strong	7/72/39/8	14%*	0%**	.53*	.48*	0.60/0.68*	.07*/72014349.23***	.51* (0.02/0.49)	39.93E+8*** (0.88/1.45)
Cognitive Behavioral Therapy for Social Anxiety Disorder	Strong	28/95/73/9	4%**	7%**	.61**	.53*	0.69/0.78*	.07*/1.32E+11***	.0002* (0.008/0.10)	37.22E+9*** (0.83/1.28)
Cognitive Processing Therapy for Post-Traumatic Stress Disorder	Strong	7/48/29/6	0%**	14%**	.78**	.84***	0.55/0.63*	.07*/5.47E+20***	14.06E+22*** (1.05/1.46)	11.00E+26*** (1.15/1.56)
Cognitive Remediation for Schizophrenia	Strong	19/43/20/4	16%*	36.84%*	.58*	.53*	0.62/0.71*	.004*/2.64*	.07* (0.001/0.21)	1.63* (0.05/0.73)
Cognitive Therapy (CT) for Bipolar Disorder	Modest	21/21/17/5	14%*	24%*	.39*	.36*	0.51/0.57*	.004*/96.04***	1.00* (0.02/0.42)	594.47*** (0.25/0.70)
Cognitive Therapy for Depression	Strong	31/55/11/4	0%***	2%***	.62**	.59*	0.66/0.75*	.06*/105.69***	.30* (0.006/0.34)	91.59*** (0.25/0.80)
Dialectical Behavior Therapy for Borderline Personality Disorder	Strong	5/15/14/4	20%*	40%*	.53*	.46*	0.77/0.87*	.12*/3.77**[x]	.11* (0.003/0.30)	4.40**[x] (0.09/0.75)
Emotion Focused Therapy for Depression	Modest	8/15/7/4	0%**	13%**	.54*	.33*	0.77/0.88*	.20*/8.72**	1.65* (0.05/0.78)	47.07*** (0.27/1.09)
Exposure and Response Pre-	Strong	6/6/6/3	0%***	0%***	.57*	.65**	0.67/0.76*	.14*/14.43***	.47*[a] (0.02/0.65)	.47*[a] (0.02/0.65)

vention for Obsessive-Compulsive Disorder										
Exposure Therapies for Specific Phobias	Strong	86/121/78/22	2%***[z]	0%**	.88***	.88***	1.05/1.20*	.14*/19.80E+7***	2878.03*** (0.26/0.70)	6.73E+36*** (1.70/2.15)
Eye Movement Desensitization and Reprocessing for Post-Traumatic Stress Disorder	Strong/ Controversial	7/26/19/6	0%*	29%*	.37*	.32*	1.07/1.23*	.12*/28287.90***	.17* (0.06/0.49)	28.28*** (0.78/1.74)
Family Focused Therapy (FFT) for Bipolar Disorder	Strong	19/22/11/4	0%**	5%**	.54*	.39*	0.69/0.79*	.18*/18.46***	71.29*** (0.22/0.83)	5064.40*** (0.39/1.00)
Family Psychoeducation for Schizophrenia	Strong	22/37/1/1	9%*[y]	0%*	.29*	.30*	0.85/0.96*	1.32*	1.32*[a] (0.04/1.06)	1.32*[a] (0.04/1.06)
Family-Based Treatment for Anorexia Nervosa	Strong	8/27/10/4	0%**	13%**	.58*	.54*	1.13/1.29*	.18*/6.23**	.36* (0.01/0.45)	17.86*** (0.17/0.78)
Family-Based Treatment for Bulimia Nervosa	Modest	0/8/0/0	--	--	--	--	0.54/0.62*	--	--	--
Friends Care for Mixed Substance Abuse/Dependence	Modest	0/37/0/0	--	--	--	--	0.43/0.49**	--	--	--
Guided Self-Change for Mixed Substance Abuse/Dependence	Modest	0/11/4/2	--	--	--	--	0.20/0.23***	.36*/3.15**[x]	.36* (0.01/0.64)	3.15**[x] (0.09/1.06)
Healthy-Weight Program for Bulimia Nervosa	Strong	14/28/16/4	0%***	0%***	.80***	.67**	0.36/0.41**	.10*/1229.18***	.08* (0.002/0.22)	6040.20*** (0.33/0.81)
Illness Management and Recovery (IMR) for Schizophrenia	Modest	0/0/0/0	--	--	--	--	--	--	--	--
Interpersonal and Social Rhythm Therapy	Modest	6/18/9/2	0%***	0%***	.52*	.38*	0.49/0.55**	.17*//69*	.45* (0.01/0.60)	.39* (0.02/0.67)

