Received: 1 April 2019 | Revised: 3 February 2020 | Accepted: 5 February 2020

DOI: 10.1111/cpsp.12331

EMPIRICAL

CLINICAL PSYCHOLOGY SCIENCE AND PRACTICE WILEY

Potentially harmful therapies: A meta-scientific review of evidential value

Alexander J. Williams[1] | **Yevgeny Botanov**[2] | **Robyn E. Kilshaw**[3] | **Ryan E. Wong**[4] | **John Kitchener Sakaluk**[4]

[1]University of Kansas, Lawrence, KS, USA

[2]Department of Psychology, Pennsylvania State University – York, York, PA, USA

[3]University of Utah, Salt Lake City, UT, USA

[4]Department of Psychology, University of Victoria, Victoria, BC, Canada

Correspondence
Alexander J. Williams, Psychology Department, University of Kansas, Regnier Hall 370B, 12610 Quivira Rd., Overland Park, KS 66213, USA.
Email: alexwilliams@ku.edu

Funding information
SSHRC Doctoral Fellowship; SSHRC Insight Development Grant

Abstract

Lilienfeld (2007, Psychological treatments that cause harm. Perspectives on Psychological Science, 2, 53) identified a list of potentially harmful therapies (PHTs). Given concerns regarding the replicability of scientific findings, we conducted a meta-scientific review of Lilienfeld's PHTs to determine the evidential strength for harm. We evaluated the extent to which effects used as evidence of harm were as follows: (a) (in)correctly reported; (b) well-powered; (c) statistically significant at an inflated rate given their power; and (d) stronger compared with null effects of ineffectiveness or evidence of benefit, based on a Bayesian index of evidence. We found evidence of harm from some PHTs, though most metrics were ambiguous. To enhance provision of ethical and science-based care, a comprehensive reexamination of what constitutes evidence for claims of harm is necessary.

KEYWORDS

evidential value, metascience, potentially harmful therapies, replicability, reproducibility

1 | INTRODUCTION

> *I am a sensitive observer, and my conclusion is that a vast majority of my patients get better as opposed to worse after my treatment.*
>
> *Prefrontal lobotomist, sometimes attributed to Walter Freeman (quoted in Dawes, 1994, p. 48)*

The mid-20th century description of lobotomies illustrates the dangers of poorly researched mental health treatments. Consequently, the paramount ethical principles for mental health practitioners and researchers (e.g., American Psychological Association [APA], 2002) are beneficence and nonmaleficence. Clinical practitioners and researchers must do good (beneficence) while also doing no harm (nonmaleficence). Understandably, the bulk of psychological intervention research has examined beneficence through efficacy and effectiveness research (e.g., Cuijpers et al., 2013; Smith & Glass, 1977; Stewart & Chambless, 2009; Wampold & Imel, 2015). Unfortunately, research on psychological interventions that are potentially maleficent or harmful has lagged behind.

Questions about potential harm from psychotherapies date back more than half a century (e.g., Eysenck, 1952). Mental health treatments have long been known to be potentially and severely harmful (e.g., Josefson, 2001). However, not until more recently (Lilienfeld, 2007) have treatments been delineated into a provisional list of *potentially harmful therapies* (PHTs). The need to better understand PHTs continues to grow as popular media erroneously critiques well-established treatments as potentially harmful (e.g., exposure therapies; Morris, 2015), while treatments accepted as harmful (e.g.,

reparative/conversion therapy; APA, 2009) continue to be made available (Mallory, Brown, & Conron, 2018).

Although Lilienfeld's (2007) review of PHTs was provocative, the emergence of the replicability movement (Open Science Collaboration, 2015) and its extensions to clinical psychological science (Tackett, Brandes, King, & Markon, 2018; Tackett et al., 2017) highlight the need to reevaluate the credibility of clinical literatures. It is particularly important to vet claims of harm (or efficacy) in an era where concerns about *p*-hacking and underpowered studies loom large (Nelson, Simmons, & Simonsohn, 2018). In the current article, we draw upon multiple methods popularized during the replicability movement for evaluating the strength of evidence in scientific literatures. We produce a synthesis of evidential value—broadly defined—for the claims of potential harm from therapies identified by Lilienfeld.

Public Health Significance

Psychological interventions designed to help people sometimes inadvertently harm them instead. In our examination—incorporating more than 70 reports—of treatments previously identified as potentially harmful, we found that the clinical trials often provided weak scientific evidence and are therefore difficult to interpret. However, some interventions showed stronger evidence for harm and only grief therapy showed promise of benefit; as such, the remaining treatments we examined require more compelling, reproducible, and replicable evidence of benefit to justify continued clinical use.

1.1 | Potentially harmful therapies

Lilienfeld (2007) highlights the difficulties in identifying criteria for harmful therapies, and the debate as to how to best do so continues to this day (e.g., Jarrett, 2008; Linden, 2013). To minimize threats to internal validity and increase our confidence that a PHT may be leading to harm, the current review only examined PHTs that Lilienfeld identified using evidence from randomized controlled trials (RCTs). Consequently, facilitated communication, attachment therapies, recovered-memory techniques, DID-oriented therapy, peer-group interventions of conduct disorder, and relaxation treatments for panic-prone individuals were not included in the current review, while the following PHTs were included:

1.1.1 | Boot camp for conduct disorder

Boot camps are modeled after military basic training and are designed to decrease disruptive or illegal behavior that is associated with conduct disorder. Despite a decrease in popularity since the 1990s (Meade & Steiner, 2010), boot camp programs are still available in some states.

1.1.2 | Critical Incident Stress Debriefing (CISD)

CISD was developed (Mitchell, 1983) as a preventative intervention to assist emergency responders exposed to a severe stressor. Debriefing (not to be confused with the ethical research procedure), psychological debriefing, and critical incident response are similar intervention strategies to CISD. Implicit to the theoretical basis of CISD and other forms of debriefing is that nearly all individuals exposed to a severe stressor can benefit from an intervention shortly thereafter. CISD proponents contend that all of the research demonstrating harm (for reviews, see Gist, 2015; Lohr, Gist, Deacon, Devilly, & Varker, 2015; Rose, Bisson, Churchill, & Wessely, 2002) is due to poor clinician training, erroneous application of the intervention to populations other than emergency responders, and using the intervention with individuals rather than small groups (Everly & Mitchell, 2000; Mitchell, n.d.). CISD continues to be an active and available intervention option with a training organization and professional meetings (e.g., see https://icisf.org/).

1.1.3 | DARE

DARE (Drug Abuse Resistance Education) and similar efforts (e.g., Sloboda et al., 2009) aim to limit underage substance use. In the original DARE curriculum, a uniformed police officer taught students about the perils of drug/alcohol use. The program has been modified in recent years in an attempt to make it more interactive and closely aligned with research-based prevention methods (Caputi & McLellan, 2017). DARE claims to be present in 75% of US school districts and over 50 countries (About DARE, 2019).

1.1.4 | Expressive-experiential psychotherapies

Lilienfeld (2007) identified a traditional yet diverse set of talk therapies (as opposed to art-, music-, or drama-based expressive therapies; Malchiodi, 2013) that have a shared mechanism of therapeutic effect: the re-experiencing and heightening of strong emotions (e.g., anger). Intensifying

the experience of painful emotions (e.g., anger is increased and prolonged) may serve as a possible source of harm (Moos, 2005).

1.1.5 | Grief counseling

Grief counseling consists of a broad group of psychotherapies designed to assist individuals in coping after the death of a close social contact. Despite concerns that grief counseling may subvert the normal grieving process into a complicated or pathological one (see Bonanno & Lilienfeld, 2008; Larson & Hoyt, 2007 for continued debate), training and continuing education in grief counseling is commonplace.

1.1.6 | Scared Straight

With the goal of deterring future criminal behavior, Scared Straight interventions (made famous through the documentary *Scared Straight*!; Shapiro, 1978) involve at-risk juveniles being confronted and intimidated by inmates who describe the dangers of prison life (e.g., violence, sexual assault, abuse). Evidence suggests that these programs may in fact increase rates of recidivism (Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013), yet Scared Straight programs continue to be used as indicated by the release of the television series *Beyond Scared Straight* (Shapiro & Coyne, 2011).

