Evaluation & the Health Professions
1-28

DOI: 10.1177/0163278718772879
journals.sagepub.com/home/ehp

Challenges Facing Evidence-Based Prevention: Incorporating an Abductive Theory of Method

W. Alex Mason[1], Jasney Cogua-Lopez[1], Charles B. Fleming[2], and Lawrence M. Scheier[3]

Abstract

Current systems used to determine whether prevention programs are "evidence-based" rely on the logic of deductive reasoning. This reliance has fostered implementation of strategies with explicitly stated evaluation criteria used to gauge program validity and suitability for dissemination. Frequently, investigators resort to the randomized controlled trial (RCT) combined with null hypothesis significance testing (NHST) as a means to rule out competing hypotheses and determine whether an intervention works. The RCT design has achieved success across numerous disciplines but is not without limitations. We outline several issues that question

[1]National Research Institute for Child and Family Studies, Boys Town, NE, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[3]LARS Research Institute, Scottsdale, AZ, USA

Corresponding Author:
W. Alex Mason, National Research Institute for Child and Family Studies, 14100 Crawford Street, Boys Town, NE 68010, USA.
Email: walter.mason@boystown.org

allegiance to the RCT, NHST, and the hypothetico-deductive method of scientific inquiry. We also discuss three challenges to the status of program evaluation including reproducibility, generalizability, and credibility of findings. As an alternative, we posit that extending current program evaluation criteria with principles drawn from an abductive theory of method (ATOM) can strengthen our ability to address these challenges and advance studies of drug prevention. Abductive reasoning involves working from observed phenomena to the generation of alternative explanations for the phenomena and comparing the alternatives to select the best possible explanation. We conclude that an ATOM can help increase the influence and impact of evidence-based prevention for population benefit.

Keywords
evidence-based prevention, methodology, randomized controlled trial, abductive inference, reproducibility, generalizability

Evidence-based prevention encourages the use of proven interventions for preventing behavioral, emotional, and physical health problems (e.g., Kellam & Langevin, 2003; Nation et al., 2003). A critical element in establishing whether a program is evidence-based is the type of scientific inference (reasoning) used to make this case. Most programs are vetted using published standards of evidence that emphasize significance testing of hypothesized program effects in evaluations (e.g., Cohen, 1994; Krantz, 1999). This approach guides scientists in the review of experimental evidence supporting program impact. Prevention science has adopted a broad set of principles for conducting program evaluation elaborated as the "standards of evidence" (Flay et al., 2005). The standards were designed to ensure that investigators rely on proper research methodology, attending to internal and external validity of program evaluations. It is only when programs have met these standards that they can "go to scale" (Gottfredson et al., 2015).

In this article, we suggest the potential impact of evidence-based prevention on reducing public health problems, particularly adolescent drug use, is constrained by a reliance on the hypothetico-deductive (HD) method of scientific inquiry. This observation is not unique to prevention science; indeed, similar critiques have been made in regard to evidence-based medicine and evidence-based economics (Cartwright, 2007; Ioannidis, Stanley, and Doucouliagos, 2017, Osimani, 2013; Stegenga, 2011; Upshar & Fuller, 2016; Worrall, 2007). The HD method is a mainstay of current scientific

reasoning with historical roots in the philosophy of science.[1] Rarely have the quantitative social and behavioral science fields raised challenges to the nature of HD reasoning or considered alternatives. As a result, its utilization as a basis for scientific reasoning has become entrenched in those fields, including prevention science.

To address this and related concerns, we have organized this article into three sections as follows. We begin by reviewing the HD method and outlining three challenges that derive from the application of its principles. We then propose incorporation of an abductive theory of method (ATOM; Haig, 2005) and discuss how its principles can address these challenges, offering recommendations to researchers that can advance evidence-based prevention. Briefly, abductive reasoning involves working from observed phenomena to the generation of alternative explanations for the phenomena and comparing the alternatives to select the best possible explanation, not unlike the process of making a differential diagnosis in medicine (Stanley & Campos, 2013). We close by discussing how broadening the epistemological framework of evidence-based prevention to include an ATOM will foster a stronger knowledge base from which to determine whether programs supporting positive health and development are truly "evidence-based."

