American Psychological Association

Journal of Abnormal Psychology

© 2019 American Psychological Association
0021-843X/19/$12.00

2019, Vol. 128, No. 6, 547–562
http://dx.doi.org/10.1037/abn0000417

Using Implementation Science to Close the Gap Between the Optimal and Typical Practice of Quantitative Methods in Clinical Science

Kevin M. King, Michael D. Pullmann,
Aaron R. Lyon, and Shannon Dorsey
University of Washington

Cara C. Lewis
Kaiser Permanente Washington Health Research Institute,
Seattle, Washington

Quantitative methods remain the fundamental approach for hypothesis testing, but in approaches to data analysis there is substantial evidence of a gap between what is optimal and what is typical. It is clear that diffusion and dissemination alone are not maximally effective at improving data analytic practices in clinical psychological science. Amid declines in quantitative psychology training, and growing demand for advanced quantitative methods, applied researchers are increasingly called upon to conduct and evaluate research using methods in which they lack expertise. This "research-to-practice" gap in which rigorously developed and empirically supported quantitative methods are not applied in practice has received little attention. In this article, we describe how implementation science, which aims to reduce the research-to-practice gap in health care, offers a promising set of methods for closing the gap for quantitative methods. By identifying determinants of practice (i.e., barriers and facilitators of change), implementation strategies can be selected to increase adoption and high-fidelity application of new quantitative methods to improve scientific inferences and policy and practice decisions in clinical psychological science.

General Scientific Summary
Making studies more replicable will require more effective use of statistics in research, but there is a large gap between how statistics are applied in psychological research and how they should be applied. The current article describes how the lessons of clinical implementation science, which has focused on getting evidence based treatments into community practice settings, may be applied to improve research in clinical psychology.

Keywords: quantitative methods, open science, implementation science, quantitative implementation

There is a gap between how psychological science might be optimally conducted and how it is typically conducted, which undermines the credibility of research findings. For example, although 97% of effects in a large scale replication effort were statistically significant in their original studies, only 36% of those effects were statistically significant in new, larger samples, with a median effect size half that of the original studies (Aarts et al., 2015), echoing earlier concerns that many published research findings may be false (Ioannidis, 2005). There are many factors that likely undermine the credibility of current research, but many of them reflect a gap between how research should be conducted and how it is conducted. Researchers engage in questionable research practices (Fiedler & Schwarz, 2016; John, Loewenstein, & Prelec, 2012) to achieve statistically significant results through selective testing, reporting, or utilizing other "researcher degrees of freedom" which can dramatically increase the number of false positive findings (Simmons, Nelson, & Simonsohn, 2011). This leads to higher than expected rates of false positives, especially when research practices are not robust (Ioannidis, 2005; Szucs, 2016). Compounding the problems, researchers then engage in hypothesizing after the results are known (i.e., HARKing) implying that post hoc findings are actually a priori (Kerr, 1998).

Clinical psychological science has long faced a similar gap between what researchers operationalize as optimal treatment and what occurs in community practice settings (Kazdin, 2008). Researchers have long known that psychotherapy has large effect sizes when conducted in controlled research settings (Weisz, Donenberg, Han, & Weiss, 1995), but fails to have much if any effectiveness in typical clinical samples and settings (Weisz & Jensen, 2001). Indeed, many fields have struggled to translate research identifying best practices into meaningful change in real-world outcomes (Damschroder et al., 2009). For example, studies have reported that it can take 10 to 25 years for biomedical

Kevin M. King, Michael D. Pullmann, Aaron R. Lyon, and Shannon Dorsey, Department of Psychology, University of Washington; Cara C. Lewis, Kaiser Permanente Washington Health Research Institute, Seattle, Washington.

We appreciate the feedback from Teague Henry and Dale Kim who provided comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Kevin M. King, Department of Psychology, University of Washington, Box 351525, Seattle, WA 98195-1525. E-mail: kingkm@uw.edu

interventions to appear in practice (Contopoulos-Ioannidis, Alexiou, Gouvias, & Ioannidis, 2008; Morris, Wooding, & Grant, 2011). In mental health services research, there is evidence for low availability of both delivery of and training in evidence-based practices (EBPs). Fewer than 10% of therapists use EBP manuals (Becker, Smith, & Jensen-Doss, 2013). One study indicated that most children with disruptive behavior problems in a community sample were exposed to a broad array of evidence based interventions (such as positive reinforcement or modeling), but most exposure was at too low of an intensity to have a therapeutic effect (Garland et al., 2010). Only a third of patients with mental health disorders who received treatment in the past year reported receiving minimally adequate (i.e., ≥ 2 visits) treatment in the National Comorbidity Survey Replication study (Wang et al., 2005). Only 17% of university based mental health training programs required both didactic training and clinical supervision in EBPs (Weissman et al., 2006), while a typical provider's caseload features a wide variety of diagnoses, and therefore might be expected to deploy a large number of EBPs (Park, Tsai, Guan, & Chorpita, 2018; Weisz et al., 1995). While a more recent study of MSW programs found that nearly 90% *reported* teaching EBPs, but the majority of EBP instruction was carried out by adjunct faculty, and program directors reported concern about the competency of core faculty to teach EBPs as well as low availability of practicum settings in which to apply them (Bertram, Charnin, Kerns, & Long, 2015).

Implementation science developed in response to repeated observations of this "research to practice" gap in clinical settings. Implementation science is a relatively new field in health care that offers a perspective and methods that could be applied to enhance the quality of quantitative methods application in psychopathology research. Evidence suggests that diffusion (the passive spread of knowledge) and dissemination (targeted efforts at increasing knowledge) of research evidence for best practices does little to change behaviors or systems (Davis, Thomson, Oxman, & Haynes, 1995). Diffusion and dissemination efforts often fail to produce broad-band improvements in practice partially because they only target knowledge, with the problematic assumption that increasing knowledge alone will produce behavior (Azjen, 1991; Jensen-Doss & Hawley, 2010). Rather, there are multiple determinants of implementation which exist at multiple levels, including characteristics of the intervention itself, intra- and interpersonal factors, and organizational and systems factors. Successful implementation is believed to be more probable when these multilevel factors align. For instance, implementation is more likely to occur when an intervention is designed to be accessible and have low barriers to success, people have favorable attitudes, beliefs, skills, and motivation to engage in a given intervention, are in a peer and organizational context that supports the use of the intervention, and are working in a system that has policies which favor implementation (Damschroder et al., 2009). Identifying and prioritizing determinants is a critical first step to then inform the selection and application of strategies that might be used to promote implementation (Powell et al., 2015). Finally, successful implementation will not just result in effective practice, but it will alter attitudes about what is acceptable practice, and produce widespread adoption and deep penetration in specific disciplines that are marked by sustainable, high fidelity practice with lower barriers to entry (Lewis et al., 2015; Proctor et al., 2011).

This relatively new field of clinical implementation science may be a useful lens to view the challenge of creating sustained, multisystems change to increase the replicability of research. To date, the "credibility revolution" has led to calls for more transparent and robust approaches to conducting science, including preregistration of hypotheses; open sharing of protocols, analytic syntax and data; and collaborative efforts to replicate prior research (Vazire, 2018). The reforms proposed to improve the credibility of psychological research (such as open syntax and data, or preregistration) have assumed that applied researchers can effectively evaluate the statistical models that are proposed or used in data analysis, and that increasing knowledge about improved methods will naturally lead to improved practice. From an implementation science perspective, the true impact of a statistical method on theory, intervention, and policy will depend in large part on the degree to which it is disseminated, used correctly, in situations where it is appropriate, and by researchers who are not experts or did not originate the analytical method. Otherwise, its application will be limited to a relatively small number of research teams who can collaborate with a methodological expert.

Herein, we discuss how the perspective and methods of implementation science may be applied to improve the research-to-practice pipeline for quantitative methods. To our knowledge, an implementation science lens has not been applied specifically to quantitative methods. We discuss how implementation science could guide research examining the determinants of successful quantitative methods implementation, outcomes of implementation effects, and strategies for implementing quantitative reforms. Finally, we offer examples from the implementation of clinical interventions to illustrate ways researchers have approached the integration of evidence-based practices in other domains.

The Connection Between Statistical Analysis and the Credibility of Research Findings

Variation in how data are processed and analyzed may influence the credibility of scientific conclusions and reduce the replicability of specific findings (Schmidt, 2009). Replicability can be operationalized in terms of *reproduction* (the same data with a different analyst), *direct replication* (the same data collection and analytic procedures with a new sample), and *conceptual replication* (repeating a hypothesis test with a different, related method). Interested readers may be directed to an excellent review of issues around replication by Zwaan, Etz, Lucas, and Donnellan (2017), as well as associated commentaries.

Replication of findings from tests of clinical interventions may add new dimensions to the conversation on replicability. Interventions are often changed as they are translated from the research lab to community environments, preventing strict tests of replication while increasing the generalizability of intervention research findings. For this reason, hybrid implementation-effectiveness trials are used to simultaneously test the effectiveness of an intervention while also testing or observing elements of implementation that might influence the installation and/or sustainability of an intervention in applied settings (Curran, Bauer, Mittman, Pyne, & Stetler, 2012). This research is critical because there may be aspects of research-developed interventions that would be difficult if not impossible to implement in community clinic settings. Implementation research also describes an "adaptable periphery" of elements that can be changed about the intervention or its delivery without modifying the core elements responsible for change. For example, effective intervention elements might be modularized to better fit with the schedule of providers (Lyon et al., 2015), or lay providers might be used in contexts where mental health professionals are scarce (Murray et al., 2011).

Decisions about data analysis influence all forms of replication. For example, even the simplest analytic decisions can influence the conclusions drawn from a given study, such the treatment of outliers and non-normal data (Aguinis, Gottfredson, & Joo, 2013; Osborne, 2002), whether to dichotomize or categorize variables (MacCallum, Zhang, Preacher, & Rucker, 2002), which covariates to include (Patel, Burford, & Ioannidis, 2015), and whether or not to adjust *p*-critical due to multiple comparisons (Cramer et al., 2016). Indeed, a recent study demonstrated that when presented with a dataset and a basic research question, 29 different data analytic teams used nearly as many different analytic approaches, with only 69% of groups concluding that the effect was significant, and effect sizes ranging widely (Silberzahn et al., 2018). In short, even decisions about data analysis that are made in good faith can produce variability in research findings.