1.2 | The replicability movement and clinical psychology

Replication of iatrogenic effects by independent research teams is central to Lilienfeld's (2007) recommendations for establishing harm. Although replication is an ostensible cornerstone value of psychological science, psychologists across subdisciplines have been forced to grapple with ascribing meaning to replication failures of both individual (e.g., Cheung et al., 2016; O'Donnell et al., 2018; Wagenmakers et al., 2016) and collections of effects (Klein et al., 2014; Open Science Collaboration, 2015) by large-scale collaborative research efforts. This startling dearth of replicability (see Nelson et al., 2018; Spellman, 2015, for reviews) has left psychological scientists questioning much of what they thought they knew about best practices for study planning (Nosek, Ebersole, DeHaven, & Mellor, 2018; Schönbrodt & Perugini, 2013), psychological measurement (Flake & Fried, 2019; Hussey & Hughes, 2018; Sakaluk, 2019), data analysis (Benjamin et al., 2018; Cumming, 2014; John, Loewenstein, & Prelec, 2012; Lakens et al., 2018; Simmons, Nelson, & Simonsohn, 2011), sharing research materials (Meyer, 2018; Nosek, Spies, & Motyl, 2012; Soderberg, 2018), synthesizing literatures (Carter, Schönbrodt, Gervais, & Hilgard, 2017; Rouder & Morey, 2011), and even academic writing (Gernsbacher, 2018).

The replicability movement discourse has now pervaded clinical psychological science (Tackett et al., 2017, 2018; Walsh et al., 2018). That clinical research might also possess limited replicability should, on its face, not be a terribly surprising possibility. Indeed, many of the same factors argued to underlie low replicability in other areas of psychology (e.g., low statistical power; Maxwell, 2004) have a long history of being acknowledged—and ignored—in clinical psychological science (Cohen, 1962). The replicability of clinical research might also be challenged by unique subdisciplinary factors, such as the burden of recruiting samples for psychotherapy trials and tightly accredited training curricula often leaving little in the way of time or encouragement for trainees to invest in deeper levels of methodological and analytical acumen (King, Pullmann, Lyon, Dorsey, & Lewis, 2018). The possibility, if not likelihood, of unreplicable findings in clinical research is all the more concerning in light of clinical science being brought to bear on the everyday lives of consumers, health-care providers, managed care administrators, and policy makers seeking evidence-based treatments.

A crux of the looming replicability issue in clinical research is the over-reliance on null hypothesis significance testing (NHST). $p < .05$, at the best of times, is a relatively low bar of evidence (Benjamin et al., 2018) for a couple of research teams to surpass in order to claim efficacy (Chambless et al., 1998) or harm (Lilienfeld, 2007); this is especially true when the "rules" of NHST are not followed, which can lead to dramatically inflated rates of false-positive findings (Nelson et al., 2018; Simmons et al., 2011). When coupled with a system of publication that fetishizes claims of a "significant" effect (Rosenthal, 1979)—to the disparagement of null effects—one could easily understand how a literature could become overrun with unreplicable findings. Clinical literatures therefore must be (re)appraised through a broader lens of evidential value, beyond an exclusive reliance on the $p < .05$ threshold. Researchers must not only consider whether a claim is supported by an effect that is "significant" or null, but also how believable, credible, or otherwise reasonable the claims are in light of other features of the research design and resulting data.

In a recent meta-scientific review, Sakaluk, Williams, Kilshaw, and Rhyner (2019) broadly evaluated the credibility of the clinical trials underpinning psychological interventions labeled "Empirically Supported Treatments" (ESTs). Often billed as encapsulating gold-standard psychological treatments (e.g., exposure therapy) for particular diagnoses (e.g., specific phobias), therapies deemed ESTs were traditionally required to meet a set of criteria demonstrating efficacy (Chambless et al., 1998) similar to Lilienfeld's (2007) criteria for harm (though see Tolin, McKay, Forman, Klonsky, & Thombs, 2015, for

recommended new EST criteria). Despite the high evidential value suggested by the labeling of these treatments, Sakaluk et al. (2019) found that research reports of many ESTs—including a number categorized as possessing particularly "strong" evidence by traditional criteria—were contaminated by reporting errors, imprecise studies, inflated rates of statistically significant effects given a particular study's precision, and ambiguous efficacy compared with a null effect. With more than 50% of the articles on ESTs they investigated performing poorly across most metrics of evidential value, Sakaluk et al.'s findings call into question what is meant by "empirically supported" when describing psychological interventions.

1.3 | The current meta-scientific review

The discourse of replicability in clinical psychological science (Tackett et al., 2017, 2018; Walsh et al., 2018) highlights the need to (re)evaluate the evidential value underlying clinical intervention trials. Strong evidence, however, is needed to legitimize not only claims of benefit via therapeutic efficacy (Sakaluk et al., 2019), but also claims of (non)maleficence via therapeutic harm (or the lack thereof). In the current meta-scientific review, we assessed the evidential value of six interventions identified by Lilienfeld (2007) as PHTs. As in Sakaluk and colleagues' review (2019), we adopt a broader conceptualization of evidential value beyond *p*-values/statistical significance. Specifically, we consider rates of misreporting, estimates of statistical power, R-Index values, Bayes factors, and posterior estimates of effect size, in order to assess the evidential value for harm from each PHT.

2 | METHOD

Below, we describe the methods for determining our sample of relevant reports, effects within reports, coding strategy, and analytic approach. Given the breadth and complexity of large meta-scientific syntheses like these, we have endeavored to make all features of our investigation transparent and reproducible whenever possible. All our decision-making documents, coded data frames, and analytic materials are available on our Open Science Framework project (https://osf.io/5g6m7/). Additionally, the majority of our inclusion/exclusion criteria and coding analytic strategies were preregistered prior to analysis, and changes to this protocol were documented throughout the duration of the project.

2.1 | Sample of reports and effects

In keeping with the previous meta-scientific synthesis of clinical intervention literature (Sakaluk et al., 2019), we used an externally determined database to dictate the scope of our review. Namely, we initially relied upon Lilienfeld's (2007) references when selecting the studies to include in our sample, and the harmful outcomes he identified when deciding the effects to code. We then expanded the scope of the review by including trials evaluated in 26 other reviews of PHTs (including many systematic reviews), 23 of which were published after Lilienfeld's review (for these reviews, see Barnett & Howard, 2018; Birur, Moore, & Davis, 2017; Caputi & McLellan, 2017; Faggiano, Minozzi, Versino, & Buscemi, 2014; Faggiano et al., 2005; Flynn, Falco, & Hocini, 2013; Forneris et al., 2013; de Graaf, Honig, Pampus, & Stramrood, 2018; Horn & Feder, 2018; Howlett & Stein, 2016; Joyce et al., 2016; Kearns, Ressler, Zatzick, & Rothbaum, 2012; Kramer & Landolt, 2011; Lapp, Agbokou, Peretti, & Ferreri, 2010; Meade & Steiner, 2010; Pack, 2013; Pan & Bai, 2009; Petrosino et al., 2013; Qi, Gevonden, & Shalev, 2016; Rose et al., 2002; Rosner, Kruse, & Hagl, 2010; Skeffington, Rees, & Kane, 2013; Welsh & Rocque, 2014; West & O'Neal, 2004; Wethington et al., 2008; Wittouck, Van Autreve, De Jaegere, Portzky, & Heeringen, 2011).

We extracted all potential RCTs (and RCTs from meta-analyses) cited in Lilienfeld (2007) as well as all RCTs related to PHTs cited in the other reviews. This included reports with randomization at levels other than that of the individual (e.g., classroom). Our efforts yielded 110 potential reports for inclusion in our review (21 from non–peer-reviewed sources). Twenty-one reports were then excluded prior to coding for the following reasons: determined to be something other than a prospectively randomized trial (quasi-experimental study, literature review, mediator study, etc.) (8); lacked relevant PHT outcomes (4); study of something other than a PHT (4); duplicates (3); could not be identified after reviewing the article and references and contacting the authors (1); and unpublished report that could not be obtained from interlibrary loan requests, requests to colleagues, and attempts to contact the authors (1). Of the remaining 89 reports, 14 were ruled ineligible during the coding process (e.g., contained no group comparison or did not provide outcome data related to a PHT). Ultimately, effects from 75 reports were included in our analyses. We deemed effects eligible if they were considered relevant for potential harms in Lilienfeld (2007; Table 1, p. 58, and additional harms specified within the text). We also included other potential harms if they had strong face validity (e.g., worsened grief as a potential harm from grief counseling).