Our goal here is to take a first step in outlining a constructive framework for improving the science of behavior change, with a particular focus on the field of adolescent drug use prevention. We do not delve into the nuanced philosophy of science that underlies current research practices; indeed, doing so would be beyond our expertise and the scope of this article. Instead, we wish to initiate what we hope will be an ongoing discussion that seeks to apply general principles drawn from an ATOM to the improvement of evidence-based prevention standards and practices. We believe that integrating these ideas, which are new to prevention science, can help move the field forward to maximize its impact on improving public health.

Principles of the Hypothetico-Deductive Method

The HD method begins with an expectation, derived from theory, about what might happen under certain conditions; it then entails the formulation of hypotheses and moves through a series of tests intended to refute or validate the theory from which the hypotheses are generated (e.g., Andersen & Hepburn, 2016). Knowledge of the developmental pathways to drug use relies heavily on the HD method (e.g., Fergusson, Boden, & Horwood, 2009; Lee, Chassin, & MacKinnon, 2010; Mason, Russo, Chmelka, Herrenkohl, & Herrenkohl, 2017) and guides formulation of prevention and

intervention research (Scheier, 2010). From a prevention standpoint, application of the HD method involves developing intervention theories with explicitly stated program logic and using this framework to guide the collection and analysis of experimental data. This includes specifying the statistical tests to examine hypothesized intervention effects on targeted proximal and distal outcomes.[2] Commonly, this method entails using null hypothesis significance testing (NHST) to guide inferences regarding program effects. In its simplest form, the null hypothesis would state that the program (e.g., instructional strategies) has no effect on a target mediator (e.g., drug refusal skills) or the program's ultimate outcome (i.e., drug use). The objective, then, is to refute or falsify the claim of "null" effects. If the program does work in the manner hypothesized (e.g., exposure to instructional strategies improves refusal skills and reduces drug use), it will warrant continued testing until no further refutable evidence can be established.

Although ubiquitous across the social and behavioral sciences, the HD method has limitations (Pepper, 1942). A primary limitation concerns the method's reductionism, which entails breaking down compound phenomena into parts and allowing for testing of simple hypotheses. A mechanistic, deterministic, and linear view of natural phenomena often accompanies reductionism (e.g., Bechtel, 2009). By contrast, natural phenomena, including people, do not function in the bits and pieces as is conceptualized using a reductionist approach; rather, humans are influenced by a dynamic confluence of contextual effects that sometimes evade detection, let alone conform to experimental conditions. The HD method also encourages evaluating theories in isolation against data (Capaldi & Proctor, 2008), which can lead to the premature rejection of theories that, although useful, have deficiencies.

There have been attempts to address these concerns while staying within the boundaries of a hypothesis-driven approach to science. Still, we suggest that the method on its own is inadequate for providing a full understanding of evidence-based prevention. Dating back to Coie et al. (1993), prevention scientists have recognized that interventions should be grounded in dynamic, developmental theories; adopt a systems perspective that situates individuals within their ecological contexts (Brook, Brook, Gordon, Whiteman, & Cohen, 1990; Szapocznik & Coatsworth, 1999); and target multiple risk and protective processes related to the multidimensional outcomes of interest (Kellam & Langevin, 2003). As we describe in more detail below, a strict HD method provides a valuable but narrow approach to evaluating the evidence base, and alternative ways of examining interventions are worthy of consideration. Combined with questionable practices that, for better or worse, have become linked with the HD method,

such as certain conventions surrounding significance testing (Cohen, 1994; Jacobson & Truax, 1991; Kline, 2004), practical challenges with epistemic roots have emerged.

Reproducibility Challenge

In this section, we begin by addressing the broad *reproducibility challenge* that social and behavioral sciences, in general, have been facing in recent years and illustrate how the issues apply to evidence-based prevention. Next, we acknowledge that one response to this challenge has been to improve the rigor (i.e., study design quality) and transparency (i.e., full and accurate reporting) of our field's methods (Collins & Tabak, 2014). This response is necessary, albeit we believe not sufficient. We illustrate this point by highlighting flaws in either the logic or the practice of significance testing, a practice that forms the bedrock of most rigorous standards of evidence.