Some work on the credibility of research findings has focused on analytic issues, such as how a researcher's flexibility in the collection or analysis of data can produce unreliable findings. For example, a seminal article introduced the notion of "researcher degrees of freedom," noting that many decisions that researchers make (such as the choice of dependent variables, sample size, covariates, or selective reporting of results) can dramatically increase the number of false positive findings (Simmons et al., 2011). Similarly, the multiplicity of decisions which occur as a part of data analysis has been described as the "garden of forking paths," such that researchers' decisions about how to analyze data often represent only a single representation of how a given hypothesis could be tested (Gelman & Loken, 2013; Rovine & Molenaar, 2005; Tomarken & Waller, 2003). There have been suggestions and attempts to change quantitative practice in order to improve the credibility of research, such as requiring confidence intervals and parameter estimation (Cumming, 2014), augmenting *p* values with confidence intervals and power estimates (Fritz, Scherndl, & Kühberger, 2013), lowering the value for *p*-critical (Ioannidis, 2018), banning *p*-critical altogether (Loftus, 1996), or moving entirely away from frequentist statistics toward Bayesian statistics (Wagenmakers et al., 2018). It is important to note, however, that nearly all of these proposed methods follow the same "implementation strategy" format: An article or author guideline is published, recommendations for improving practice are provided, and presumably researchers are expected to read, understand, and successfully implement the recommendations.

Moreover, the conversation on replicability has largely focused on fields where experimental analysis dominates, with less attention paid to clinical science (Tackett et al., 2016). Data analyses in clinical psychology can be very complex, because the field is often focused on areas such as rare behaviors, high-risk samples, and mechanisms of change, and extensively uses intervention and observational designs. For example, intervention or experimental psychopathology researchers might mix experimental methods with longitudinal data analytic methods to study how treatment effects or psychopathology unfolds over both short (Tang, DeRubeis, Beberman, & Pham, 2005) and long (Laurenceau, Hayes, & Feldman, 2007) periods of time. Researchers who study psychopathology might use structural equation (King, Luk, et al., 2018), latent class (Witkiewitz et al., 2013), or network analysis (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011) to study the structure of psychopathology, or use longitudinal models (King, Littlefield et al., 2018) to study how psychopathology changes over time. Many clinical researchers must use methods that account for nesting of observations within clusters (such as multiple observations of the same people, or multiple clients nested within providers nested within organizations; McNeish, Stapleton, & Silverman, 2016). Most researchers who collect data over time must account for data missing due to dropout (Schafer & Graham, 2002), and clinical psychologists must contend with heterotypic continuity, where the same disorder might be expressed with different patterns of symptoms over time. Moreover, because clinical psychologists study disorders, symptoms, or behaviors that are infrequent, they often encounter data that are non-normally distributed (such as binary, count, log-normal, or zero-inflated distributions). Finally, clinical psychologists often analyze data from small, hard to reach samples, which can both limit power and the ability to use "optimal" statistical methods which may rely on large-sample assumptions.

The Gap Between Optimal and Typical Applications of Quantitative Methods

We propose that the "intervention" in the implementation science of methods is the specific quantitative method being disseminated and implemented. There is substantial evidence of a "research to practice gap," or a large difference between optimal and typical data analytic interventions, that mirrors the gap in health interventions. New statistical methods take a long time to appear in applied journals, and most statistical methods are rarely used in the applied literature. Methodological articles are both highly cited and broadly ignored: Some influential articles manage to receive a wide audience, but the majority of developments in quantitative methodology go unnoticed by substantive researchers. Sharpe (2013) highlighted this contradiction, noting that 16 articles published in *Psychological Methods* accounted for *half* of all citations for that journal as of 2012; nearly all of these articles were reviews that provided concrete examples and made recommendations about how analyses should be conducted. This is not new. in 1992, seven of the 10 most highly cited articles in psychology were focused on the application of statistics (Sternberg, 1992). An earlier study in the biomedical literature indicated that articles introducing an existing technique to a new audience can reach hundreds of citations in only 4 to 5 years (Altman & Goodman, 1994). On the other hand, these articles represent the minority of quantitative research. Altman and Goodman (1994) showed that new statistical techniques take an average of 4 to 6 years to be cited 25 times in the biomedical literature. A review noted that the median number of quantitative articles cited in substantive (i.e., nonmethodological) articles was one (Mills, Abdulla, & Cribbie, 2010), while 39% of articles did not reference a single quantitative article. This article suggested that across 15 years, a typical article published in a quantitative journal would be expected to be cited only twice by substantive articles. Taken together, this suggests that new developments in statistical methods are slowly adopted by applied researchers, but this process is greatly facilitated by expository articles.

Unfortunately, it is hard to estimate the extent to which even highly cited methodological articles impact practice, as evidence suggests that basic statistical principles are poorly understood by most psychologists. Several examples exist. Researchers commonly misinterpret *p* values to represent the strength of an observed effect, or that a nonsignificant *p* value means that there is no effect (Nickerson, 2000). They use Cronbach's alpha to provide information on the unidimensionality of a scale, yet a scale could have a high alpha while being comprised of two perfectly uncorrelated subscales that, within subscale, have high alpha (Schmitt, 1996; Sijtsma, 2009). Researchers misinterpret confidence intervals as meaning that 95% of replications of an original study will contain the estimate, when in fact on average 83% will because both original and replication means vary around the true effect size (Belia, Fidler, Williams, & Cumming, 2005; Cumming, Williams, & Fidler, 2004). Researchers often believe that linear models (such as multiple regression) require normally distributed data, but the only assumption of normality is for residuals (Williams, Grajales, & Kurkiewicz, 2013).

Problems in the interpretation of statistical principles are likely driven in part by researchers' reliance on statistical rules of thumb and archetypes rather than statistical thinking, which reflects an active process where a theoretical problem is connected with real data though a statistical model (see Wild & Phannkuch, 1999 for an in-depth discussion). Instead, researchers rely heavily on "mindless rituals" (Gigerenzer, 2004) and "magic numbers" (such as for sample size, Kar & Ramalingam, 2013; or for model fit, Marsh, Hau, & Wen, 2004) to perform and interpret statistical models rather than applying statistical thinking. Null-hypothesis significance testing (NHST) is one clear example of a mindless ritual, where researchers inspect *p* values to make a yes/no decision about their hypothesis (Gigerenzer, 2004). Another long-held rule of thumb is that a sample size of 30 per cell was sufficient to detect differences across groups (Wilson Von Voorhis & Morgan, 2007). Rituals, rules, and magic numbers can be helpful checks against common errors in human judgment (Tversky & Kahneman, 1974), but when they are used inflexibly or in a context where developments in the quantitative literature are largely ignored, they can become quickly outdated. For example, researchers tend to rely on strict cutoffs to adjudicate model fit, such as CFI $>$ .95 or RMSEA $<$ .05 (D. L. Jackson, Gillaspy, & Purc-Stephenson, 2009). The highly cited article on model fit by Hu and Bentler (1999) is most often cited as providing strict cutoffs for determining model fit. Yet that article only simulated model fit for a specific factor model (where loadings = .70), and subsequent research has suggested that what comprises "good" model fit can very much depend on the characteristics of the true model (Chen, Curran, Bollen, Kirby, & Paxton, 2008; McNeish, An, & Hancock, 2018), meaning that the results of Hu and Bentler's (1999) original work may not generalize to many cases of SEM where it is cited. In other words, "golden rules" for model fit should be taken to be neither golden nor rules (Marsh et al., 2004). Indeed, statistical "myths" are so common that entire books have been dedicated to falsifying them (Lance & Vandenberg, 2008).

There are frequent errors in reporting descriptive and inferential results in published literature. For example, a review of social psychology journals found that half of a sample of articles contained at least one sample mean or standard deviation inconsistent with the reported sample size, and 20% contained multiple inconsistencies (Brown & Heathers, 2017). For example, 14% of *p* values were inconsistent with the reported statistics (such as *t* or *z* test values) in a review of two psychiatry journals (Berle & Starcevic, 2007). Similar studies have reported error rates of 11% in high impact journals (*Nature* and *British Medical Journal*; (García-Berthou & Alcaraz, 2004) and 18% in a broad survey of psychology (Bakker & Wicherts, 2011). A recent survey of all articles published in APA flagship journals from 1985 to 2013 indicated that half of articles that used NHST had at least one *p* value that was inconsistent with the test statistics reported, and inconsistencies that would have altered statistical conclusions of the article ranged from 6–16% across journals (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). In short, even simple statistics such as *p* values are frequently misreported, often in directions that alter the study's conclusions in favor of the researcher's hypotheses (Bakker & Wicherts, 2011). This suggests that the current approach to peer review is insufficient to detect and correct errors, even in articles using basic statistics.

There are no parallel studies on the frequency of analytic or reporting errors of more complex statistical models that often characterize clinical psychological science. However, there are many examples of methodological reviews that note common and longstanding errors in the application of methods. For example, Miller and Chapman (2001) described decades of methodological commentaries on the difficulties of adjusting predictor effects for covariates (that is, "Lord's paradox"; Lord, 1967), and showed how misunderstandings persisted in the applied literature. Indeed, researchers still struggle to interpret covariate effects, confounding and suppressor variables (MacKinnon, Krull, & Lockwood, 2000; Westfall & Yarkoni, 2016). Cohen (1962) famously described how small sample sizes in many psychological studies left them underpowered to detect small or medium effects, and 27 years later, Sedlmeier and Gigerenzer (1989) showed that power was almost ubiquitously absent of mention in psychological studies. Fiske and Campbell (1992) noted that in spite of over 2,000 citations of their influential article on construct validity (Campbell & Fiske, 1959), little progress had been made in improving the link between psychological methods and psychological constructs. Indeed, the debate over the use of NHST (Nickerson, 2000) highlights the challenges of changing common statistical practices of applied researchers.

There are many examples of the limited effects of diffusion and dissemination in the methodological literature. Often the introduction or reintroduction of a method is followed by a burst of research applications, then by a cautionary review of methodological problems with those applications. For example, Nagin (latent class growth, Nagin, 1999) and Muthén (growth mixture, Muthén & Shedden, 1999) introduced exploratory variants of latent trajectory modeling that classified individuals into latent classes based on commonalities in their patterns of change over time. Research on latent class growth and growth mixture took off rapidly, with hundreds of applied examples in the years following the publication of these articles (Bauer, 2007). Subsequently, a series of methodological articles began to suggest that the solutions obtained from latent class and growth mixture models were not as deterministic as applied authors reported them. For example, Bauer and Curran (2003a, 2003b) showed that latent class solutions were influenced by skew and non-normality in the data. Jackson and Sher showed that class solutions were influenced by the number and timing of assessments, using different indicators of the same outcome, or by the threshold that was used to define a cut score (K. M. Jackson & Sher, 2005, 2006, 2008). Finally, Sher, Jackson, and Steinley (2011) showed that studies repeatedly identified the same four trajectories of change over time (i.e., the "cat's cradle" pattern of steady high, steady low, increasing, and decreasing classes). In short, latent class/growth mixture solutions seemed to be strongly influenced by methodological factors, and generally should be taken as exploratory rather than confirmatory, yet applied researchers have generally ignored these cautions (Bauer, 2007).