2.2 | Coding strategy

The last author, who was primarily responsible for the analyses, divided the reports among the other four authors, who carried out the coding. Each article was independently reviewed by two authors, and all coding disputes were resolved via an archived

TABLE 1 Potential harms coded

Intervention	Potential harms
Critical incident stress debriefing	Heightened risk for posttraumatic symptoms
	Heightened risk for anxiety symptoms
Scared straight interventions	Exacerbation of conduct problems
	Increased probability of offending
Grief counseling	Increases in depressive symptoms
	Increases in grief
Expressive-experiential psychotherapies	Exacerbation of painful emotions
	Increased anger
Boot camp interventions for conduct disorder	Exacerbation of conduct problems
	Increased recidivism
	Increased antisocial behavior
DARE programs	Increased intake of alcohol and other substances (e.g., cigarettes)

Note: Potential harms from Table 1 (p. 58) and text of Lilienfeld (2007).

Slack channel. Any discrepancies were initially addressed by the two authors assigned to a given article. If they could not resolve the dispute, the other authors were invited to help reach a consensus decision. Additionally, we coded study design features (e.g., type of control); however, there was little variability in these features. Thus, the bulk of our analyses focused on the remainder of the content we coded, which corresponded to the statistical elements reported in each study (e.g., sample size(s), test statistic types and values, degrees of freedom, *p*-values).

2.3 | Analytic strategy

As in Sakaluk et al. (2019), we did not consider any one metric of evidential value as the absolute arbiter of truth for a claim of harm or efficacy, but rather, we deemed consistency across metrics as providing stronger evidence for (or against) an intervention's harm (see Hu & Bentler, 1999; Mayo, 2018; Ruscio, Haslam, & Ruscio, 2006, for similar sentiments). We also drew on the heuristic cutoffs for each metric described by Sakaluk et al. (2019) as a reasonable starting place for categorizing and contextualizing the evidential value of the PHT literature.

2.3.1 | Misreporting

First, we evaluated the extent that inferential statistics were misreported (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2016) for each PHT. As misreported or insufficiently reported statistics cannot be verified, we propose that higher levels of misreporting constitute weaker evidence for/against an intervention's harm than do effects that can be verified as accurately reported. We appraised the rates of gross (affecting claims of statistical significance) and minor (not affecting claims of statistical significance) misreporting using Schönbrodt's *p*-checker application (2018); when present, the last author manually verified effects that were flagged as gross errors.

2.3.2 | Power-related indexes

We also calculated two different metrics of statistical power for studies to detect effects of harm (or efficacy). First, for pairwise comparisons between treatment and control conditions at a given measurement point (e.g., post-treatment), we calculated the median smallest effect size (Cohen's *d*) that could be reliably detected at 80% power given a study's reported sample size. We propose that well-powered studies that can reliably detect reasonable and clinically appropriate effect sizes constitute stronger evidence than low-powered studies that can only detect large and implausible effects. We calculated two variations of this metric—a one-tailed version (testing harm only) and a two-tailed version (testing harm or efficacy) using the *pwr* package (Champely, 2018) for *R* (R Core Team, 2018).

The second power-related index of evidential value we calculated was Schimmack's Replicability Index (R-Index; 2016). The R-Index estimates and then compares the median observed power for a set of studies against the rate of statistical significance for that same set of effects. These two rates should be equal in the long run, so rates of significance in excess of typical power levels is taken as a metric of inflation. The final R-Index is then computed as the difference between median observed power and the inflation rate, effectively attempting to correct or penalize sets of studies for excessive significance (see also Schimmack, 2012). We submit that studies that are either typically low-powered or contain excessive rates of significance constitute weaker statistical evidence than do studies with good power and with appropriate levels of statistical significance given their power. We estimated R-Index using Schönbrodt's *p*-checker application (2018).

2.3.3 | Bayes factors

The final metrics of evidential value we calculated within a given PHT were individual and meta-analyzed Bayes factors (Rouder & Morey, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). These metrics estimate the relative probabilities of two hypotheses (i.e., harm vs. no effect/efficacy) and were computed using a noninformative prior (i.e., the Jeffreys–Zellner–Siow, or JZS, prior) to allow the data to determine the range of possible probabilities. Similar to

Sakaluk et al. (2019), we meta-analyzed both "optimistic" effect selections (i.e., in the direction of efficacy or neutrality) and "pessimistic" effect selections (i.e., in the direction of harm). This provided independent sets of effects across samples as well as estimates covering the range of what might be reasonable to infer about a particular PHT given particular assumptions. We also estimated meta-analytic posterior distributions for the effect of harm based on these competing samples of effects. These analyses were performed twice: first using what could be considered a "one-tailed" assessment of Bayes factors with a hypothesized interval of effects that prioritized the assessment of harm for a particular PHT, and second using a "two-tailed" assessment of Bayes factors with a hypothesized interval of effects that left open the possibility of harm or efficacy for a particular PHT. We see these separate analyses as speaking to two different questions of evidential value: the first about the evidential value underlying Lilienfeld's (2007) original claims of harm, and the second about an effect of harm (a possibility) or an effect of efficacy (the original intention of the intervention).

We submit that studies characterized by Bayes factors that are ambiguous with respect to harm or efficacy/equivalence (see Jeffreys, 1961) constitute weaker evidence than do studies characterized by stronger Bayes factors in favor of harm or efficacy/equivalence. Moreover, posterior distributions of effect sizes for a PHT that are concentrated in an ambiguous range of effect size (e.g., $-0.10 < d < 0.10$) constitute weaker evidence for a PHT than do posterior distributions with the bulk of their mass in an informative range of effect sizes (e.g., $d < -0.10$, $d > 0.10$). All of these analyses were performed using the *BayesFactor* package (Morey & Rouder, 2018) for *R* (R Core Team, 2018), which required us to use *t*-statistics for pairwise comparisons between PHTs and control conditions at post-treatment and follow-up waves of an RCT.

2.3.4 | Rationale for metric selection

Multiple considerations guided our selection of misreporting rates, power-related indexes, and Bayes factors as metrics of evidential value, in lieu of other systems of determining evidential value (e.g., Chambless & Hollon, 1998; Guyatt et al., 2011; Guyatt et al., 2008). Early methods to examine evidential value of psychotherapies involved appraising whether statistical comparisons of treatment versus control reached $p < .05$ in two independent RCTs (see, e.g., Lilienfeld's (2007) "probable harm" classification for PHTs and Chambless and Hollon's (1998) "efficacious" designations for ESTs). Although a clearly positive, watershed development in terms of increasing the scientific rigor with which psychotherapies were evaluated (Tolin et al., 2015), subsequent methodological scholarship has demonstrated that $p < .05$ is a relatively weak evidentiary threshold, for which the utility can be further (and easily) undermined through the use of relatively commonplace statistical practices (John et al., 2012; Simmon et al., 2011). In comparison, our selection of evidential value metrics ensures that (a) no one metric is exalted to serve as *the* metric of evidential value (unlike $p < .05$), and (b) we can take into consideration the degree to which the rules of inference with *p*-values may have been compromised (vis-a-vis *R*-Index's correction for inflated rates of significance).

In recent years, the appraisal of the evidential value of ESTs has shifted to new systems of increasing complexity and rigor (Tolin et al., 2015), such as the Grading of Recommendations, Assessment, Development, and Evaluations (GRADE) approach (Guyatt et al., 2008). However, proponents of GRADE acknowledge that it involves a high degree of subjectivity in its application, as "[t]wo persons evaluating the same body of evidence might reasonably come to different conclusions about its certainty" (Siemieniuk & Guyatt, 2019, para. 5), a feature that may threaten its reproducibility. By comparison, our metrics of evidential value, when applied to the same body of evidence (i.e., sample of effects), are fully reproducible in their output. That is, although two hypothetical reviewers might disagree to what degree a certain level of misreporting, power, *R*-Index, or Bayes factor is desirable or problematic, our approach would at least ensure that the hypothetical reviewers would derive the same numbers. Moreover, for many of our metrics, there are consensus (Cohen, 1962; Jeffreys, 1961) or expert opinions (e.g., the consultants who contributed to Sakaluk et al., 2019) to inform qualitative interpretations of evidential value (e.g., concluding evidence is "Strong" for power > 0.80). For GRADE, however, the two reviewers could render substantively different classifications (e.g., "Low" and "High") based on the application of the same system.