Ioannidis (2005) argues that scientific inquiry is plagued by a high prevalence of false positive findings. This results in the scenario in which theories (or interventions) that appear to survive initial tests and that are published in peer-reviewed journals, lack replication in studies with a similar design and sample population and, more broadly, lack reproducibility in studies across trials with varying designs and that use samples drawn from different populations. Ioannidis outlines several conditions, whereby false findings may outnumber true findings. These include having (1) a small number of studies within a field, (2) small true effect sizes, (3) the potential for multiple post hoc comparisons (i.e., multiple hypothesis testing), (4) flexibility in design and analysis decisions, (5) financial and other conflicts of interest, and (6) numerous research teams "in chase of statistical significance" on common research questions. All of these conditions are notably present in prevention science.[3] For example, effect sizes of preventive interventions often are small relative to drug treatment interventions (e.g., Sandler et al., 2014). Also, conflicts of interest in the form of financial rewards and professional status are common considerations, since developers of interventions often oversee their own program evaluations (Gorman, 2005, 2014). Program evaluators typically test the effects of preventive interventions on a wide range of potential outcomes and have flexibility in design, measurement, and analysis decisions. Finally, multiple researchers are often evaluating similar interventions that share theoretical foundations, putative mediators, and outcomes, a practice commonly encountered with drug use prevention programs. As a result of these

conditions, findings in support of evidence-based interventions (EBIs) may fail to be reproduced across diverse trials (Gandhi, Murphy-Graham, Petrosino, Chrismer, & Weiss, 2007).

The standards of evidence-based prevention (Gottfredson et al., 2015) and various program registry websites[4] emphasize rigorous and transparent evaluation methods to address reproducibility concerns (Shadish, Cook, & Campbell, 2002). The assumption underlying these standards is that improvements in current research practices will be used in practice and generate reliable and valid information. Yet there are reasons to question this assumption. A primary reason concerns current practices and misunderstandings regarding NHST. Levine, Weber, Hullett, Park, and Lindsey (2008) organized the most common criticisms of significance testing into four themes addressing sample size, the specification of the null hypothesis, statistical power, and common misunderstandings. We briefly review these.

Sample size influences the determination of statistical significance, with larger sample studies more likely to report statistically significant results, other things being equal. Design and analysis considerations related to sampling and measurement can affect the p values of statistical tests, as can the magnitude of the effect itself. This property of significance testing being a multidetermined decision rule for judging the acceptability of a result has generated valuable practice recommendations (e.g., improve the reliability and validity of measures; Hansen & Collins, 1994); however, significance testing still presents challenges. For example, increasing the sample size of a study increases the likelihood that trivial effects will be statistically significant. This has led to recommendations to report meaningful effect sizes and make a determination of the practical significance of results (Jacobson & Truax, 1991; Vacha-Haase & Thompson, 2004). However, these recommendations have often left applied researchers befuddled because many fields lack a basis for establishing a reasonable effect size or determining whether it is practically important (Kraemer & Kupfer, 2005). As a result, significance testing prevails, including the perpetuation of practices, such as p-hacking (Head, Holman, Lanfear, Kahn, & Jennions, 2015) to find more statistically significance results (e.g., selectively reporting on multiple nonindependent tests), thereby increasing the likelihood that program evaluation results will not be reproducible.

Furthermore, in many disciplines, the null hypothesis, against which an alternative hypothesis is tested, is routinely false (Meehl, 1967). For instance, in prevention science, it is common to frame the null hypothesis as the difference between two experimental conditions corresponding to a specific value, typically zero. This is contrasted with an alternative

hypothesis that the difference between experimental conditions is nonzero and, as a modest extension of the prognostic value, either positive or negative in direction (Cohen [2017] termed this "directional support"). As a result, NHST as commonly practiced does not provide much information, particularly as sample sizes increase, thereby tending to suggest, when there is a statistically significant result, that obviously false null hypotheses are just that, false. Although federal funders of research require some indication in grant applications of adequate statistical power for the primary hypotheses of proposed studies, many behavioral science experiments have small samples due to budgetary and other practical considerations, causing them to suffer from low statistical power (Maxwell, 2004). Low statistical power increases the likelihood of a Type II error or failing to reject the null hypothesis when an effect is actually present (Cohen, 1988). Across a field of study, low power also increases the proportion of statistically significant findings that are Type I errors (Ioannidis, 2005) and, even when statistically significant findings reflect true effects, results in inflated estimates of effect size, what Gelman and Carlin (2014) refer to as a Type M error.