We do not argue that clinical science has failed to adopt novel quantitative methods, but that the adoption of novel methods is not followed by a substantial increase in the quality of research using those methods. In other words, new methods may be adopted, but adoption is slow, partial, and without fidelity. Often the process of adoption, even with high fidelity, produces new problems (e.g., McNeish et al., 2016; Sher, Jackson, & Steinley, 2011) and threatens the development of a cumulative clinical science. From an implementation science perspective, this may be because the current model of spread is largely reliant on diffusion and dissemination (defined below), with little attention to the determinants that shape the implementation of new practices or the strategies that might improve their implementation. Replicability in clinical psychology will not be improved by mere transparency if reviewers and readers are ill-equipped to understand statistical models. Statistical directives such as changing *p* value cutoffs (Benjamin et al., 2018) or requiring effect size estimates and confidence intervals (Cumming, 2014) or Bayes factors (Wagenmakers et al., 2018), will not be effective if the average reader or reviewer cannot understand whether the model that produced those statistics is trustworthy. New methods will not be integrated if universities do not support statistical packages that facilitate these methods, or faculty/instructor professional development to bring appropriate training and consultation in-house.

How Can We Improve the Application of Quantitative Methods in Research?

Diffusion and dissemination. In order to build toward a replicable, cumulative science of clinical psychology, it is critical to improve how applied researchers use, interpret, and report statistical analyses (which we generally refer to below as "methods"). To date, researchers have tried a range of strategies representing diffusion, dissemination, and implementation activity to promote change in quantitative methods. Diffusion—or the passive spread of innovations—is the most common strategy. Much like in the clinical literature where treatment developers would rely on manuals to diffuse information to providers, quantitative researchers rely on articles in methodological journals to describe methodological innovations. Dissemination—delivery of targeted information to specific audiences—is also common, where researchers write tutorials explaining innovations for applied audiences, or offer training through websites, workshops, preconferences, or summer institutes. Similar to the clinical intervention literature where continuing education workshops are the status quo, these brief trainings are insufficient for achieving competent application. For example, MacCallum's cautions against dichotomization of continuous variables is highly cited (MacCallum et al., 2002), and has likely influenced a generation of researchers to avoid this questionable practice. However, it is also important to note that this advice followed articles one (Hunter & Schmidt, 1990) and two (Cohen, 1983) decades prior that essentially delivered the same cautions, and itself was followed by further research attempting to understand why researchers might dichotomize a continuous variable (DeCoster, Iselin, & Gallucci, 2009), highlighting the slow and incomplete nature of dissemination.

Active efforts to integrate new practices in routine settings, has been more rare. Similar to treatment guideline efforts, the most popular discrete implementation strategy is the mandate or guideline. Indeed, some organizations have offered analysis or publication guidelines or standards around statistical methods. For example, experts in clinical trials methodology developed the CONSORT guidelines to improve the reporting of randomized control trials (RCTs; Moher et al., 2012), which was simultaneously published in dozens of medical journals. The *APA Publication Manual* provides standards for reporting many basic statistics (American Psychological Association, 2010), and editors and editorial boards have tried to shape how authors report studies, such as banning or requiring more stringent *p* values (Finch et al., 2004; Ioannidis, 2018), or implementing incentives to promote specific practices such as data or materials sharing (Kidwell et al., 2016).

Systematic attempts to improve the quality of data analysis in clinical science will require multifaceted strategies. First, to create or choose effective strategies to implement statistical methods that will result in widespread adoption and sustainment of new statistical methods in appropriate ways, it is important to understand the determinants, or barriers and facilitators, of methods implementation. Second, methodologists will need to develop ways to measure methods "fidelity" (i.e., adherence to best practices and competency in using a statistical method), and ways to communicate this information to target stakeholders (e.g., trainees, scientists, reviewers, editors, funders, organizations). To understand success, it may be important to consider other implementation outcomes beyond fidelity, such as adoption, sustainment, and or penetration, to evaluate implementation efforts. Given that much of quantitative research focuses on *efficacy* (i.e., whether a method provides accurate parameter estimates under varying simulated conditions), it will be less important to define and measure whether or not a method works, but instead to understand how well a suggested method works when used by applied researchers in real data. Finally, it will be critical to develop evidence-based implementation strategies that can effectively change practices by targeting the known determinants of methods implementation.

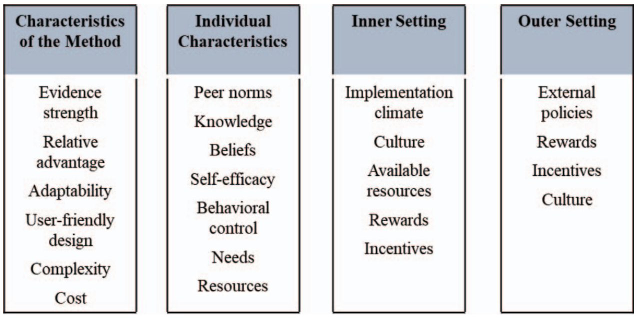
One important principle of applying implementation science to improving methods is that it is best applied to methods that applied researchers will frequently use. Moreover, many approaches to disseminating research methods are aimed at developing competent users, rather than experts. Much like the community therapist who sees a wide variety of patients with diagnoses that might require a wide variety of EBPs (Chorpita, Daleiden, & Weisz, 2005b; Weisz et al., 1995), a typical applied user may be required to competently use or evaluate a wide variety of data analytic methods. Thus, an implementation approach considers how to support a generalist approach to data analysis for most applied researchers. In this way, a quantitative implementation science perspective can coexist with the team science approach, where projects are conducted by teams of subject matter experts rather than generalists, and which would argue against the need for broad implementation of statistical methods (Bennett & Gadlin, 2012; Börner et al., 2010). Indeed, some have argued for subject matter experts for psychological disorders that are rare or involve highly complex treatments that are difficult to sustain with fidelity (Comer & Barlow, 2014), and it will be similarly important to identify statistical methods that have a low cost-benefit ratio for many to learn and implement, and therefore might more efficiently rely on expert consultants.

However, it is still important that all researchers understand fundamental statistical principles like probability, correlations, and *t* tests, as well as the tools used for inference like NHST, effect sizes, and confidence intervals. Without this basic understanding, applied psychologists would be unable to develop a sense of a cumulative literature, as their ability to review literature would be limited to checklists of which studies did or did not find significant effects, ignoring direction, magnitude, precision, or potential moderators of effects. Moreover, if it is true that well-funded clinical trials are likely to have a dedicated methodologist, several reviews suggest that the reporting of both preclinical (Landis et al., 2012) and clinical trials (Turner, Shamseer, Altman, Schulz, & Moher, 2012) are inadequate, even with widely disseminated reporting and data-analytic guidelines for the analysis of clinical trials (e.g., CONSORT, Schulz, Altman, & Moher, 2010). In short, even in science teams, all researchers will at least need generalist training in the basics of quantitative methods.

Implementation Determinants

There are many factors that might influence the adoption of new methods or implementation of new methods with fidelity. Damschroder and colleagues integrated a large body of existing knowledge on organizational change to catalog the multilevel influences on implementation of innovations in their Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009). This perspective recognizes that there are multiple systems and stakeholders at many levels that influence the implementation of any innovation, and that characteristics of the intervention, the individual who applies it, and the context in which it is applied all may influence implementation success. Although implementation science has not developed a formal unifying theory of behavior change, all frameworks, theories, and models describe implementation as being influenced by multiple levels of the ecological system in which it occurs emphasizing the influence of interactions between individuals (and their knowledge, motivations, and beliefs) and their environments to ultimately influence behavior and behavior change (Glanz & Bishop, 2010; Richard, Gauvin, & Raine, 2011). In other words, it is not enough to increase applied researchers' knowledge, skills, or motivation to use a new analytic method if their research environment does not facilitate or support the use of that method, if the method itself is too difficult or expensive to apply, or if journal or grant reviewers expect or request that inferior methods are used. Figure 1 lists the CFIR constructs across multiple levels of influence.

Characteristics of the method. Implementable interventions are likely to have an *evidence base* that suggests the intervention provides some *relative advantage* to existing methods (Cadarette et al., 2017). In terms of methods, this may be evidence that a method can answer questions that are of interest to researchers, that will match (or is robust to) the types of data that they face, and be better than what they are currently doing. An interesting example of this is that hierarchical linear models almost entirely dominate the analysis of nonindependent data, even while alternative methods (such as generalized estimating equations) can provide a more parsimonious, robust, and often analytically equivalent way of analyzing clustered data (McNeish et al., 2018). Making the intervention *adaptable* to a researchers' specific needs, and that have a *user-friendly design* with low *complexity* will also influence adoption; interventions that are seen as flexible and relatively easy to use are much more likely to be adopted than those which are inflexible or have a high skill or knowledge barrier to use. The popularity of the PROCESS macros for testing mediation and moderation (Hayes, 2012), and Excel spreadsheets (Dawson, 2014) or online tools (McCabe, Kim, & King, 2018; Preacher, Curran, & Bauer, 2006) for probing interactions, highlight the importance of reducing complexity as a means of increasing adoption of quantitative methods. In a review of statistical practices in the published literature, Kirk (1996) noted that some statistical information was more common (such as R^2 as a measure of effect size in regression) than others because it was commonly included in statistical output. Finally, the *cost* of a new quantitative method will certainly influence adoption, either in terms of the financial cost of software or the time-cost required to learn a method in a given software package.



Characteristics of the Method	Individual Characteristics	Inner Setting	Outer Setting
Evidence strength	Peer norms	Implementation climate	External policies
Relative advantage	Knowledge	Culture	Rewards
Adaptability	Beliefs	Available resources	Incentives
User-friendly design	Self-efficacy	Rewards	Culture
Complexity	Behavioral control	Incentives	
Cost	Needs		
	Resources		

Figure 1. Multilevel determinants of quantitative methods implementation. See the online article for the color version of this figure.