A final consideration motivating our selection of evidential value metrics is that they are, for the most part, intuitive in their interpretation. GRADE, in contrast, relies on the extraction and interpretation of more cumbersome metrics. Namely, when appraising *Imprecision* (Guyatt et al., 2011), reviewers applying GRADE are encouraged to interpret confidence interval width as a signal of evidential value, with narrower CIs being characterized as stronger evidence. Although an exhaustive explication of our concerns with this approach to appraising evidential value via examining effect (im)precision is beyond the scope of the current review, we briefly note the ample evidence that confidence intervals (a) are often mistakenly interpreted and poorly implemented (e.g., Belia, Fidler, Williams, & Cumming, 2005; Hoekstra, Morey, Rouder, & Wagenmakers, 2014); (b) are subject to having their meaning eroded by the very same violations of frequentist inference rules that commonly compromise *p*-value-reliant inference (e.g., Sakaluk, 2016); and (c) contain, in their width, more information than merely their estimation

precision, but also capture some degree of true heterogeneity in effect sizes (Kenny & Judd, 2019). By comparison, our selection of rates of misreporting, statistical power, etc., seem much less likely to be misinterpreted or misapplied by the average reader.

3 | RESULTS

In total, we identified 562 effects that met our inclusion criteria for providing evidence for/against harm for boot camp interventions (*number of effects* = 31), CISD (*number of effects* = 100), DARE (*number of effects* = 243), expressive-experiential psychotherapies (*number of effects* = 6), grief counseling (*number of effects* = 131), and Scared Straight interventions (*number of effects* = 51). Apart from some of the effects from boot camp interventions and DARE, the bulk of these effects were from comparisons of treatment against no-treatment control groups (see Figure 1). Consistency of coding across our reviewers was excellent, as computing our metrics using data from either the first or second reviewer produced relatively trivial differences (see https://osf.io/nbc5h/ for comparable tables of metrics).

Evidential value metrics for each PHT are listed in Table 2. Most of the effects described in the PHT literature were not sufficiently reproducible to allow for calculation of our evidential value metrics. In other words, the majority of studies included the outcome of a statistical test comparing the PHT to a control condition (e.g., "statistically significant," "$p < .05$"), but otherwise did not report a sufficient amount of accompanying information (e.g., omitting descriptive statistics, test statistics, degrees of freedom) to allow us to verify accurate reporting, estimate power, or compute Bayes factors.

3.1 | Evidential value of individual PHTs

With few exceptions, there was little evidence of misreported statistical tests across PHTs. However, the low rates were largely a byproduct of insufficient reporting of statistics, which undermined our ability to verify the (in)accuracy of the reporting. In the case of DARE, for example, of the 127 potentially verifiable effects, only six tests (5%) described the necessary statistical elements needed for stat-checking. For expressive-experiential therapies, the state of reproducible reporting was so low that none of the included tests could be evaluated.

A pattern of consistently low *R-Index* values across PHTs was also evident, indicating that tests of PHTs were either chronically underpowered, characterized by appreciable degrees of inflation of statistical significance, or both. Our analyses of the typical effect sizes that could be detected within the literatures of each PHT tell a more nuanced—and potentially more reliable (given the greater availability of usable effects)—story. Specifically, studies assessing boot camp interventions and DARE appear to have been designed, on

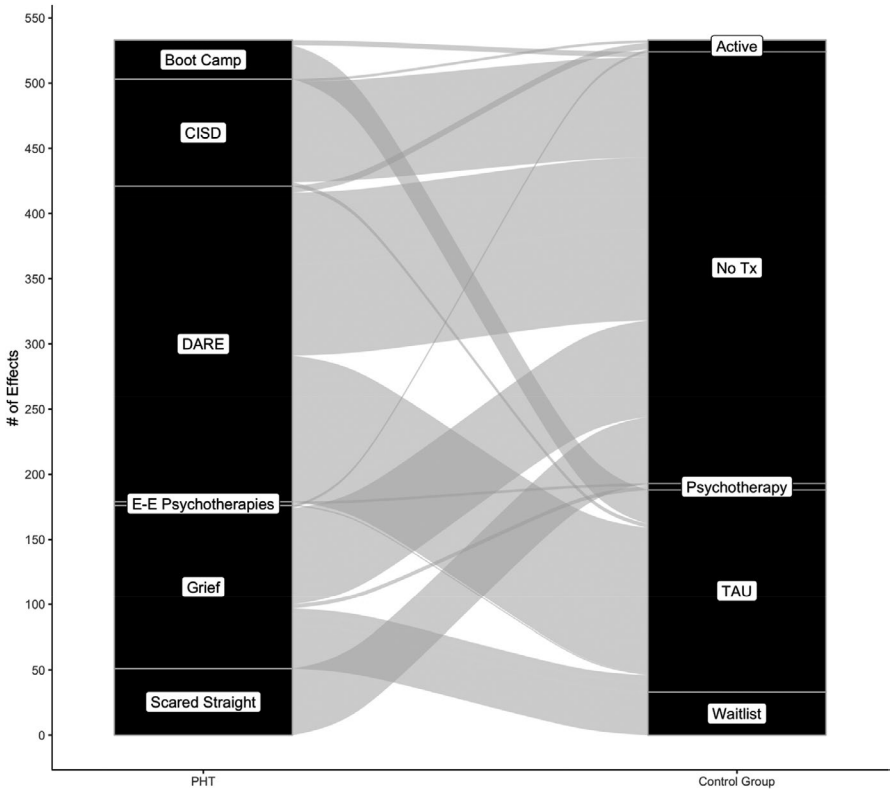

FIGURE 1 Alluvial plot (Brunson, 2018) of effects coded for each potentially harmful therapy and the control group used in the statistical comparison. Plot depicts the number of effects coded for each PHT and to what extent each type of control was used for that PHT portion of effects in the total sample. DARE, for example, yielded the largest collection of effects, which utilized mostly No Tx or TAU controls, with a select number of comparisons with Active controls

TABLE 2 Evidential value statistics for potentially harmful therapies

PHT	# of Effects (Total)	Reproducibility-friendly tests	% Gross reported	% Minor reported	R-Index	# of Usable effects (Power)	80% power to detect *d* (1-tailed)	80% power to detect *d* (2-tailed)	BF_{10} range	# of Usable effects (BF_{10} Meta)	BF_{10} meta Optim.	BF_{10} meta Pess.
Boot camp interventions for conduct disorder	31	4/18	0	0	0.23	28	−0.245	0.276	–	0	–	–
CISD	100	8/26	12.50	0	0.28	60	−0.469	0.53	0.062–17.902	7/7	0.058	8.575
DARE programs	243	6/127	0	17	0.43	120	−0.141	0.158	0.024–2.247	4/4	0.013	0.468
Expressive-experiential psychotherapies	6	0/5	–	–	–	4	−0.543	0.614	–	0	–	–
Grief counseling for normal bereavement	131	11/82	0	0	0.48	108	−0.739	0.838	0.044–1.96	10/10	0.016	0.042
Scared Straight interventions	51	12/32	0	0	0.00	47	−0.555	0.628	0.085–9.509	2/2	0.228	4.496

Note: Reproducible-Friendly = fully reported p-checkable tests/ostensibly p-checkable tests. Gross reporting errors = misreports impacting claims of statistical significance. Minor reporting errors = misreports not impacting claims of statistical significance. Dividing 1 by $BF_{10} < 1$ will render BF_{01} in favor of no harm.

average, to reliably detect modest (and therefore reasonable) psychological effects (~$d = 0.15$–0.25) of efficacy or harm at posttest or follow-up. Studies of the remaining PHTs, meanwhile, were poised to reliably detect only effects of efficacy or harm that were substantial and potentially beyond what might be considered reasonable (e.g., Hedges' $g = 0.25$ in pill placebo-controlled studies of psychological treatments for major depressive disorder; Cuijpers et al., 2014) for psychological interventions.

Individual Bayes factors for harm of PHTs ranged from providing very strong evidence against the possibility of harm to approaching strong evidence of harm. Meta-analytically synthesizing across Bayes factors for grief counseling, we found consistent strong evidence against harm for both pessimistic ($BF_{01} = 23.80$) and optimistic ($BF_{01} = 62.50$) effect selections. Our results were more ambiguous for DARE, as we found that there was either relatively weak ($BF_{01} = 2.14$) or very strong ($BF_{01} = 76.92$) evidence against harm, using pessimistic and optimistic effect selections, respectively.

In contrast, meta-analytic findings for Scared Straight interventions and CISD supported the possibility of harm. For Scared Straight interventions, we found moderate evidence in favor of harm under pessimistic effect selection ($BF_{10} = 4.50$) and moderate evidence against harm under optimistic effect selection ($BF_{01} = 4.39$). Evidence against harm for CISD, meanwhile, was strong under optimistic effect selection ($BF_{01} = 17.24$), but in favor of harm under pessimistic effect selection ($BF_{10} = 8.56$).