(IPSRT) for Bipolar Disorder									
Interpersonal Psychotherapy for Binge Eating Disorder	Strong	5/31/20/4	0%***	0%***	.25*	.10*	0.42/0.48**	.06*/638.05***	.13* (0.005/0.34)	13.59***[x] (0.19/0.78)
Interpersonal Psychotherapy for Bulimia Nervosa	Strong	20/40/7/1	0%***	0%***	.50*	.50*	0.75/0.85*	.08*/638.05***	638.05***[a] (0.83/2.11)	638.05***[a] (0.83/2.11)
Interpersonal Psychotherapy for Depression	Strong	0/61/21/8	--	--	--	--	0.72/0.82*	.03*/3.27E+11***	.03* (0.0009/0.10)	2.44* (0.19/0.58)
Mentalization-Based Treatment for Borderline Personality Disorder	Modest	3/16/11/5	0%***	0%***	.65**	.30*	0.79/0.90*	3.41**/450546.31***	955.19*** (0.31/0.85)	17821383*** (0.69/1.23)
Moderate Drinking for Alcohol Use Disorders	Modest	13/5/0/0	0%**	8%**	.56*	.50*	0.61/0.69*	--	--	--
Motivational Interviewing, Motivational Enhancement Therapy (MET), and MET plus CBT for Mixed Substance Abuse/Dependence	Strong	0/14/5/2	--	--	--	--	0.20/0.23***	.07*/1.22*	.14* (0.004/0.22)	1.22* (0.02/0.32)
Multi-Component Cognitive Behavioral Therapy for Fibromyalgia	Strong	1/13/11/2	0%***	0%***	.53*	.47*	0.98/1.12*	.14*/.86*	.17* (0.005/0.56)	.86* (0.03/1.18)
Multi-Component Cognitive Behavioral Therapy for Rheumatologic Pain	Strong	4/25/0/0	0%***	0%***	.36*	.22*	0.66/0.75*	--	--	--
Paradoxical Intention for Insomnia	Strong	11/14/7/2	0%***	0%***	.77**	.90***	0.87/.99*	.34*/22.08***	.33* (0.01/0.79)	22.08*** (0.41/1.75)
Present-Centered Therapy	Strong	12/18/10/6	0%**	8%**	.44*	.46*	0.78/0.88*	.07*/3.09E+31***	44.90E+19 (0.90/1.28)	46.87E+23 (0.98/1.34)

for Post-Traumatic Stress Disorder										
Prize-Based Contingency Management for Alcohol Use Disorders	Modest	4/4/0/0	0%***	0%***	.58*	.42*	0.78/0.89*	--	--	--
Prize-Based Contingency Management for Cocaine Dependence	Modest	12/39/17/13	0%**	17%**	.69**	.63**	0.55/0.61*	.07*/1.16E+28***	1135877.00 (0.26/0.52)	1.20E+12*** (0.42/0.69)
Prize-Based Contingency Management for Mixed Substance Abuse/Dependence	Strong	2/5/1/1	0%***	0%***	.75***	.50*	0.24/0.28***	1113.06***	1113.06***[a] (0.23/0.61)	1113.06***[a] (0.23/0.61)
Problem-Solving Therapy for Depression	Strong	18/45/36/11	0%***	0%***	.95***	.96***	0.56/0.64*	.007*/4.59E+27***	227.4*** (0.21/0.54)	2.36E+07*** (0.46/0.83)
Prolonged Exposure Therapy for Post-Traumatic Stress Disorder	Strong	33/62/37/12	0%**	18%**	.89***	.93***	0.71/0.80*	.01*/2.55E+11***	36.46*** (0.15, 0.52)	32999766*** (0.61/1.00)
Psychoanalytic Treatment for Panic Disorder	Modest/ Controversial	1/1/1/0	0%***	0%***	.88***	.75**	0.72/0.82*	37.41***	--	--
Psychoeducation for Bipolar Disorder	Strong (Mania) Modest (Depression)	12/42/4/2	0%***	0%***	.63**	.67**	0.56/0.64*	.65*/53.27***	.65* (0.02/0.66)	53.27*** (0.29/1.10)
Psychological Debriefing for Post-Traumatic Stress Disorder	No Evidence/ Potentially Harmful	0/0/0/0	--	--	--	--	--	--	--	--
Rational Emotive Behavioral Therapy for Depression	Modest	3/8/6/0	0%***	0%***	.15*	.00*	0.47/0.53**	.07*/9.52**	--	--
Relaxation Training for Insomnia	Strong	14/86/62/4	0%***	0%***	.66**	.39*	0.71/0.81*	.06*/135.08***	.11* (0.003/0.31)	567.29*** (0.42/1.19)