Individual characteristics. Implementation science has drawn upon multiple models of behavior change (Azjen, 1991; Gerrard, Gibbons, Houlihan, Stock, & Pomery, 2008; Glanz & Bishop, 2010; W. R. Miller & Rose, 2009), to understand inter- and intrapersonal factors that influence behavior change, and we believe it may also be applied to the implementation of research methods. The presence of rules of thumb and heuristics about data analysis indicate that *injunctive and descriptive peer norms,* or the belief that other researchers, reviewers, editors, or funders use certain methods, and believe that using some methods are acceptable, likely influence the adoption of new quantitative methods. This can both foster or dissuade good practices; for example, a review of management researchers suggested that many questionable research practices (such as HARKing or selective reporting of results) were actually encouraged by reviewers or editors (Banks et al., 2016). Researchers will be more likely to adopt an innovative method if they have *knowledge and positive beliefs about the method, self-efficacy* to use the method themselves, they are *have perceived behavioral control that they can adopt a new method*, and the new method matches their *needs and available resources*

(Azjen, 1991; Cook, Lyon, Kubergovic, Browning Wright, & Zhang, 2015). Finally, motivation and goals are critical. Researchers should believe it is *important* and *feasible* for them to adopt and use a new method, be interested in adopting new analytic methods (W. R. Miller & Rose, 2009), and see a relative benefit of adopting a new method that is not in conflict with their personal goals (such as publishing and obtaining funding). Motivations and goal conflicts may loom large, especially when there are discrepancies between what is best for science (building a credible evidence base) and what is best for scientists (publications, grants, and awards; J. Sakaluk, personal communication, October 19, 2018). These conflicts may be especially strong when a field has yet to adopt new norms around open and transparent practices in science.

Inner setting. Finally, it is important to acknowledge that implementation will be more widespread when the organizational context where the research is conducted (i.e., the inner setting) has a positive *implementation climate*, defined as one that expects, supports, and rewards implementation of the innovation (Klein & Sorra, 1996). The *culture* of scientific contexts (i.e., the values, norms, and basic assumptions of an organizational context) will influence successful implementation. For example, support for teaching advanced (or improved) quantitative classes in training programs will be affected by perceptions of faculty about the need for statistical reform, as well as the availability of expert resources (such as quantitative psychologists) to provide training and consultation. As noted above, the evidence suggests that the availability of quantitative experts in doctoral programs is declining even as the complexity of analytic methods is increasing (Aiken, West, & Millsap, 2008). *Organizational rewards and incentives*, such as hiring, promotion, and tenure criterion that incentivize best practices will also influence implementation. Unfortunately, models of research success that emphasize individual scientific contributions (e.g., acting as a principle investigator for research grants or serving as the first or senior author for research articles) explicitly disincentivize activities that might promote improved implementation of methods in an organization. For example, individuals that work to bridge the gap between quantitative and applied science may be underrecognized by organizations that do not reward collaborative research efforts (Sharpe, 2013).

Outer setting. The larger context (such as professional organizations, journals, and funding agencies) in which research occurs will also influence implementation, and some aspects of the inner setting (such as rewards, incentives, and culture) will also be observed in the outer setting. *External policies* that are enacted by journals (i.e., Finch et al., 2004; Kidwell et al., 2016) or funding agencies (such as the new "Rigor and Reproducibility" section for NIH grants), and *incentives* (Maner, 2014; such as recognition for open science practices) will be critical to promoting improved statistical practices.

Strategies to Improve Implementation

Although there are many strategies that can be used to improve implementation (Powell et al., 2015), we focus on a subset here that are especially relevant to quantitative methods given the known determinants that are relatively ubiquitous: exnovation, shaping, and augmentation; training; consultation and supervision; feedback; and guidelines and standards. Importantly, implementation strategies (such as checklists or feedback tools) should not be viewed as substitutes for improving statistical learning among researchers, but rather tools that can augment existing knowledge to maintain fidelity to best practices. Implementation will be maximized when applied researchers' statistical thinking is improved, and when systems, contexts, and tools facilitate optimal statistical practices.

De-implementation, shaping, and augmentation. Improving the implementation of quantitative methods will involve finding methods to encourage researchers to abandon ineffective or inferior practices (*de-implementation*), shape or augment existing practices to improve them, as well as increase the use of novel and/or more effective practices (Prasad & Ioannidis, 2014). A de-implementation goal might be to stop using null hypothesis testing (Loftus, 1996), or to stop interpreting coefficient alpha as a test of a scale's unidimensionality and validity (Sijtsma, 2009). One example of shaping is reflected in developments in mediation. As awareness of the shortcomings of the Sobel test for estimating the significance of mediated effects has grown (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002), researchers have increasingly used bias-corrected bootstrap standard errors, joint significance and Monte Carlo methods, although the Sobel test is still frequently applied (Hayes & Scharkow, 2013). This reflects changes in some aspects of how tests of mediation are conducted (estimating significance and confidence intervals), but not others (computing the mediated effect using the product of coefficients). An example of augmentation, adding materials or components to improve an existing quantitative method, is represented in the development of graphical tools for testing and interpreting interactions which plot confidence intervals, regions of significance, and even the underlying data (McCabe et al., 2018; Preacher et al., 2006), but do not represent a substantive change in the computation or interpretation of interactions.

Training. There are numerous problems with how statistics are taught to undergraduate and graduate students. Classes are often taught by statisticians rather than applied analysts, who bring an overly mechanical focus on mathematics. When combined with students' own anxiety about math and statistics, this leads to low statistical literacy and difficulty generalizing statistical knowledge to new substantive content (Tishkovskaya & Lancaster, 2012). Students often have limited access to faculty who have primary training in statistics or methods as well as a shortage of quantitative training programs (Aiken, West, Reno, Kazdin, & Sherman, 1990). Graduate programs infrequently require full courses in advanced methods (Aiken et al., 2008), and often provide only limited training in statistical approaches and methodologies specific to clinical psychology (Bransford, Brown, & Cocking, 2000).

The shortage of methods training is especially problematic for students in APA-accredited clinical psychology doctoral programs. APA accreditation standards are vague as the to the level of statistical competence students must achieve. The implementing regulations state only that doctoral trainees should be exposed to "doctoral level training in research methods, statistical analysis, and psychometrics," but acknowledge that programs will vary in what they include in this topic of discipline specific knowledge (Commission on Accreditation, 2017). One study of clinical psychology programs reported findings in line with the work of Aiken and colleagues (Aiken et al., 1990; 2008): Nearly all programs required courses in experimental design and analysis, regression, and introductory measurement, but few programs required methodological courses beyond these foundational courses, and <10% of clinical program directors believed this training sequence was insufficient (Rossen & Oakland, 2008). This is compounded by the time demands placed on students both in terms of the typical amount of clinical training required to become proficient in clinical practice, as well as the number of additional courses that are required by the APA to obtain a breadth of training in all of psychology. This substantially limits clinical doctoral students' ability to obtain more advanced training in research methods and to effectively prepare them for a career in clinical psychological science.

There are also a variety of professional development opportunities, workshops, and courses offered as preconference meetings, stand-alone daylong (or multiday) workshops, or online courses, but there is no research on the quality or efficacy of professional development for quantitative methods, or even descriptive studies of their content or instructional approaches. The literature on adult professional education, and specifically of postgraduate medical and mental health education, suggests that the impact is likely to be limited, particularly if they emphasize didactic instruction and do not provide opportunities for supervised practice with feedback (Davis et al., 1995; Lyon, Stirman, Kerns, & Bruns, 2011). A substantial body of evidence has suggested that passive learning methods, such as reading textbooks or articles or attending didactic workshops, may produce small increases in knowledge, but generally fail to produce lasting changes in practices that are sustained over time, or produce mastery of new skills (Beidas & Kendall, 2010; Herschell, Kolko, Baumann, & Davis, 2010).

For both pre- and postgraduate training, it will be critical to apply best practices in education to statistics training. Lovett and Greenhouse (2000) outlined principles of best practices in statistics education, such as that (a) students learn best what they practice, (b) knowledge is specific to the context of learning, (c) learning is more efficient when errors are corrected quickly, (d) new knowledge must be integrated into existing knowledge, and (e) efficiency of learning decreases under higher cognitive load (pp. 196–197). These principles mirror the broader evidence base that effective learning is influenced not only by the specific practices of teaching, but also by the multiple contexts in which learning occurs (Bransford et al., 2000; Rakovshik & McManus, 2010). Statistical courses and workshops thus may be maximally effective when students are able to practice their new knowledge on data sets that are familiar to them (e.g., by writing the results section for an empirical research paper as the final assignment for a course), and when they are able to get feedback from instructors on their work, but it will be critical to empirically evaluate this hypothesis.

Consultation and supervision. A common challenge with statistical education is that there is some time between when knowledge is acquired and when it is applied. Thus, ongoing consultation and supervision is often critical to consolidate knowledge, and to ensure and maintain ongoing fidelity. Even trainings that do produce increases in knowledge or skills (such as training in motivational interviewing) have been shown to produce reversals of such gains when they were not paired with ongoing consultation or supervision (W. R. Miller, Yahne, Moyers, Martinez, & Pirritano, 2004). Moreover, the content of consultation and supervision is important. Models of supervision from clinical research suggest that "gold standard" supervision is comprised of fidelity monitoring, skill building, review of actual practices and symptom monitoring (Dorsey et al., 2013). Behavioral rehearsal may be an especially effective skill building tool as it efficiently combines a means of fidelity monitoring, skill building, and review of practices (Dorsey et al., 2017, 2013). However, it is important to note that in community practice settings, behavioral rehearsal and other methods for observing clinician skill are rarely used (Dorsey et al., 2018). However, these methods may be used more often in organizations with more positive implementation climates (Lucid et al., 2018). This literature would suggest that the optimal quantitative or methodological consultation and supervision would consist of some version of behavioral rehearsal (such as coreviewing code, or analyzing data with a trainee).

Feedback. Feedback, or giving providers data on their performance, is thought to be critical to improving professional practice. A meta-analysis of 118 RCTs suggested that feedback can produce small to medium sized effects on the performance of professionals in health care settings (Jamtvedt, Young, Kristoffersen, O'Brien, & Oxman, 2010). Feedback on quantitative methods may be easy to develop, in one way, in that many diagnostic tools (such as residual plots) and interpretative guidelines could be automatically provided in statistical output. Feedback may be more successful when it focuses on improving task motivation and learning or when paired with a clear goal, while praise or discouraging feedback may actually worsen performance (Kluger & DeNisi, 1996).