3.2 | Magnitude of potential harm/efficacy effects

Using our Bayesian meta-analytic models, we sampled from posterior distributions of effect sizes for each PHT ($n = 10{,}000$) in order to estimate the plausible extent of harm a given PHT might entail (see Figure 2). Median posterior effect sizes and 95% credibility intervals are reported in Table 3. Harm for both DARE and grief counseling appeared unlikely. The plausible extent of harm for CISD and Scared Straight interventions, meanwhile, appeared more consistent, and under pessimistic effect selection specifically, could be substantial. We were unable to compute Bayes factors for boot camp interventions due to insufficiently reported statistical detail.

As an accompanying exploratory analysis, we re-specified these meta-analytic models to evaluate the possible extent of harm (the concern espoused in Lilienfeld (2007)) *or* efficacy (the presumed original intent of the intervention; see Figure 3). Median posterior effect sizes and 95% credibility intervals are reported in Table 4. Out of the four examinable PHTs, only grief counseling appeared to have any substantive

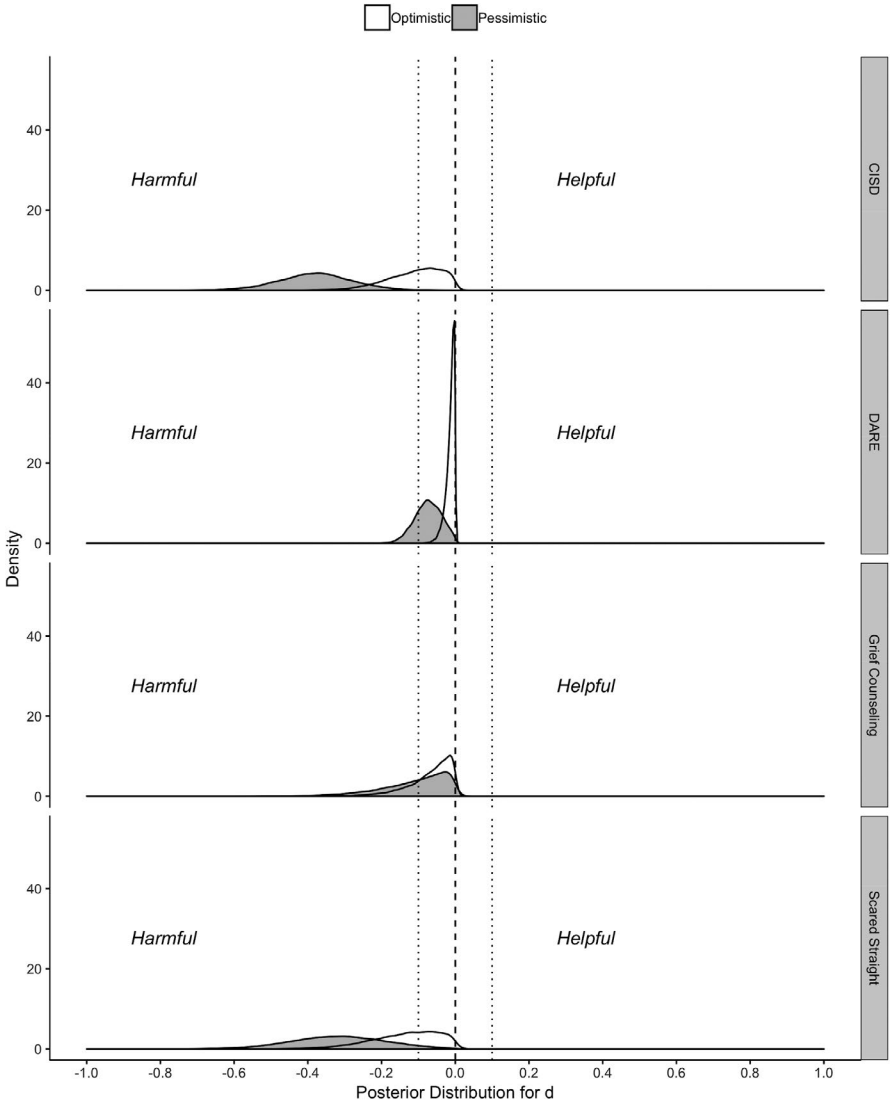

FIGURE 2 Meta-analyzed posterior distributions for four potentially harmful therapies based on optimistic and pessimistic effect selections. Dashed line represents $d = 0$, and dotted line represents interval for a potentially trivial effect. Hypothesis specified a directional test of harm

efficacious potential. DARE, alternatively, demonstrated minimal efficacious and iatrogenic potential—its posteriors largely suggest it is without effect for better or worse. CISD and Scared Straight interventions, finally, appear to be ineffective at best, and appreciably harmful at worst, depending on optimistic or pessimistic effect selection.

4 | GENERAL DISCUSSION

As the psychologist Richard Gist noted, "[W]e must especially work to discourage yielding to the desperate need to do *something* by acquiescing to the compulsion to use *anything*" (Gist, 2001, p. 16). Lilienfeld's (2007) list of PHTs was a seminal contribution that drew attention to maleficence as an unintended consequence of the pursuit of beneficence. However, concerns about the replicability of clinical science (e.g., Tackett et al., 2017, 2018) demand comprehensive inquiry into the evidential value of both claims of therapeutic efficacy (e.g., Sakaluk et al., 2019) and potential harms from interventions. To this end, we performed a meta-scientific review of six PHTs described by Lilienfeld.

More consistently than strong evidence of harm (or benefit), our analyses suggest the evidence underlying many PHTs is weak or ambiguous across a variety of metrics. Findings from PHT trials are limited partially due to insufficient controls (e.g., over-reliance on no-treatment control groups and randomization inconsistencies), as well as spartan statistical reporting prohibiting verification of results. The literatures for most PHTs were also both underpowered to detect plausibly sized effects of psychological interventions and featured inflated rates of statistically significant results. Finally, Bayes factors for many PHTs were either impossible to calculate due to insufficient information or indicated neither strong evidence for harm nor benefit (e.g., DARE). There were some notable, albeit isolated, exceptions to these overall ambiguous findings. Evidence for potential harm most clearly emerged for Scared Straight interventions and CISD. Namely, the pessimistic Bayes factors suggested moderate

TABLE 3 Meta-analytic posterior median and 95% credibility interval for harm

PHT	Effect selection	*d*	95% CR LL	95% CR UL
CISD	Optimistic	−0.038	−0.133	−0.002
CISD	Pessimistic	−0.215	−0.359	−0.073
DARE	Optimistic	−0.01	−0.044	0
DARE	Pessimistic	−0.073	−0.146	−0.01
Grief	Optimistic	−0.012	−0.057	0
Grief	Pessimistic	−0.027	−0.114	−0.001
Scared Straight	Optimistic	−0.12	−0.34	−0.007
Scared Straight	Pessimistic	−0.316	−0.565	−0.075

Note: Negative sign indicates the direction of harm

to strong evidence of harm, with the posterior distributions suggesting that, at best, these interventions are potentially ineffective.

4.1 | Implications for science and practice

We strongly advise that patterns of ambiguous findings for a given PHT should not be taken as evidence that the intervention is not harmful. Ambiguous findings indicate just that ambiguity as to whether the therapy will harm, or potentially help. And even if harm is not likely, the absence of harm is not compelling evidence of benefit. A goal of this review is to highlight that ethical research and practice dictates comprehensive examination of not only effective treatments but also potentially iatrogenic interventions as well. Given the probability for harm detected in the present review, Scared Straight and CISD deserve the most attention and are the most deserving candidates for *psychological reversal* (see Sakaluk et al., 2019). The probability of harm suggested by our Bayesian meta-analysis distinguishes CISD and Scared Straight from the other PHTs analyzed. We remain open to the possibility that harm from Scared Straight and CISD may be mitigated in very specific samples but underscore the need for a high evidential bar to justify future trials or use in clinical practice.

A crucial takeaway from our review is that the traditional model for psychological outcome research is limited. Conducting trials with small teams, in siloed laboratories, and with overly flexible research methods is insufficient to address contemporary inquiries of clinical practice. Registered (Nosek et al., 2018), large-scale collaborative trials and crowdsourced scientific initiatives (Uhlmann et al., 2018) verified for accuracy and reproducibility (Sakaluk, Williams, & Biernat, 2014) are needed to accelerate our scientific understanding of health treatments. The field must also be accepting of informative null findings (Chambers, 2013). Additionally, we recommend that, prior to conducting a treatment trial or initiating treatment, researchers and clinicians investigate the scientific standing of the underlying putative therapeutic mechanisms. When putative mechanisms do not converge with basic science findings, trials and treatments should be discouraged.