Levine et al. (2008) describe a disconcerting scenario in significance testing where small sample studies increase the likelihood of overlooking important findings, but increasing the sample size increases the likelihood of rejecting obviously false null hypotheses and revealing trivial effects, with little basis for determining the practical significance of results. Combining the above-mentioned concerns with the numerous misunderstandings and false beliefs that researchers have been shown to hold about significance testing (Kline, 2004), Levine et al. (2008) state, "A decision rule in which errors are probable and difficult to avoid cannot be a useful tool" (p. 178). Yet, this decision rule serves as the primary basis on which practitioners and policy makers render judgments about the efficacy or effectiveness of preventive interventions, using this information to determine which interventions are "evidence-based" and ready for large-scale dissemination.

Concerns about the reproducibility of findings, including critiques of significance testing, cut to the heart of evidence-based prevention. This strategy ultimately compromises the impact of evidence-based prevention *as traditionally practiced*, particularly if research findings generated by the field cannot be trusted as reliable. Solutions to some of these concerns can be found in improving current research practices (i.e., adhering to standards of rigor and transparency; Gottfredson et al., 2015); however, many concerns cannot be remedied by merely doing better randomized prevention

trials. As discussed below, even when such trials have been well conducted, their generalizability to real-world practice settings often is limited.

Generalizability Challenge

One important challenge for the field of prevention concerns the limited degree to which EBIs have been translated into routine practice (Mason, Fleming, Thompson, Haggerty, & Snyder, 2014). It typically takes about 17 years for an EBI to move from the initial development, pilot testing, and field trial stages into applied service settings (Meffert, Neylan, Chambers, & Verdeli, 2016). This lengthy time frame reflects the traditional movement from efficacy to effectiveness trials. A traditional research cycle requires testing for favorable program effects under ideal, tightly controlled conditions (efficacy trial) and then expanding the application to different settings, samples, or conditions (effectiveness trials) conducted in less tightly controlled real-world settings (Glasgow, Lichtenstein, & Marcus, 2003). The next stage involves dissemination trials (e.g., Frantz, Stemmler, Halweg, Plück, & Heinrichs, 2015), which evaluate the best ways to scale up supported interventions, with an emphasis on randomized controlled trial (RCTs) conducted at each stage. The 17-year benchmark hides the fact that, with some notable exceptions (e.g., multisystemic therapy, Triple P), many EBIs are not adopted for routine use in community settings on a large scale (e.g., school-based drug use prevention programs; Ringwalt et al., 2011).

Prevention science has taken great strides to map out theories and strategies for scaling up EBIs to bridge the science-to-practice gap (Spoth et al., 2013). These Type 2 translation efforts represent significant steps forward but may not realize their full potential as long as the field remains committed to the efficacy–effectiveness–dissemination cycle and an overreliance on RCTs (Glasgow & Emmons, 2007). Efficacy trials are often conducted with good intent but lack a specific large-scale delivery vehicle for the program. This leaves many interventions ill-prepared or impractical for implementation in applied service settings (Rotherham-Borus & Duan, 2003), a conclusion reinforced by Glasgow et al. (2003), who suggest that dissemination and implementation are often afterthoughts to intervention development and efficacy testing.

The tightly controlled nature of the RCT and its emphasis on reductionism and standardization suggest this design tool may be poorly equipped for providing verisimilitude with real-world field settings (e.g., Brown & Liao, 1999; Cartwright, 2007). The RCT, on its own, cannot show that supported preventive interventions can be disseminated and implemented successfully

in practice. This is particularly true for complex interventions needed to prevent significant and widespread behavioral, emotional, and physical health problems at a population level (e.g., Frantz et al., 2015). Despite clear overtures that reinforce the importance of RCTs (Gottfredson et al., 2015), a growing body of scholarship questions the status of the RCT as the gold standard tool for evidence-based prevention and treatment (Tucker & Roth, 2006; Williams, 2010). One concern relates to randomization itself. It can be impossible or undesirable, and even unethical at times, to randomly assign individuals, families, schools, or communities to experimental conditions. Those assigned to a no-intervention control condition do not receive potentially beneficial programming, possibly making it unethical to hold back a program in light of current public health mandates (Henry, Tolan, Gorman-Smith, & Schoeny, 2017). For prevention, this may be a more relevant concern in the evaluation of selective and indicated interventions targeting high-risk individuals or groups. Added to this is the potential in real-world settings for contamination and compensatory rivalry, both threats to internal validity. There are available alternative designs that attempt to ameliorate these threats. These include the dynamic wait-listed design that phases-in delivery of the intervention for all participants (Brown, Wyman, Guo, & Peña, 2006). However, the appearance of denying timely services to those in need within the context of a trial can generate resistance from both study participants (e.g., disengagement) and intervention providers (e.g., sabotage, if the study cannot be fully blinded).