Guidelines and standards. Manuals and practice guidelines have become popular methods for disseminating standards of evidence based practice (Corrigan, Steiner, McCracken, Blaser, & Barr, 2001). Many statistical methods have excellent textbooks that serve as resources to practitioners, and guidelines are similarly widespread. Lang and Altman (2016) developed a comprehensive set of guidelines for reporting standards of most statistical methods reported in the biomedical literature. The Enhancing the Quality and Transparency of Health Research (EQUATOR) network (www.equator-network.org) compiles reporting guidelines for a wide variety of study designs. Some research on reporting guidelines suggests that they can indeed improve reporting practices in journals, but also suggests that these improvements are relatively modest (Peng, Chen, Chiang, & Chiang, 2013). However, in the same way that most clinical psychologists do not adopt and follow treatment manuals exactly (Chorpita, Becker, Daleiden, & Hamilton, 2007), statistical manuals and guidelines must be implemented in a way that maximizes their efficacy. Moreover, to improve sustainability, their efficacy should be evaluated, as effective guidelines or manuals may be more likely to be sustained over time. During Geoff Loftus' term as editor of *Memory and Cognition*, for example, *p* values were banned, but almost immediately returned in the subsequent editor's term (Finch et al., 2004), suggesting that the guidelines may have been effective in terms of reducing reports with *p* values, but were unsustainable without broader support from the journal or community of scientists publishing in that journal. Another review of 150 articles in the *Journal of Applied Psychology* suggested that APA guidelines on statistical reporting were inconsistently applied (Finch, Cumming, & Thomason, 2001). Guidelines are less likely to be followed when they require new knowledge or skills, or when they conflict with the norms of an organization, while those that are easy to understand and implement which can utilize existing resources are more likely to be followed (Burgers et al., 2003; Francke, Smit, De Veer, & Mistiaen, 2008).

It may be that more active and adaptive deployment of innovations would improve their efficacy. For example, Chorpita's managing and adapting practice (MAP) model leverages existing intervention content elements drawn from hundreds of trials and packages them in an adaptive system that provides resources to aid practitioners to deliver optimal, evidence based practices for their patient population (Chorpita, Daleiden, & Collins, 2014; Southam-Gerow et al., 2014). Redesign of the typical ways statistical methods are packaged and made available to users could similarly guide applied researchers toward a statistical model that fits a particular problem they are trying to solve. However, if there is not an exact match (e.g., the distributional properties of the data do not fit the proposed analysis), there could be suggested analyses or methods to augment analysis, to provide some optimal balance between rigidly applying some analytical standard and ignoring problems in data or analyses (Chorpita, Daleiden, & Weisz, 2005a).

Checklists are another example of guidelines that may be especially effective in ensuring quality practice. Checklists have been proposed to aid in peer review (Moher et al., 1995; Parker et al., 2018), to improve qualitative research (Power & Williams, 2001), or to evaluate RCTs (Turner et al., 2012). For example, Lang and Altman (2016) published the Statistical Analysis and Methods in the Published Literature (SAMPL) guidelines, intended to serve as a generic "guide to authors" for reporting a wide variety of statistical analyses common to biomedical research). Checklists can serve as memory aides even for overlearned tasks to ensure the consideration of all steps in a complicated process. However, it is important to note that some evaluations have suggested that checklists may have a limited impact, and it may depend on the degree to which checklists are deployed passively or actively. One study suggested that journals that enforced reporting standards for RCTs (the CONSORT guidelines; Schulz et al., 2010) exhibited improved, but still suboptimal, reporting of clinical trials. An RCT suggested that a statistical peer reviewer improved the quality of analyses in peer reviewed research, but suggesting a checklist of reporting guidelines to reviewers did not (Cobo et al., 2007). Finally, job aids, which are well-crafted, visualized forms of performance support tools, achieve many of the same functions of checklists by prompting specific behaviors at the point where they are typically performed. Job aids may be digital or printed and should address some of the most common, avoidable mistakes that tend to occur in the application of statistical methods.

Implementation Outcomes

It will be important to measure outcomes of implementation efforts in order to explain successes and failures in implementing methodological changes in practice. Proctor and colleagues developed a taxonomy of outcomes of implementation efforts, which can be thought of as the effects of deliberate efforts to implement new interventions (Proctor et al., 2011). Outcomes include fidelity (composed primarily of adherence and competence), acceptability, appropriateness, feasibility, cost, adoption, penetration, and sustainability, and are distinguished from the effectiveness of an intervention and whether or not applying the intervention solves a particular problem (Crable et al., 2018). These are the primary outcomes of implementation research, and it is important that psychometrically sound instruments are used to measure implementation outcomes in order to more clearly understand the relative effectiveness of specific implementation strategies (Lewis et al., 2015).

To measure fidelity, it may be most important to target *adherence to* and *competence with* a methodological approach, or the degree to which and how well an intervention is implemented in the manner in which it was intended. Attending to adherence and competence allows for insights into how and why implementation failures occurred and whether failures are due to the method or its application (Schoenwald et al., 2011). For example, adherence to principles of motivational interviewing can be coded using well validated measures (W. R. Miller, Moyers, Ernst, & Amrhein, 2008). Although no measures of adherence or competence for specific quantitative methods exist, adherence is a regular target of narrative reviews in quantitative methods which describe how well common published research follows best practices. For example, Credé and Harms (2018) showed that a substantial proportion of research in the management literature reported the results of confirmatory factor analyses in a way that underemphasized potential model misfit (such as selectively ignoring chi-square, reporting degrees of freedom inconsistent with the specified model, ignoring model residuals, or failing to present an alternative measurement model).

Quantitative scientists must develop easy to use adherence guides for specific methods, measures of competency to allow self or external evaluation of an applied researchers' skillset, and systems and standards (such as open sharing of syntax, model results, and data) that facilitate external evaluation. For example, the book *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (Hancock & Mueller, 2010) provides a broad overview of statistical methods to aide readers' evaluation of the use of methods in research. Of course, no guide will be perfect or be able to cover all research problems or all data sets, but it will be crucial to decide on what best practices might be for a given method *for most circumstances and most data*. It may also be difficult for methodologists to agree on definitions of adherence or competence. For example, a special issue of *Personality and Individual Differences* in 2007 published conflicting perspectives from multiple authors on model fit for SEM. However, if quantitative scientists do not provide adherence or competency guides, then it will be impossible for anyone but methodological experts to effectively conduct or evaluate research, and the evaluation may depend on who the expert is. This "oracle model" limits the potential reach of any given quantitative method, and fails to improve implementation of all methods.

Although valid adherence measures have been developed for some mental health interventions, they have often failed to become widespread. This may be in part because more valid measures of fidelity (such as coding of actual practice or behavioral rehearsal) are also more time-intensive and difficult to administer in practice (Dorsey et al., 2017). It will be critical to develop measures of fidelity that maximize efficiency while not sacrificing validity (Schoenwald et al., 2011).

Some implementation outcomes focus on perceptions held by end-users and other individuals who influence the use of a method. *Acceptability* is the degree to which a particular intervention is viewed as agreeable or reasonable. For quantitative methods, acceptability can be viewed from multiple perspectives, such as the researcher, the reader, funders, editors, and reviewers. For example, in the experience of the first author, it was common in the mid-2000s to need to justify the use of bootstrapped standard error tests of indirect effects by supplementing these analyses using the Baron and Kenny (1986) causal steps-approach to mediation), suggesting that reviewers preferred the latter approach. In current practice, using bootstrapped standard errors to test indirect effects is deemed acceptable, and the Baron and Kenny approach inferior. *Appropriateness* is how well a given intervention is seen to fit a particular problem. Proctor and colleagues (2009) distinguished between appropriateness and acceptability, arguing that interventions that might be appropriate might be viewed as unacceptable, or vice versa. For example, Grimm (2007) illustrated how three different types of longitudinal models (bivariate latent growth curve, latent growth with a time-varying covariate, and bivariate latent change score models) could be fitted to the same dataset while answering slightly different types of questions. *Feasibility* of interventions is the extent to which a new intervention can be used in a given setting. New methods that are implemented in R (R Core Team & R Development Core Team, 2016), a statistical software that requires a great deal of computer programming to run effectively, may not be seen as feasible for researchers with little to no background in programming, while software that runs on a "point-and-click" basis (e.g., in SPSS) may be far more feasible. For example, we recently developed a point and click R-based Shiny app to facilitate optimized plots of interactions (McCabe et al., 2018). *Cost* is the cost in terms of both time and money required to effectively implement a new intervention. Cost can be influenced by the monetary cost of statistical software or training (both can be hundreds of dollars per user), as well as the time it takes for an applied researcher to learn how to use the new intervention. Again, R is a good example of a software that has low financial costs to acquire but a higher time investment cost to individuals to use and master.

Other implementation outcomes reflect use in organizational settings (such as labs, research groups, centers, departments, or professional organizations). *Adoption* refers to the degree to which a new intervention is tried by individuals or organizations. For example, epidemiological researchers showed that two approaches to handling confounding (disease risk score [DRS] and high dimensional propensity scores [hdPS]) had two dramatically different rates of adoption over time (Cadarette et al., 2017). DRS were first introduced in 1976, but did not see widespread adoption until 2000, while hdPS were first introduced in 2009 and rapidly adopted. *Penetration* is how broadly a new intervention is used within specific organizational settings. Penetration could be measured in terms of how many researchers in a lab, research group or department use a particular intervention, or it could be measured at the level of the journal or subdiscipline. Linear mixed models have much higher penetration in psychology compared to GEEs (McNeish et al., 2016). *Sustainability* is the degree to which a new intervention is maintained over time within specific settings, which often recognizes that sustainable interventions will need to be dynamic and adaptable to changing organizational or individual circumstances (Shelton, Cooper, & Stirman, 2018).

Theoretically, successful implementation (i.e., positive implementation outcomes) is the precursor to the effectiveness of interventions in practice. Without implementation success, a quantitative intervention may be effective, in that it can produce appropriate and accurate answers, but if it is not perceived to be acceptable, feasible, or sustainable, adoption may be low or it could be widely adopted but used with poor fidelity, so it is not very effective. In other words, successful implementation is defined by widespread, appropriate, and accurate use of a method.

It is important to acknowledge that it is not immediately obvious who should be expected to take responsibility for driving these changes. Structural changes at the level of journals, funding agencies, departments, and scientific organizations may be enacted by any number of roles, but may be inhibited by diffusion of responsibility. Quantitative psychologists are rarely hired or promoted on the basis of their ability to translate basic research to applied settings, but rather (like other psychologists) on their original contributions to science. Thus, it will also be important to incentivize research around improving best practices in order to facilitate change.