Reminiscence/Life Review Therapy for Depression	Modest	5/14/13/5	0%***	0%***	1.00***	1.00***	0.75/0.85*	.10*/1.15E+28***	53.58E+16*** (1.71/2.27)	46.73E+21*** (1.87/2.42)
Schema-Focused Therapy for Borderline Personality Disorder	Modest	0/1/1/0	--	--	--	--	0.54/0.61*	2.34E+24***	--	--
Seeking Safety for Mixed Substance Abuse/Dependence	Modest	3/45/16/2	0%*	33%*	.73**	.79**	1.01/1.49*	1.05*/509753.24***	1.05* (0.04/1.40)	509753.20*** (2.36/3.90)
Seeking Safety for PTSD with Substance Use Disorder	Strong	15/97/37/4	0%**	6%**	.41*	.48*	0.89/1.01*	.14*/509753.24***	.16* (0.004/0.45)	1187.52*** (1.00/1.95)
Self-Management/Self-Control Therapy for Depression	Strong	42/75/48/11	5%**[y]	5%**	.55*	.51*	1.13/1.29*	.11*/63.52***	.05* (0.001/0.15)	1058.71*** (0.33/0.82)
Self-System Therapy for Depression	Modest	2/0/9/0	0%***	0%***	.22*	.45*	0.75/0.86*	.32*/.79*	--	--
Short-Term Psychodynamic Therapy for Depression	Modest	0/2/2/2	--	--	--	--	0.42/0.47**	.06*/.10*	.06* (0.001/0.20)	.09* (0.002/0.25)
Sleep Restriction Therapy for Insomnia	Strong	6/74/74/4	0%***	0%***	.59*	.19*	0.74/0.84*	.007*/19.39***	.0004* (0.0009/0.11)	7.41** (0.16/1.07)
Smoking Cessation with Weight Gain Prevention	Modest	13/34/8/2	0%**	13%**	.10*	.13*	0.34/0.39**	.05*/15.45***	.06* (0.002/0.19)	15.54***[x] (0.14/0.67)
Social Learning/Token Economy Programs for Schizophrenia	Strong	0/0/0/0	--	--	--	--	--	e--	--	--
Social Skills Training (SST) for Schizophrenia	Strong	34/36/1/1	0%**	6%**	.59*	.48*	0.97/1.10*	1.37*	1.37*[a] (0.08/1.95)	1.37*[a] (0.08/1.95)
Stimulus Control Therapy for Insomnia	Strong	9/61/49/7	22%*	0%*	.58*	.56*	1.16/1.32*	.0009*/244.81***	.11* (0.002/0.29)	66071.82*** (0.95/1.87)

Stress Inoculation Training for Post-Traumatic Stress Disorder	Modest	18/41/25/3	0%***	0%***	.55*	.49*	1.04/1.18*	.12*/82.42***	3.31**[x] (0.10/1.10)	262.06*** (0.56/1.55)
Supported Employment for Schizophrenia	Strong	32/38/14/4	0%***	0%***	.75**	.72**	0.33/0.38**	2.30*/31855.54***	37.12*** (0.15/0.60)	1276346*** (0.46/0.92)
Systematic Care for Bipolar Disorder	Strong (Mania)	4/3/0/0	0%***	0%***	.58*	.41*	0.24/0.28***	--	--	--
Transference-Focused Therapy for Borderline Personality Disorder	Strong/ Controversial	3/5/1/1	0%***	0%***	.59*	.53*	0.49/0.55**	.29*	.29*[a] (0.01/0.50)	.29*[a] (0.01/0.50)

Note. Number of usable effects (p-checker/pwr/BayesFactor (individual)/BayesFactor (meta). [z]Results initially flagged as misreported were confirmed to be run as one-tailed tests. [y]Effect described as marginally significant was misreported. [x]Bayes Factor indicates weaker threshold of evidence if wider prior is adopted. [a]Same effects used for pessimistic/optimistic meta-analytic estimates.
NA: Not available due to estimation problems. *Controversial. **Modest. ***Strong (based upon our suggested thresholds).