Field trials also make it difficult to standardize the nature of the control group. In numerous instances, particularly school-based drug prevention trials, control participants end up receiving some form of relevant prevention or treatment services, sometimes even EBIs. This is because of federal legislative mandates to use EBIs and achieve the benchmarks detailed in the No Child Left Behind Act and its successor, the Every Student Succeeds Act. When this happens, tendering some form of treatment to the control group can downwardly bias intervention effects and make it more challenging to pool findings across studies in meta-analyses (Lipsey & Wilson, 2001). Furthermore, a randomized design can create problems of self-selection and motivation in which only certain individuals or groups may be willing to participate in a trial (Tucker & Roth, 2006). High-risk families, for instance, often face multiple stressors and are reluctant to participate in family interventions that are often delivered outside the home. Their exclusion biases the sample that is participating in an intervention trial because they do not reflect qualities of the target population (Hill, Goates, & Roseman, 2010).

As a design strategy, RCTs can strip away meaningful contextual information in tests of interventions, particularly those that are systems-oriented, multifaceted, adaptive, and targeted toward heterogeneous populations. In these types of scenarios, it may be prudent to leave intact certain groups and capture variation that reflects temporal or contextual effects. There are a number of examples where researchers can capitalize on quasi-experimental designs including environmental strategies and policy studies as well as community-based drug use interventions. Environmental strategies that are implemented with whole communities can utilize a comparator-based time series design as a practical way to isolate the effects of a change in policy (Biglan, Ary, & Wagenaar, 2000). These types of interventions are common in community-based prevention trials, which frequently demand complex, multitiered, and adaptive or tailored interventions to address complex problems.

In summary, a randomized controlled efficacy trial encourages narrowly defined interventions with strict adherence to protocols in order to demonstrate favorable program effects under ideal conditions. By contrast, the applied service setting encourages multicomponent and adaptive interventions that can generate desired changes within a heterogeneous population of individuals, often involving intact families, schools, peer networks, and communities. An intervention designed to clear the efficacy hurdle may have little relevance for the real world.

Credibility Challenge

Evidence-based prevention suffers a credibility challenge by being part of the larger scientific enterprise that has been criticized for lacking reproducibility (Ioannidis, 2005). The notion of an objective science has been challenged by concerns about research practices (e.g., p-hacking) and reward structures (e.g., publish or perish) that contribute to bias and, at times, fraud (e.g., Fanelli, 2010). Prevention science has also been the subject of targeted challenges that serve to cast doubt on the credibility of the field (Gorman, 2003). These critiques have been applied to school-based (Gorman, 2002, 2005) and family-based drug prevention (Gorman, Conde, & Huber, 2007). Gandhi et al. (2007) suggest there is weak evidence for many school-based drug use prevention programs, including ones that appear on evidence-based program registries. Recent reports documenting a failure to replicate positive results for some programs (e.g., Gorman, 2017; Robling et al., 2016) compound this problem and call attention to the

potential biases that may result when program developers also serve as evaluators (Eisner, 2009).

Of course, critiques of prevention science as well as replication studies vary in quality and require scrutiny, but we should not ignore the concerns raised. Collectively, these concerns undermine confidence in the proposition that a sufficient knowledge base exists to identify programs that, if implemented with fidelity, will produce the desired benefits. The practical consequence of a credibility challenge is that prevention science may fall short of its goal of influencing policy and practice to make significant gains eradicating behavioral, psychological, physical, and social problems on a large scale.