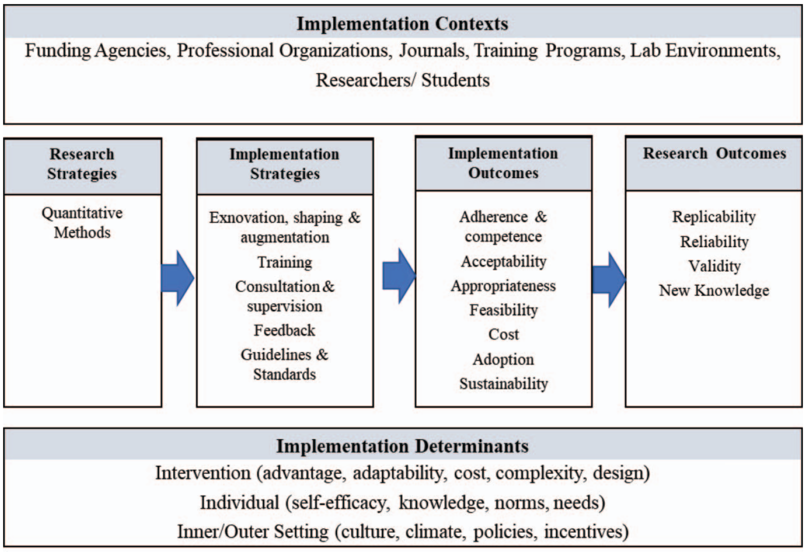
Future Directions

A wave of reform aims to change how research is conducted in the psychological sciences, including clinical psychology (Tackett et al., 2016; Vazire, 2018). Above, we have outlined the parallels in the research-to-practice gap for clinical interventions and quantitative methods, and how the solutions applied to clinical interventions might translate to quantitative methods. We have refrained from proposing concrete solutions, as one important lesson of implementation science is that success will depend on interventions that effect change across multiple systems in multiple ways. Given the paucity of research on how to best implement quantitative methods, we view this article as an outline of future research areas rather than a prescription for change. Figure 2 summarizes our proposed model of quantitative implementation science, which we adapted from the proposed framework of Proctor and colleagues (2009).

An important caveat is that although the conversation around replicability has only just begun in clinical psychological science (Tackett, Brandes, King, & Markon, 2019; Tackett et al., 2016), it has yet to reach implementation science. Thus, it is possible that much of the research now being conducted based on ideas from implementation science is based on research findings that are themselves not replicable. That said, one norm that is very common in implementation science is publishing trial protocol articles before the onset of implementation trials (e.g., Dorsey et al., 2013; Hartzler et al., 2017), which, similar to preregistration, may promote more transparent research practices and increase confidence in the conclusions of implementation studies. However, as researchers apply the principles of implementation science to methods research, it will also be important to attend to the replicability of implementation science methods themselves.

It is important to recognize that responsibility for narrowing the gap between optimal and typical practices should not only lie with methodologists. Given the incentives for innovation in the field, it is understandable that quantitative psychology has focused more on producing novel solutions to analytical problems rather than improving the conduct or interpretation of existing analyses. Given few consequences for not ensuring that data analyses follow best practices (e.g., shortage of quantitative reviewers), the difficulty staying current with the methodological literature, and the incentives for innovation in data analysis, it is understandable that applied researchers have not always been motivated to change.

The lessons of other disciplines as filtered through the lens of implementation science are clear: Rebellions are not built on


Implementation Contexts
Funding Agencies, Professional Organizations, Journals, Training Programs, Lab Environments, Researchers/ Students
Research Strategies
Quantitative Methods
Implementation Strategies
Exnovation, shaping & augmentation
Training
Consultation & supervision
Feedback
Guidelines & Standards
Implementation Outcomes
Adherence & competence
Acceptability
Appropriateness
Feasibility
Cost
Adoption
Sustainability
Research Outcomes
Replicability
Reliability
Validity
New Knowledge
Implementation Determinants
Intervention (advantage, adaptability, cost, complexity, design)
Individual (self-efficacy, knowledge, norms, needs)
Inner/Outer Setting (culture, climate, policies, incentives)

Adapted from Proctor, Landsverk, Aarons, Chambers, Glisson, & Mittman, 2009

Figure 2. A proposed model for research on quantitative methods implementation. See the online article for the color version of this figure.

merely on hope (cf. Kennedy, Shearmur, Emanuel, & Edwards, 2016). Instead, the gap between optimal and typical practice will only be narrowed by the collective efforts of researchers applying the principles of implementation science, and the funding agencies that promote such research. As noted above, there is little to no research on any of the determinants, strategies, or outcomes that might promote implementation of quantitative methods in applied research. In the same way that clinical trials research has evolved to study how to deliver high quality treatment in realistic clinical settings, we argue that quantitative research should begin to study how to conduct high quality analyses in realistic research settings, and how to best train and support the research workforce to be capable of such analyses.

Researchers should aim to understand how applied researchers conduct, interpret, and evaluate different research methods, from the perspective of both the data analyst and the consumer of research. Future research should aim to identify the determinants of variation in quantitative methods practices, to understand what implementation strategies might be successful, how certain strategies might influence specific implementation outcomes, and how context plays a role in all dimensions of quantitative implementation. Systematic efforts will be required to produce lasting change in how data are analyzed, and in turn in the reliability, validity, and replicability of clinical psychological science.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16,* 270–301. http://dx.doi.org/10.1177/1094428112470848

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50. http://dx.doi.org/10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Reno, R. R., Kazdin, A. E., & Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology a survey of PhD programs in North America. *American Psychologist, 45,* 721–734. http://dx.doi.org/10.1037/0003-066X.45.6.721

Altman, D. G., & Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *Journal of the American Medical Association, 272,* 129–132. http://dx.doi.org/10.1001/jama.1994.03520020055015

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Azjen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50,* 179–211. http://dx.doi.org/10.1016/0749-5978(91)90020-T

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. http://dx.doi.org/10.3758/s13428-011-0089-5

Banks, G. C., O'Boyle, E. H., Jr., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management, 42,* 5–20. http://dx.doi.org/10.1177/0149206315619011

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. http://dx.doi.org/10.1037/0022-3514.51.6.1173

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42,* 757–786. http://dx.doi.org/10.1080/00273170701710338

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

Bauer, D. J., & Curran, P. J. (2003b). Overextraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods, 8,* 384–393. http://dx.doi.org/10.1037/1082-989X.8.3.384

Becker, E. M., Smith, A. M., & Jensen-Doss, A. (2013). Who's using treatment manuals? A national survey of practicing therapists. *Behaviour Research and Therapy, 51,* 706–710. http://dx.doi.org/10.1016/j.brat.2013.07.008

Beidas, R. S., & Kendall, P. C. (2010). Training therapists in evidence-based practice: A critical review of studies from a systems contextual perspective. *Clinical Psychology: Science and Practice, 17,* 1–30. http://dx.doi.org/10.1111/j.1468-2850.2009.01187.x

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396. http://dx.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., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2,* 6–10. http://dx.doi.org/10.1038/s41562-017-0189-z

Bennett, L. M., & Gadlin, H. (2012). Collaboration and team science: From theory to practice. *Journal of Investigative Medicine, 60,* 768–775. http://dx.doi.org/10.2310/JIM.0b013e318250871d

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16,* 202–207. http://dx.doi.org/10.1002/mpr.225

Bertram, R. M., Charnin, L. A., Kerns, S. E. U., & Long, A. C. J. (2015). Evidence-based practices in North American MSW curricula. *Research on Social Work Practice, 25,* 737–748. http://dx.doi.org/10.1177/1049731514532846

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., . . . Uzzi, B. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine, 2,* 49cm24. http://dx.doi.org/10.1126/scitranslmed.3001399

Borsboom, D., Cramer, A. O. J., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS ONE, 6,* e27407. http://dx.doi.org/10.1371/journal.pone.0027407

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies Press.

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological & Personality Science, 8,* 363–369. http://dx.doi.org/10.1177/1948550616673876

Burgers, J. S., Grol, R. P. T. M., Zaat, J. O. M., Spies, T. H., van der Bij, A. K., & Mokkink, H. G. A. (2003). Characteristics of effective clinical guidelines for general practice. *The British Journal of General Practice, 53,* 15–19.

Cadarette, S. M., Ban, J. K., Consiglio, G. P., Black, C. D., Dubins, D., Marin, A., & Tadrous, M. (2017). Diffusion of Innovations model helps interpret the comparative uptake of two methodological innovations: Co-authorship network analysis and recommendations for the integration of novel methods in practice. *Journal of Clinical Epidemiology, 84,* 150–160. http://dx.doi.org/10.1016/j.jclinepi.2016.12.006

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36,* 462–494. http://dx.doi.org/10.1177/0049124108314720

Chorpita, B. F., Becker, K. D., Daleiden, E. L., & Hamilton, J. D. (2007). Understanding the common elements of evidence-based practice: Misconceptions and clinical examples. *Journal of the American Academy of Child & Adolescent Psychiatry, 46,* 647–652. http://dx.doi.org/10.1097/chi.0b013e318033ff71

Chorpita, B. F., Daleiden, E. L., & Collins, K. S. (2014). Managing and adapting practice: A system for applying evidence in clinical care with youth and families. *Clinical Social Work Journal, 42,* 134–142. http://dx.doi.org/10.1007/s10615-013-0460-3

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005a). Identifying and selecting the common elements of evidence based interventions: A distillation and matching model. *Mental Health Services Research, 7,* 5–20. http://dx.doi.org/10.1007/s11020-005-1962-6

Chorpita, B. F., Daleiden, E. L., & Weisz, J. R. (2005b). Modularity in the design and application of therapeutic interventions. *Applied & Preventive Psychology, 11,* 141–156. http://dx.doi.org/10.1016/j.appsy.2005.05.002

Cobo, E., Selva-O'Callagham, A., Ribera, J. M., Cardellach, F., Dominguez, R., & Vilardell, M. (2007). Statistical reviewers improve reporting in biomedical articles: A randomized trial. *PLoS ONE, 2,* e332. http://dx.doi.org/10.1371/journal.pone.0000332

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

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253. http://dx.doi.org/10.1177/014662168300700301

Comer, J. S., & Barlow, D. H. (2014). The occasional case against broad dissemination and implementation: Retaining a role for specialty care in the delivery of psychological treatments. *American Psychologist, 69,* 1–18. http://dx.doi.org/10.1037/a0033582