Table 3

Evidential Value Metrics for Empirically Supported Treatments Grouped By Div. 12 Strength of Evidence Category

Div. 12 Evidence Category	# of Usable Effects	% Gross Errors	% Minor Errors	Median Post Hoc Power	R-Index	80% Power to detect Median *d* (1-tailed/2-tailed)	BF_{10} Range	Pessimistic Meta BF_{10} (Posterior 95% CR: LL_d/UL_d)	Optimistic Meta BF_{10} (Posterior 95% CR: LL_d/UL_d)
Modest	182/535/267/45	2%**	10%**	.53*	.48*	0.71/0.80*	.005*/1.16E+28***	12.03E+10*** (0.22/0.36)	47.45E+61*** (0.63/0.77)
Strong	523/1873/824/	2%**	7%**	.64**	.67**	0.73/0.84*	.002*/30.85E+30***	8.65** (0.02/0.13)	81.60E+62*** (0.45/0.56)
Modest (Controversial)	4/11/5/0	25*	0%*	.26*	.27*	0.86/0.98*	.02*/29.07E+7***	--	--
Strong (Controversial)	10/31/20/3	0%**	10%**	.50*	.51*	1.07/1.23*	.14*/28.29E+3***	.13* (0.004/0.31)	.19* (0.005/0.40)

Note. Number of usable effects (p-checker/pwr/BayesFactor (individual)/BayesFactor (meta).
*Controversial. **Modest. ***Strong (based upon our suggested thresholds).

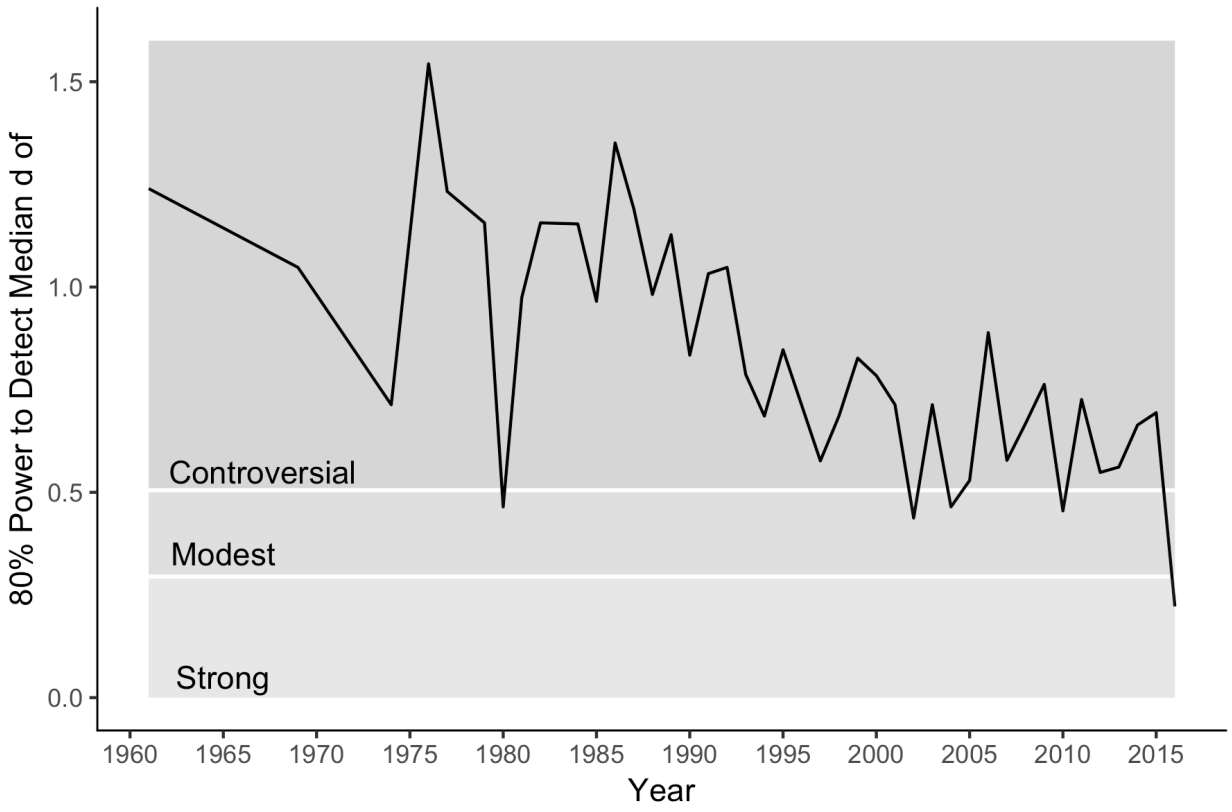

Figure 1. Median smallest pairwise difference (*d*) between EST and control that could be reliably detected (80% power), by year of publication time. Shaded regions correspond to our proposed cut-offs for levels of evidential value.