Principles of the ATOM

We now introduce abductive reasoning as an alternative scientific logic on which to base program evaluation findings (Douven, 2017). Charles Peirce coined the term abduction (Burks, 1946), and its concepts have been elaborated over the years (Thagard, 1978), recently by Haig (2005, 2009) in a formal ATOM. Peirce conceptualized abduction as one of the three types of reasoning, in addition to induction and deduction (Burks, 1946). Of the three, only abductive reasoning is concerned with the logic of discovery in the sense of generating new hypotheses about observed phenomena. Capaldi and Proctor (2008) termed this "novel hypothesis abduction." Abductive reasoning more commonly refers to the process of selecting among the best available hypotheses, what Capaldi and Proctor (2008) have termed "competing theories abduction." In essence, "abductive inference involves reasoning from a claim about a presumed effect (the empirical phenomenon) to its explanation in terms of underlying causal mechanisms" (the explanatory theory; Vertue & Haig, 2008, p. 1051). The ATOM begins with a process of *phenomena detection*, where phenomena are conceptualized as stable and recurrent "empirical regularities" (e.g., events or processes, such as the presumed outcomes of a preventive intervention) that require explanation. Haig (2005) contrasts phenomena with *data*, the latter of which are recordings or reports that serve as evidence of phenomena but are less stable and more context-dependent (e.g., mean scores at pretest and posttest in a prevention trial). Data are valuable and contain an intrinsic ability to provide reliable evidence for the presence of phenomena. As a result, ATOM emphasizes *strategies* for rigorous data collection (e.g., limiting measurement error, experimentally or statistically controlling for confounding, conducting replications). ATOM also emphasizes a multistage

framework of analysis *methods* involving (a) initial data analysis to ensure quality, (b) exploratory data analysis (e.g., confirmatory factor analysis, stepwise regression, and latent class analysis) to suggest patterns, (c) close replication to confirm patterns, and (d) constructive replication to demonstrate generalization of results (Haig, 2005, p. 376).

Following phenomena detection, ATOM turns to the process of *theory construction*, the goal of which is to develop an explanation of the causal mechanisms that might have generated observed data patterns (regularities) presumed to reflect empirical phenomena. Haig (2005) notes that theory construction in ATOM involves three phases including (1) theory generation, (2) theory development, and (3) theory appraisal. Collectively these phases describe the process of moving from tentative theories with initial plausibility to an "inference to the best explanation" (IBE) or a determination of the particular theory that offers the most coherent explanation of the evidence compared to its rival theories (Capaldi & Proctor, 2008; Thagard, 1978; i.e., "competing theories abduction").

Haig (2009) proposed a hybrid HD method extended by the principles of abductive reasoning to achieve the goals of both predictive success and explanatory worth, and others have made similar efforts to refine the basic tenets of the HD method (e.g., the pragmatist paradigm; Reiss, 2015). Haig offers, as an example, the practice of model comparisons in structural equation modeling in which model selection is guided by both global fit statistics (as indicators of the goodness-of-fit between the data and the model) and parsimony (as one indicator of explanatory worth). Below, we indicate how this hybrid approach can help the field move toward resolving the challenges facing evidence-based prevention. In each case, we begin with recommendations that derive more directly from traditional research practices and then offer recommendations grounded in abductive principles. To guide this discussion, Table 1 provides an overview of each challenge and a summary of recommendations.

Recommendations for Addressing the Reproducibility Challenge

Because ATOM emphasizes the importance of phenomena detection as well as rigorous data collection and analysis, Haig (2005) highlights the need for strategies, including the RCT, that control for confounds. Thus, *Recommendation 1* is to support and expand efforts to increase the quality of randomized trials (see Table 1). This should involve mindfulness of existing standards of prevention science (Gottfredson et al., 2015), as well

Table 1. Challenges and Recommendations Guided by an Extension of the Hypothetico-Deductive Method With Abductive Reasoning.

Challenges	Recommendations	Examples
Reproducibility The extent to which prevention science findings are sound and can be *reproduced*.	1. Support and expand efforts to increase the rigor and transparency of randomized trials	– Standards of evidence for prevention – Consolidated standards of reporting trials – Good clinical practice training – Trial registration
	2. Increase the use of well-designed quasi-experiments when true randomization is undesirable or infeasible	– Regression discontinuity – Matched comparison groups – Adjustment for preintervention differences
	3. Implement methods that demonstrate the robustness of evidence-based intervention (EBI) effects across settings, samples, and other conditions with a de-emphasis on significance testing	– Meta-analysis – Big data techniques – Grounded theory

(continued)

Table 1. (continued)