Commission on Accreditation. (2017). *Implementing regulations Section C: IRs related to the guidelines and principles for doctoral graduate programs for internship programs for postdoctoral residency programs*. Retrieved from https://www.apa.org/ed/accreditation/section-c-soa.pdf

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C., & Ioannidis, J. P. (2008). Medicine. Life cycle of translational research for medical interventions. *Science, 321,* 1298–1299. http://dx.doi.org/10.1126/science.1160622

Cook, C. R., Lyon, A. R., Kubergovic, D., Browning Wright, D., & Zhang, Y. (2015). A Supportive beliefs intervention to facilitate the implementation of evidence-based practices within a multi-tiered system of supports. *School Mental Health, 7,* 49–60. http://dx.doi.org/10.1007/s12310-014-9139-3

Corrigan, P. W., Steiner, L., McCracken, S. G., Blaser, B., & Barr, M. (2001). Strategies for disseminating evidence-based practices to staff who treat people with serious mental illness. *Psychiatric Services, 52,* 1598–1606. http://dx.doi.org/10.1176/appi.ps.52.12.1598

Crable, E. L., Biancarelli, D., Walkey, A. J., Allen, C. G., Proctor, E. K., & Drainoni, M.-L. (2018). Standardizing an approach to the evaluation of implementation science proposals. *Implementation Science; IS, 13,* 71. http://dx.doi.org/10.1186/s13012-018-0770-5

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., . . . Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23,* 640–647. http://dx.doi.org/10.3758/s13423-015-0913-5

Credé, M., & Harms, P. (2018). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology, 34,* 18–30.

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

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311. http://dx.doi.org/10.1207/s15328031us0304_5

Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M., & Stetler, C. (2012). Effectiveness-implementation hybrid designs: Combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical Care, 50,* 217–226. http://dx.doi.org/10.1097/MLR.0b013e3182408812

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4,* 50. http://dx.doi.org/10.1186/1748-5908-4-50

Davis, D. A., Thomson, M. A., Oxman, A. D., & Haynes, R. B. (1995). Changing physician performance. A systematic review of the effect of continuing medical education strategies. *Journal of the American Medical Association, 274,* 700–705. http://dx.doi.org/10.1001/jama.1995.03530090032018

Dawson, J. F. (2014). Moderation in management research: What, why, when, and how. *Journal of Business and Psychology, 29,* 1–19. http://dx.doi.org/10.1007/s10869-013-9308-7

DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14,* 349–366. http://dx.doi.org/10.1037/a0016956

Dorsey, S., Kerns, S. E. U., Lucid, L., Pullmann, M. D., Harrison, J. P., Berliner, L., . . . Deblinger, E. (2018). Objective coding of content and techniques in workplace-based supervision of an EBT in public mental health. *Implementation Science, 13,* 19. http://dx.doi.org/10.1186/s13012-017-0708-3

Dorsey, S., Lyon, A. R., Pullmann, M. D., Jungbluth, N., Berliner, L., & Beidas, R. (2017). Behavioral rehearsal for analogue fidelity: Feasibility in a state-funded children's mental health initiative. *Administration and Policy in Mental Health, 44,* 395–404. http://dx.doi.org/10.1007/s10488-016-0727-4

Dorsey, S., Pullmann, M. D., Deblinger, E., Berliner, L., Kerns, S. E., Thompson, K., . . . Garland, A. F. (2013). Improving practice in community-based settings: A randomized trial of supervision - study protocol. *Implementation Science, 8,* 89. http://dx.doi.org/10.1186/1748-5908-8-89

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological & Personality Science, 7,* 45–52. http://dx.doi.org/10.1177/1948550615612150

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the journal of applied psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210. http://dx.doi.org/10.1177/00131640121971167

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., . . . Goodman, O. (2004). Reform of statistical inference in psychology: The case of memory & cognition. *Behavior Research Methods, Instruments, & Computers, 36,* 312–324. http://dx.doi.org/10.3758/BF03195577

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395. http://dx.doi.org/10.1037/0033-2909.112.3.393

Francke, A. L., Smit, M. C., de Veer, A. J. E., & Mistiaen, P. (2008). Factors influencing the implementation of clinical guidelines for health care professionals: A systematic meta-review. *BMC Medical Informatics and Decision Making, 8,* 38. http://dx.doi.org/10.1186/1472-6947-8-38

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23,* 98–122. http://dx.doi.org/10.1177/0959354312436870

García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology, 4,* 13. http://dx.doi.org/10.1186/1471-2288-4-13

Garland, A. F., Brookman-Frazee, L., Hurlburt, M. S., Accurso, E. C., Zoffness, R. J., Haine-Schlagel, R., & Ganger, W. (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services, 61,* 788–795. http://dx.doi.org/10.1176/ps.2010.61.8.788

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf%5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/a0037714

Gerrard, M., Gibbons, F. X., Houlihan, A. E., Stock, M. L., & Pomery, E. (2008). A dual-process approach to health risk decision making: The prototype willingness model. *Developmental Review, 28,* 29–61. http://dx.doi.org/10.1016/j.dr.2007.10.001

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33,* 587–606. http://dx.doi.org/10.1016/j.socec.2004.09.033

Glanz, K., & Bishop, D. B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual Review of Public Health, 31,* 399–418. http://dx.doi.org/10.1146/annurev.publhealth.012809.103604

Grimm, K. J. (2007). Multivariate longitudinal methods for studying developmental relationships between depression and academic achievement. *International Journal of Behavioral Development, 31,* 328–339. http://dx.doi.org/10.1177/0165025407077754

Hancock, G. R., & Mueller, R. O. (2010). *The reviewer's guide to quantitative methods*. San Francisco, CA: Health.

Hartzler, B., Lyon, A. R., Walker, D. D., Matthews, L., King, K. M., & McCollister, K. E. (2017). Implementing the teen marijuana check-up in schools-a study protocol. *Implementation Science; IS, 12,* 103. http://dx.doi.org/10.1186/s13012-017-0633-5

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science, 24,* 1918–1927. http://dx.doi.org/10.1177/0956797613480187

Herschell, A. D., Kolko, D. J., Baumann, B. L., & Davis, A. C. (2010). The role of therapist training in the implementation of psychosocial treatments: A review and critique with recommendations. *Clinical Psychology Review, 30,* 448–466. http://dx.doi.org/10.1016/j.cpr.2010.02.005

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

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349. http://dx.doi.org/10.1037/0021-9010.75.3.334

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 0696–0701.

Ioannidis, J. P. A. (2018). The proposal to lower p value thresholds to .005. *Journal of the American Medical Association, 319,* 1429–1430. http://dx.doi.org/10.1001/jama.2018.1536

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14,* 6–23. http://dx.doi.org/10.1037/a0014694

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19,* 339–351. http://dx.doi.org/10.1037/0893-164X.19.4.339

Jackson, K. M., & Sher, K. J. (2006). Comparison of longitudinal phenotypes based on number and timing of assessments: A systematic comparison of trajectory approaches II. *Psychology of Addictive Behaviors, 20,* 373–384. http://dx.doi.org/10.1037/0893-164X.20.4.373

Jackson, K. M., & Sher, K. J. (2008). Comparison of longitudinal phenotypes based on alternate heavy drinking cut scores: A systematic comparison of trajectory approaches III. *Psychology of Addictive Behaviors, 22,* 198–209. http://dx.doi.org/10.1037/0893-164X.22.2.198

Jamtvedt, G., Young, J., Kristoffersen, D., O'Brien, M., & Oxman, A. (2010). Audit and feedback: Effects on professional practice and health care outcomes. *Cochrane Database of Systematic Reviews*. Advance online publication. http://dx.doi.org/10.1002/14651858.CD000259

Jensen-Doss, A., & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology, 39,* 885–896. http://dx.doi.org/10.1080/15374416.2010.517169

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. http://dx.doi.org/10.1177/0956797611430953

Kar, S. S., & Ramalingam, A. (2013). Is 30 the magic number? Issues in sample size. *National Journal of Community Medicine, 4,* 175–179.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63,* 146–159. http://dx.doi.org/10.1037/0003-066X.63.3.146

Kennedy, K., Shearmur, A., Emanuel, S. (Producers), & Edwards, G. (Director) (2016). *Rogue One: A Star Wars Story* [*Motion Picture*]. United States of America: Walt Disney Studios.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14,* e1002456. http://dx.doi.org/10.1371/journal.pbio.1002456

King, K. M., Littlefield, A. K., McCabe, C. J., Mills, K. L., Flournoy, J., & Chassin, L. (2018). Longitudinal modeling in developmental neuroimaging research: Common challenges, and solutions from developmental psychology. *Developmental Cognitive Neuroscience, 33,* 54–72. http://dx.doi.org/10.1016/j.dcn.2017.11.009

King, K. M., Luk, J. W., Witkiewitz, K., Racz, S., McMahon, R. J., & Wu, J. (2018). Externalizing behavior across childhood as reported by parents and teachers: A partial measurement invariance model. *Assessment, 25,* 744–758. http://dx.doi.org/10.1177/1073191116660381

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759. http://dx.doi.org/10.1177/0013164496056005002

Klein, K. J., & Sorra, J. S. (1996). the challenge of innovation implementation. *Academy of Management Review, 21,* 1055–1080. http://dx.doi.org/10.5465/amr.1996.9704071863

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

Lance, C. E., & Vandenberg, R. J. (2008). *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. New York, NY: Routledge.

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., . . . Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490,* 187–191. http://dx.doi.org/10.1038/nature11556

Lang, T. A., & Altman, D. G. (2016). Statistical analyses and methods in the published literature: The SAMPL guidelines. *Medical Writing, 25,* 31–36.

Laurenceau, J.-P., Hayes, A. M., & Feldman, G. C. (2007). Some methodological and statistical issues in the study of change processes in psychotherapy. *Clinical Psychology Review, 27,* 682–695. http://dx.doi.org/10.1016/j.cpr.2007.01.007

Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science, 10,* 155. http://dx.doi.org/10.1186/s13012-015-0342-x

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171. http://dx.doi.org/10.1111/1467-8721.ep11512376

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305. http://dx.doi.org/10.1037/h0025105

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54,* 196–206.