4.2 | Limitations

This review is not designed to be an exhaustive analysis of the evidential value of PHTs. Whereas we strove to include as many RCTs as possible, conducting six systematic reviews is beyond the scope of this article. We relied on currently available reviews, and 19% of the reports we extracted from them were unpublished theses and other forms of non–peer-reviewed reports. However, we cannot comment on the state of evidential support for PHTs that were not derived from RCTs, cannot be found in the extant review literature due to publication bias, or were not included in Lilienfeld (2007). Additionally, we have not provided an evaluation of the seven other PHTs identified by Lilienfeld due to a dearth of RCTs. This should not be taken as tacit evidence for lack of harm (or evidence of efficacy). We chose to leverage the internal validity that accompanies RCTs, but other compelling sources of harm should be carefully considered as well, especially when an RCT would not be ethical or pragmatic (e.g., an RCT of conversion/reparative therapy).

Another potential limitation is that our results are relegated to a particular quantitative operationalization of harm (average iatrogenic effects, based on Ttreatment vs. control comparisons, in the full sample of participants for a given RCT), when alternative methods to assess harm are available (e.g., American Geriatrics Society, 2015). From our perspective, our use of a consistent operationalization of harm was desirable in order to render comparable the evaluations of the PHTs that we reviewed, as well as map onto ways in which efficacy has alternatively been considered and evaluated (Chambless & Hollon, 1998; Sakaluk et al., 2019). Further, the operationalization that we used (and relatedly, our selection of metrics of evidential value) *ought* to have been the most feasibly implementable. The computation of our metrics required fairly low-level statistical elements (*M*s,

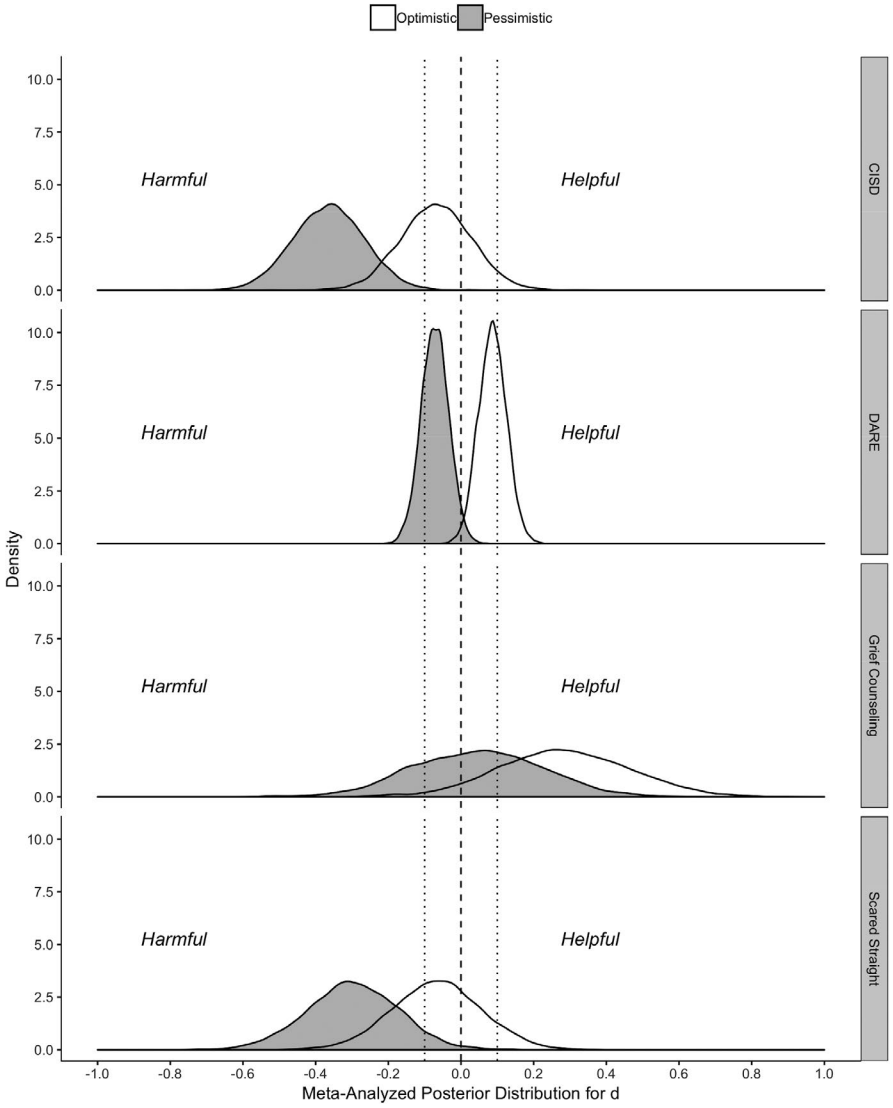

FIGURE 3 Meta-analyzed posterior distributions for four potentially harmful therapies based on optimistic and pessimistic effect selections. Dashed line represents $d = 0$, and dotted line represents interval for a potentially trivial effect. Hypothesis specified a nondirectional test of harm or efficacy

*SD*s, *n*s, test statistic values, and degrees of freedom) to be consistently reported, and reporting standards in many journals (e.g., APA, 2020) ostensibly require these elements to be included in published articles. Thus, while we encourage future efforts to consider alternative operationalizations of harm, we would simultaneously caution those embarking on such an endeavor to be aware that the state of statistical reporting might be even more spartan and inconsistent than in our review, thereby rendering such an investigation even more logistically challenging.

Finally, there is an inherent limitation in conducting a quantitative review of treatments that were at one time deemed potentially harmful. Treatments are often altered when early trials indicate a lack of efficacy or a potential for

TABLE 4 Meta-analytic posterior median and 95% credibility interval for harm or efficacy

PHT	Effect selection	*d*	95% CR LL	95% CR UL
CISD	Optimistic	0.02	−0.108	0.148
CISD	Pessimistic	−0.207	−0.35	−0.065
DARE	Optimistic	0.087	0.013	0.162
DARE	Pessimistic	−0.072	−0.145	0.003
Grief	Optimistic	0.29	0.147	0.434
Grief	Pessimistic	0.083	−0.067	0.23
Scared Straight	Optimistic	−0.064	−0.308	0.171
Scared Straight	Pessimistic	−0.294	−0.543	−0.048

Note: Negative sign indicates the direction of harm.

harm. Furthermore, significant drops in research trials can follow early failures. For example, our review includes nine new CISD RCTs that were not included in Lilienfeld (2007). However, the most recent was published in 2008. We also could not identify any new RCTs of Scared Straight since Lilienfeld's review. Therefore, if CISD, Scared Straight, or any other PHTs have significantly modified their treatment protocols, and the modified protocols have not been examined in RCTs, then we must limit our conclusions: We can only speak to the evidential value of harm based on the protocols followed in available RCTs. It is unclear how PHTs substantially altered in contemporary practice relate to our findings. But without clinical trials (ideally RCTs) supporting the amended protocols, their efficacy and lack of harm are still open questions that should be eyed skeptically by clinicians and researchers.

5 | CONCLUSION

Clinicians are ethically obligated to provide safe treatments. Identifying potentially iatrogenic treatments, as Lilienfeld (2007) notably did, supports clinicians in fulfilling this obligation. Concerns about the replicability of clinical psychological science, including by Lilienfeld himself (Tackett et al., 2017), motivated the present metascientific review of the clinical trial literature underlying PHTs. Across a variety of metrics, the evidential value for studies of PHTs was low; the most common pattern we observed was that effects in the PHT literature were not reported with sufficient detail to render them usable for evaluating their evidential value. However, amidst the ambiguity, there were a few indicators of potential for harm, most notably for CISD and Scared Straight interventions. Considering the potential for harm and the lack of strong evidence for benefit from most PHTs in this study—an exception being grief therapy—we recommend practitioners refrain from using them without better evidence. Moreover, researchers should carefully consider the evidence for benefit/harm before engaging in future trials of PHTs, and they should cease trials of CISD and Scared Straight unless especially impressive evidence indicates a lack of harm. Finally, this review does not serve as an exhaustive evaluation of all PHTs or as the final arbiter of potential efficacy or harm. We encourage clinicians to examine varied forms of evidence (e.g., case reports, uncontrolled studies) for the potential of harm, and to critically evaluate it (e.g., question whether a therapeutic mechanism is scientifically plausible or contorts what we already know about human behavior). Similarly, because our methods are not completely bias-free, researchers should continue evaluating treatment outcome research while developing and applying novel methods to assess evidential value, potential biases, and validity of findings. Increasing the reproducibility and evidential value of future research on potentially iatrogenic treatments will serve clinicians and researchers well in their efforts to avoid unintentionally harming those they are ethically responsible to help.

ACKNOWLEDGMENTS

This research was supported by a SSHRC Insight Development Grant awarded to Dr. Sakaluk, and a SSHRC Doctoral Fellowship awarded to Robyn Kilshaw.

ORCID

Alexander J. Williams https://orcid.org/0000-0001-9541-7981
Yevgeny Botanov https://orcid.org/0000-0001-8005-7346
John Kitchener Sakaluk https://orcid.org/0000-0002-2515-9822

REFERENCES

About D. A. R. E. (2019). Retrieved from https://dare.org/about/

American Geriatrics Society (2015). American Geriatrics Society 2015 updated Beers Criteria for potentially inappropriate medication use in older adults. *Journal of the American Geriatrics Society*, *63*(11), 2227–2246. https://doi.org/10.1111/jgs.13702

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073. https://doi.org/10.1037/0003-066X.57.12.1060

American Psychological Association (2009). *Report of the American Psychological Association Task Force on Appropriate Therapeutic Responses to Sexual Orientation*. Retrieved from http://www.apa.org/pi/lgbc/publications/therapeutic-resp.html

American Psychological Association (2020). *Publication manual of the American Psychological Association*, 7th ed. Washington, DC: American Psychological Association.

Barnett, G. D., & Howard, F. F. (2018). What doesn't work to reduce reoffending? A review of reviews of ineffective interventions for adults convicted of crimes. *European Psychologist*, *23*(2), 111–129. https://doi.org/10.1027/1016-9040/a000323

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Cesarini, D., Chamgers, C. D., Clyde, M., Cook, T. D., Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Birur, B., Moore, N. C., & Davis, L. L. (2017). An evidence-based review of early intervention and prevention of posttraumatic stress disorder. *Community Mental Health Journal*, *53*(2), 183–201. https://doi.org/10.1007/s10597-016-0047-x

Bonanno, G. A., & Lilienfeld, S. O. (2008). Let's be realistic: When grief counseling is effective and when it's not. *Professional Psychology: Research and Practice*, *39*, 377–378. https://doi.org/10.1037/0735-7028.39.3.377

Brunson, J. C. (2018). *ggalluvial: Alluvial Diagrams in 'ggplot2'. R package version, (9), 1*. Retrieved from https://CRAN.R-project.org/package=ggalluvial

Caputi, T. L., & McLellan, A. T. (2017). Truth and DARE: Is DARE's new Keepin' it REAL curriculum suitable for American nationwide implementation? *Drugs: Education, Prevention and Policy*, *24*, 49–57. https://doi.org/10.1080/09687637.2016.1208731

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2017). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices. Psychological Science*, *2*(2), 115–144. https://doi.org/10.31234/osf.io/9h3nu

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., … Woody, S. R. (1998). Update on empirically validated therapies. II. *The Clinical Psychologist*, *51*, 3–16. https://doi.org/10.1037/e619622010-001

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18. https://doi.org/10.1037/0022-006X.66.1.7

Champely, S. (2018). *pwr: Basic functions for power analysis. R package version, 1(2-2)*. Retrieved from https://CRAN.R-project.org/package=pwr

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*, 750–764. https://doi.org/10.1177/1745691616664694

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

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *The Canadian Journal of Psychiatry*, *58*, 376–385. https://doi.org/10.1177/070674371305800702

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: A meta-analysis. *Psychological Medicine*, *44*, 685–695. https://doi.org/10.1017/S0033291713000457

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966 https://doi.org/10.1177/0956797613504966

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York, NY: Free Press.

de Graaff, L. F., Honig, A., van Pampus, M. G., & Stramrood, C. A. (2018). Preventing post-traumatic stress disorder following childbirth and traumatic birth experiences: A systematic review. *Acta Obstetricia Et Gynecologica Scandinavica*, *97*(6), 648–656. https://doi.org/10.1111/aogs.13291

Everly, G. S., & Mitchell, J. T. (2000). The debriefing "controversy" and crisis intervention: A review of lexical and substantive issues. *International Journal of Emergency Mental Health*, *2*, 211–226.

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*, 319–324. https://doi.org/10.1037/h0063633

Faggiano, F., Minozzi, S., Versino, E., & Buscemi, D. (2014). Universal school-based prevention for illicit drug use. *Cochrane Database of Systematic Reviews*, https://doi.org/10.1002/14651858.CD003020.pub3

Faggiano, F., Vigna-Taglianti, F., Versino, E., Zambon, A., Borraccino, A., & Lemma, P. (2005). School-based prevention for illicit drugs' use. *Cochrane Database of Systematic Reviews*, https://doi.org/10.1002/14651858.CD003020.pub2

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them, PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Flynn, A. B., Falco, M., & Hocini, S. (2015). Independent evaluation of middle school–based drug prevention curricula: A systematic review. *JAMA Pediatrics*, *169*(11), 1046–1052. https://doi.org/10.1001/jamapediatrics.2015.1736

Forneris, C. A., Gartlehner, G., Brownley, K. A., Gaynes, B. N., Sonis, J., Coker-Schwimmer, E., … Lohr, K. N. (2013). Interventions to prevent post-traumatic stress disorder: A systematic review. *American Journal of Preventive Medicine*, *44*(6), 635–650. https://doi.org/10.1016/j.amepre.2013.02.013

Gernsbacher, M. A. (2018). Writing empirical articles: Transparency, reproducibility, clarity, and memorability. *Advances in Methods and Practices in Psychological Science*, *1*, 403–414. https://doi.org/10.1177/2515245918754485

Gist, R. (2001). A message of caution. *American Psychological Society Observer*, *14*(8), 16–17.

Gist, R. (2015). Psychological debriefing. In R. L. Cautin, & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (Vol. 4, pp. 2303–2308). Hoboken, NJ: Wiley.

Guyatt, G. H., Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., … Schunemann, H. (2011). GRADE guidelines 6. Rating the quality of evidence—imprecision. *Journal of Clinical Epidemiology*, *64*(12), 1283–1293. https://doi.org/10.1016/j.jclinepi.2011.01.012

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Horn, S. R., & Feder, A. (2018). Understanding resilience and preventing and treating PTSD. *Harvard Review of Psychiatry*, *26*(3), 158–174. https://doi.org/10.1097/HRP.0000000000000194

Howlett, J. R., & Stein, M. B. (2016). Prevention of trauma and stressor-related disorders: A review. *Neuropsychopharmacology*, *41*(1), 357. https://doi.org/10.1038/npp.2015.261

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Hussey, I., & Hughes, S. (2018). Hidden invalidity among fifteen commonly used measures in social and personality psychology, PsyArXiv. 10.31234/osf.io/7rbfp

Jarrett, C. (2008). When therapy causes harm. *Psychologist*, *21*, 10–12.

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. https://doi.org/10.1177/0956797611430953

Josefson, D. (2001). Rebirthing therapy banned after girl died in 70 minute struggle. *BMJ*, *322*, 1014. https://doi.org/10.1136/bmj.322.7293.1014/e

Joyce, S., Modini, M., Christensen, H., Mykletun, A., Bryant, R., Mitchell, P. B., & Harvey, S. B. (2016). Workplace interventions for common mental disorders: A systematic meta-review. *Psychological Medicine*, *46*(4), 683–697. https://doi.org/10.1017/S0033291715002408

Kearns, M. C., Ressler, K. J., Zatzick, D., & Rothbaum, B. O. (2012). Early interventions for PTSD: A review. *Depression and Anxiety*, *29*(10), 833–842. https://doi.org/10.1002/da.21997

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*(5), 578–589. https://doi.org/10.1037/met0000209

King, K. M., Pullmann, M. D., Lyon, A. R., Dorsey, S., & Lewis, C. C. (2018). Using implementation science to close the gap between the optimal and typical practice of quantitative methods in clinical science, PsyArXiv. https://doi.org/10.31234/osf.io/n2v68

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr, Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cernalcilar, Z., Chandler, J., Cheong, W., Davis, W., Devos, T., Elsner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Kramer, D. N., & Landolt, M. A. (2011). Characteristics and efficacy of early psychological interventions in children and adolescents after single trauma: A meta-analysis. *European Journal of Psychotraumatology*, *2*(1), 7858. https://doi.org/10.3402/ejpt.v2i0.7858

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lapp, L. K., Agbokou, C., Peretti, C. S., & Ferreri, F. (2010). Management of post traumatic stress disorder after childbirth: A review. *Journal of Psychosomatic Obstetrics & Gynecology*, *31*(3), 113–122. https://doi.org/10.3109/0167482X.2010.503330

Larson, D. G., & Hoyt, W. T. (2007). What has become of grief counseling? An evaluation of the empirical foundations of the new pessimism. *Professional Psychology: Research and Practice*, *38*, 347. https://doi.org/10.1037/0735-7028.38.4.347

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53–70. https://doi.org/10.1111/j.1745-6916.2007.00029.x

Linden, M. (2013). How to define, find and classify side effects in psychotherapy: From unwanted events to adverse treatment reactions. *Clinical Psychology and Psychotherapy*, *20*, 286–296. https://doi.org/10.1002/cpp.1765

Lohr, J. M., Gist, R., Deacon, B., Devilly, G. J., & Varker, T. (2015). Science and non-science based treatments for trauma related stress disorders. In S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Eds.), *Science and pseudoscience in clinical psychology* (2nd ed., pp. 277–321). New York, NY: The Guilford Press.