Lucid, L., Meza, R., Pullmann, M. D., Jungbluth, N., Deblinger, E., & Dorsey, S. (2018). Supervision in community mental health: Understanding intensity of EBT focus. *Behavior Therapy, 49,* 481–493. http://dx.doi.org/10.1016/j.beth.2017.12.007

Lyon, A. R., Bruns, E. J., Ludwig, K., Stoep, A. V., Pullmann, M. D., Dorsey, S., . . . McCauley, E. (2015). The Brief Intervention for School Clinicians (BRISC): A mixed-methods evaluation of feasibility, acceptability, and contextual appropriateness. *School Mental Health, 7,* 273–286. http://dx.doi.org/10.1007/s12310-015-9153-0

Lyon, A. R., Stirman, S. W., Kerns, S. E. U., & Bruns, E. J. (2011). Developing the mental health workforce: Review and application of training approaches from multiple disciplines. *Administration and Policy in Mental Health, 38,* 238–253. http://dx.doi.org/10.1007/s10488-010-0331-y

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. http://dx.doi.org/10.1037/1082-989X.7.1.19

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1,* 173–181. http://dx.doi.org/10.1023/A:1026595011371

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83–104. http://dx.doi.org/10.1037/1082-989X.7.1.83

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editor) must change with them. *Perspectives on Psychological Science, 9,* 343–351. http://dx.doi.org/10.1177/1745691614528215

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science, 1,* 147–165. http://dx.doi.org/10.1177/2515245917746792

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100,* 43–52. http://dx.doi.org/10.1080/00223891.2017.1281286

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 58,* 7250–7257.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40–48. http://dx.doi.org/10.1037/0021-843X.110.1.40

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2008). *Manual for the motivational interviewing skill code (MISC)*. Unpublished manuscript.

Miller, W. R., & Rose, G. S. (2009). Toward a theory of motivational interviewing. *American Psychologist, 64,* 527–537. http://dx.doi.org/10.1037/a0016830

Miller, W. R., Yahne, C. E., Moyers, T. B., Martinez, J., & Pirritano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology, 72,* 1050–1062. http://dx.doi.org/10.1037/0022-006X.72.6.1050

Mills, L., Abdulla, E., & Cribbie, R. A. (2010). Quantitative methodology research: Is it on psychologists' reading lists? *Tutorials in Quantitative Methods for Psychology, 6,* 52–60. http://dx.doi.org/10.20982/tqmp.06.2.p052

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2012). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery, 10,* 28–55. http://dx.doi.org/10.1016/j.ijsu.2011.10.001

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16,* 62–73. http://dx.doi.org/10.1016/0197-2456(94)00031-W

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104,* 510–520. http://dx.doi.org/10.1258/jrsm.2011.110180

Murray, L. K., Dorsey, S., Bolton, P., Jordans, M. J. D., Rahman, A., Bass, J., & Verdeli, H. (2011). Building capacity in mental health interventions in low resource countries: An apprenticeship model for training local providers. *International Journal of Mental Health Systems, 5,* 30. http://dx.doi.org/10.1186/1752-4458-5-30

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469. http://dx.doi.org/10.1111/j.0006-341X.1999.00463.x

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4,* 139–157. http://dx.doi.org/10.1037/1082-989X.4.2.139

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. http://dx.doi.org/10.1037/1082-989X.5.2.241

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

Osborne, J. W. (2002, April). The effects of minimum values on data transformations. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Park, A. L., Tsai, K. H., Guan, K., & Chorpita, B. F. (2018). Unintended consequences of evidence-based treatment policy reform: Is implementation the goal or the strategy for higher quality care? *Administration and Policy in Mental Health, 45,* 649–660. http://dx.doi.org/10.1007/s10488-018-0853-2

Parker, T. H., Griffith, S. C., Bronstein, J. L., Fidler, F., Foster, S., Fraser, H., . . . Nakagawa, S. (2018). Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution, 2,* 929–935. http://dx.doi.org/10.1038/s41559-018-0545-z

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology, 68,* 1046–1058. http://dx.doi.org/10.1016/j.jclinepi.2015.05.029

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25,* 157–209. http://dx.doi.org/10.1007/s10648-013-9218-2

Powell, B. J., Waltz, T. J., Chinman, M. J., Damschroder, L. J., Smith, J. L., Matthieu, M. M., . . . Kirchner, J. E. (2015). A refined compilation of implementation strategies: Results from the Expert Recommendations for Implementing Change (ERIC) project. *Implementation Science; IS, 10,* 21. http://dx.doi.org/10.1186/s13012-015-0209-1

Power, R., & Williams, B. (2001). Checklists for improving rigour in qualitative research. Never mind the tail (checklist), check out the dog (research). *British Medical Journal, 323,* 514–515. http://dx.doi.org/10.1136/bmj.323.7311.514b

Prasad, V., & Ioannidis, J. P. A. (2014). Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices. *Implementation Science, 9,* 1–5. http://dx.doi.org/10.1186/1748-5908-9-1

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31,* 437–448. http://dx.doi.org/10.3102/10769986031004437

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research, 36,* 24–34.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., . . . Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health, 38,* 65–76. http://dx.doi.org/10.1007/s10488-010-0319-7

Rakovshik, S. G., & McManus, F. (2010). Establishing evidence-based training in cognitive behavioral therapy: A review of current empirical findings and theoretical guidance. *Clinical Psychology Review, 30,* 496–516. http://dx.doi.org/10.1016/j.cpr.2010.03.004

R Core Team, & R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richard, L., Gauvin, L., & Raine, K. (2011). Ecological models revisited: Their uses and evolution in health promotion over two decades. *Annual Review of Public Health, 32,* 307–326. http://dx.doi.org/10.1146/annurev-publhealth-031210-101141

Rossen, E., & Oakland, T. (2008). Graduate preparation in research methods: The current status of APA-accredited professional programs in psychology. *Training and Education in Professional Psychology, 2,* 42–49. http://dx.doi.org/10.1037/1931-3918.2.1.42

Rovine, M. J., & Molenaar, P. C. M. (2005). Relating factor models for longitudinal data to quasi-simplex and NARMA models. *Multivariate Behavioral Research, 40,* 83–114. http://dx.doi.org/10.1207/s15327906mbr4001_4

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. http://dx.doi.org/10.1037/1082-989X.7.2.147

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90–100. http://dx.doi.org/10.1037/a0015108

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353. http://dx.doi.org/10.1037/1040-3590.8.4.350

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health, 38,* 32–43. http://dx.doi.org/10.1007/s10488-010-0321-0

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials, 11,* 32. http://dx.doi.org/10.1186/1745-6215-11-32

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316. http://dx.doi.org/10.1037/0033-2909.105.2.309

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18,* 572–582. http://dx.doi.org/10.1037/a0034177

Shelton, R. C., Cooper, B. R., & Stirman, S. W. (2018). The sustainability of evidence-based interventions and practices in public health and health care. *Annual Review of Public Health, 39,* 55–76. http://dx.doi.org/10.1146/annurev-publhealth-040617-014731

Sher, K. J., Jackson, K. M., & Steinley, D. (2011). Alcohol use trajectories and the ubiquitous cat's cradle: Cause for concern? *Journal of Abnormal Psychology, 120,* 322–335. http://dx.doi.org/10.1037/a0021813

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1,* 580. http://dx.doi.org/10.1177/2515245918810511

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

Southam-Gerow, M. A., Daleiden, E. L., Chorpita, B. F., Bae, C., Mitchell, C., Faye, M., & Alba, M. (2014). MAPping Los Angeles County: Taking an evidence-informed model of mental health care to scale. *Journal of Clinical Child and Adolescent Psychology, 43,* 190–200. http://dx.doi.org/10.1080/15374416.2013.833098

Sternberg, R. J. (1992). *Psychological Bulletin's* top 10 "hit parade. " *Psychological Bulletin, 112,* 387–388. http://dx.doi.org/10.1037/0033-2909.112.3.387

Szucs, D. (2016). A tutorial on hunting statistical significance by chasing N. *Frontiers in Psychology, 7,* 1444. http://dx.doi.org/10.3389/fpsyg.2016.01444

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*. Advance online publication. http://dx.doi.org/10.1146/annurev-clinpsy-050718-095710

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2016). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742–756.

Tang, T. Z., DeRubeis, R. J., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology, 73,* 168–172. http://dx.doi.org/10.1037/0022-006X.73.1.168

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21stcentury: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*. Advance online publication. http://dx.doi.org/10.1080/10691898.2012.11889641

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology, 112,* 578–598. http://dx.doi.org/10.1037/0021-843X.112.4.578

Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F., & Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews, 1,* 60. http://dx.doi.org/10.1186/2046-4053-1-60

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13,* 411–417. http://dx.doi.org/10.1177/1745691617751884

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57. http://dx.doi.org/10.3758/s13423-017-1343-3

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005, June). Twelve-month use of mental health services in the United States. *Archives of General Psychiatry, 62,* 629–640. http://dx.doi.org/10.1001/archpsyc.62.6.629

Weissman, M. M., Verdeli, H., Gameroff, M. J., Bledsoe, S. E., Betts, K., Mufson, L., . . . Wickramaratne, P. (2006). National survey of psychotherapy training in psychiatry, psychology, and social work. *Archives of General Psychiatry, 63,* 925–934. http://dx.doi.org/10.1001/archpsyc.63.8.925

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 688–701. http://dx.doi.org/10.1037/0022-006X.63.5.688

Weisz, J. R., & Jensen, A. L. (2001). Child and adolescent psychotherapy in research and practice contexts: Review of the evidence and suggestions for improving the field. *European Child & Adolescent Psychiatry, 10,* I12–I18. http://dx.doi.org/10.1007/s007870170003

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE, 11,* e0152719. http://dx.doi.org/10.1371/journal.pone.0152719

Wild, C. J., & Phannkuch, M. (1999). Statistical thinking in empirical inquiry. *International Statistical Review, 67,* 223–248. http://dx.doi.org/10.1111/j.1751-5823.1999.tb00442.x

Williams, M., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation, 18,* 1–14.

Wilson Von Voorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3,* 43–50. http://dx.doi.org/10.20982/tqmp.03.2.p043

Witkiewitz, K., King, K., McMahon, R. J., Wu, J., Luk, J., Bierman, K. L., . . . Pinderhughes, E. E. (2013). Evidence for a multi-dimensional latent structural model of externalizing disorders. *Journal of Abnormal Child Psychology, 41,* 223–237. http://dx.doi.org/10.1007/s10802-012-9674-z

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*. Advance online publication. http://dx.doi.org/10.1017/s0140525x17001972

Received August 14, 2018
Revision received December 21, 2018
Accepted January 3, 2019 ■