Malchiodi, C. A. (Ed.) (2013). *Expressive therapies*. New York, NY: Guilford Publications.

Mallory, C., Brown, T. N., & Conron, K. J. (2018). Conversion therapy and LGBT youth. Williams Institute, UCLA School of Law. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Conversion-Therapy-LGBT-Youth-Jan-2018.pdf

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

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Meade, B., & Steiner, B. (2010). The total effects of boot camps that house juveniles: A systematic review of the evidence. *Journal of Criminal Justice*, *38*, 841–853. https://doi.org/10.1016/j.jcrimjus.2010.06.007

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*, 131–144. https://doi.org/10.1177/2515245917747656

Mitchell, J. T. (n.d.). *Critical incident stress debriefing (CISD)*. Retrieved from http://www.info-trauma.org/flash/media-f/mitchellCriticalIncidentStressDebriefing.pdf

Mitchell, J. T. (1983). When disaster strikes: The critical incident stress debriefing process. *Journal of Emergency Medical Services*, *13*(11), 49–52.

Moos, R. H. (2005). Iatrogenic effects of psychosocial interventions for substance use disorders: Prevalence, predictors, prevention. *Addiction*, *100*, 595–604. https://doi.org/10.1111/j.1360-0443.2005.01073.x

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

Morris (2015). *Trauma post trauma*. Slate. Retrieved from https://slate.com/technology/2015/07/prolonged-exposure-therapy-for-ptsd-the-vas-treatment-has-dangerous-side-effects.html

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

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 2600–2606. https://doi.org/10.1073/pnas.1708274114

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

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

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., … Zrubka, M. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*, 268–294. https://doi.org/10.1177/1745691618755704

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

Pack, M. J. (2013). Critical incident stress management: A review of the literature with implications for social work. *International Social Work*, *56*(5), 608–627. https://doi.org/10.1177/0020872811435371

Pan, W., & Bai, H. (2009). A multivariate approach to a meta-analytic review of the effectiveness of the DARE program. *International*

Journal of Environmental Research and Public Health, *6*(1), 267–277. https://doi.org/10.3390/ijerph6010267

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013). 'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency. *Cochrane Database of Systematic Reviews*. https://doi.org/10.4073/csr.2013.5

Qi, W., Gevonden, M., & Shalev, A. (2016). Prevention of post-traumatic stress disorder after trauma: Current evidence and future directions. *Current Psychiatry Reports*, *18*(2), 20. https://doi.org/10.1007/s11920-015-0655-0

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rose, S. C., Bisson, J., Churchill, R., & Wessely, S. (2002). Psychological debriefing for preventing post traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD000560

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638. https://doi.org/10.1037/0033-2909.86.3.638

Rosner, R., Kruse, J., & Hagl, M. (2010). A meta-analysis of interventions for bereaved children and adolescents. *Death Studies*, *34*(2), 99–136. https://doi.org/10.1080/07481180903492422

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin and Review*, *18*, 682–689. https://doi.org/10.3758/s13423-011-0088-7

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Ruscio, J., Haslam, N., & Ruscio, A. M. (2006). *Introduction to the taxometric method: A practical guide*. New York, NY: Routledge.

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, *66*, 47–54. https://doi.org/10.1016/j.jesp.2015.09.013

Sakaluk, J. K. (2019). Expanding statistical frontiers in sexual science: Taxometric, invariance, and equivalence testing. *The Journal of Sex Research*, *56*(4–5), 1–36. https://doi.org/10.1080/00224499.2019.1568377

Sakaluk, J., Williams, A., & 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

Sakaluk, J. K., Williams, A. J., Kilshaw, R., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology*, *128*(6), 500–509. https://doi.org/10.1037/abn0000421

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

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

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Shapiro, A. (Director). (1978). *Scared Straight! [Motion Picture]*. USA: Golden West Television.

Shapiro, A. (Producer), & Coyne, P. (Producer) (2011). *Beyond Scared Straight [Television Series]*. USA: A&E.

Siemieniuk, R., & Guyatt, G. (2019). *What is GRADE?* Retrieved from https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/what-is-grade

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

Skeffington, P. M., Rees, C. S., & Kane, R. (2013). The primary prevention of PTSD: A systematic review. *Journal of Trauma & Dissociation*, *14*(4), 404–422. https://doi.org/10.1080/15299732.2012.753653

Sloboda, Z., Stephens, R. C., Stephens, P. C., Grey, S. F., Teasdale, B., Hawthorne, R. D., … Marquette, J. F. (2009). The adolescent substance abuse prevention study: A randomized field trial of a universal substance abuse prevention program. *Drug and Alcohol Dependence*, *102*, 1–10. https://doi.org/10.1016/j.drugalcdep.2009.01.015

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752. https://doi.org/10.1037/0003-066X.32.9.752

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science*, *1*(1), 115–120. https://doi.org/10.1177/2515245918757689

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886–899. https://doi.org/10.1177/1745691615609918

Stewart, R. E., & Chambless, D. L. (2009). Cognitive–behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology*, *77*, 595–606. https://doi.org/10.1037/a0016032

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2018). Psychology's replication crisis and clinical psychological science, PsyArXiv. https://doi.org/10.31234/osf.io/kc8xt

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

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*(4), 317–338. https://doi.org/10.1111/cpsp.12122

Uhlmann, E. L., Chartier, C. R., Ebersole, C. R., Errington, T., Kidwell, M. C., Lai, C., … Nosek, B. A. (2018). Scientific utopia: III. *Crowdsourcing Science*. https://psyarxiv.com/vg649/

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. Jr, Allard, E. S., Benning, S. D., Blouin-Hudon, L., Bulnes, T., Cladwell, T., Calin-Jageman, R., Capaldi, C., Carfagno, N., Chasten, K., Cleeremans, A., Connell, T., DeCicco, J., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Walsh, C. G., Xia, W., Li, M., Denny, J. C., Harris, P. A., & Malin, B. A. (2018). Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: Current practices and future challenges. *Advances in Methods and*

Practices in Psychological Science, *1*(1), 104–114. https://doi.org/10.1177/2515245917749652

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. New York, NY: Routledge.

Welsh, B. C., & Rocque, M. (2014). When crime prevention harms: A review of systematic reviews. *Journal of Experimental Criminology*, *10*(3), 245–266. https://doi.org/10.1007/s11292-014-9199-2

West, S. L., & O'Neal, K. K. (2004). Project DARE outcome effectiveness revisited. *American Journal of Public Health*, *94*(6), 1027–1029. https://doi.org/10.2105/AJPH.94.6.1027

Wethington, H. R., Hahn, R. A., Fuqua-Whitley, D. S., Sipe, T. A., Crosby, A. E., Johnson, R. L., … Chattopadhyay, S. (2008). The effectiveness of interventions to reduce psychological harm from traumatic events among children and adolescents: A systematic review. *American Journal of Preventive Medicine*, *35*(3), 287–313. https://doi.org/10.1016/j.amepre.2008.06.024

Wittouck, C., Van Autreve, S., De Jaegere, E., Portzky, G., & van Heeringen, K. (2011). The prevention and treatment of complicated grief: A meta-analysis. *Clinical Psychology Review*, *31*(1), 69–78. https://doi.org/10.1016/j.cpr.2010.09.005

How to cite this article: Williams AJ, Botanov Y, Kilshaw RE, Wong RE, Sakaluk JK. Potentially harmful therapies: A meta-scientific review of evidential value. *Clin Psychol Sci Pract*. 2020;00:e12331. https://doi.org/10.1111/cpsp.12331