Challenges	Recommendations	Examples
Generalizability The extent to which prevention science findings are relevant in real-world contexts and can be *generalized*.	4. Adopt methods for improving and assessing the external validity of randomized controlled trials (RCTs)	– RE-AIM – Weighting to match the population
	5. Use alternatives to the traditional evaluation research cycle (efficacy–effectiveness–dissemination)	– Effectiveness–implementation hybrid trials – Test widely used, understudied interventions
	6. Implement a wider array of evaluation methods beyond the RCT	– Multivariate longitudinal designs – Nonlinear analysis methods – Mixed quantitative–qualitative methods – Best evidence systematic review
Credibility The extent to which prevention science findings are influential to stakeholders and have *credibility*.	7. Acknowledge and embrace uncertainty in evaluation research results	– Inference to the best possible explanation
	8. Take action in applying evaluation research results before they are conclusive	– Begin implementing preventive interventions in applied settings as soon as it can be judged that such interventions plausibly work as intended and do no harm
	9. Display humility in the application of evidence-based prevention	– Report and learn from failures – Be transparent about strengths and limitations when drawing implications

as those that cut across disciplines, such as the consolidated standards of reporting trials criteria (Schulz, Altman, & Moher, 2010). Rigorous data collection strategies required for phenomena detection also suggest utilization of quasi-experimental studies as a potential evaluation tool (Cook, Shadish, & Wong, 2008). Partly through both design (e.g., regression discontinuity designs) and statistical modeling (e.g., propensity score matching), methodologists are developing an understanding of the conditions under which well-conducted quasi-experimental studies can provide results comparable to RCTs (Cook et al., 2008). Such conditions include, for example, closely matched or statistically equated nonrandomized groups. Thus, *Recommendation 2* is to increase the use of well-designed quasi-experiments when true randomization is undesirable or infeasible (see Table 1). We agree with West and his colleagues (2008), who state, "When RCTs cannot be implemented in settings or with participants of interest, it is far better to use a strong alternative design than to change the treatment . . . or study population so that an RCT may be implemented" (p. 1363). In this respect, standards of evidence require revision to include criteria that give proper consideration of RCT alternatives.

High-quality RCTs and quasi-experiments with an overreliance on significance testing are unlikely to help evidence-based prevention realize its full potential. Instead, the abductive method encourages both close and conceptual replications of observed phenomena over the demonstration of statistical significance in single studies. Haig (2005) notes that "researchers should seek the generalizability of relationships rather than their statistical significance" (p. 376) and goes on to call for both experimental and observational studies conducted across multiple data sets that vary on features such as setting and sampling. This position is at the heart of a recent call for greater replication in prevention science (Valentine et al., 2011) and a need to move beyond the benchmark of significance testing (Anderson & Maxell, 2016). Therefore, *Recommendation 3* is to implement methods that demonstrate the robustness of EBI effects across settings, samples, and other conditions while at the same time reducing the weight given to statistically significant results in a given trial (see Table 1). Certain traditional methods, such as meta-analysis (Wilson & Tanner-Smith, 2014), that operate within the boundaries of the HD method already take steps in this direction. Meta-analysis pools observed intervention effect sizes across multiple studies using a common metric, providing greater precision than those obtained from individual trials (Lipsey & Wilson, 2001).

Still, the field as a whole could benefit from developments in other areas compatible with abduction, including big data analytics. Big data

techniques prototypically involve the search for reliable patterns or signals within a large data array (Harlow & Oswald, 2016). In many situations, particularly in circumstances in which randomization is difficult or impossible (e.g., evaluation of prevention-oriented local, state, and national policies), big data techniques can aid the search for reliable phenomena in the form of desired policy effects. Similarly, Derzon (2014) describes a synthetic differences-in-differences approach for collecting intervention data on a large scale and synthesizing results for comparison using alternative practices. Qualitative data also play an equally important role in an abductive approach to evidence-based prevention. Grounded theory (Glaser & Strauss, 1967), for instance, draws on qualitative data to aid in theory development and the identification of recurrent themes and data patterns in data usually collected from focus groups. This approach can be used, for example, to understand consumer and provider preferences that play a role in the adoption of EBIs (Spoth et al., 2013).

Recommendations for Addressing the Generalizability Challenge

Abductive principles also suggest recommendations for addressing the generalizability challenge. Recall that the goal of abductive reasoning is to infer patterns, ideally detected from analyses of multiple datasets that vary on relevant conditions, plausible explanations for the effects that underlie phenomena of interest (e.g., robust intervention-induced outcome effects leading to improvements in population health). Here, the combined emphasis on broad sets of data conditions and rigorous data analysis, which includes experimental control for confounding when feasible, suggests *Recommendation 4*, which is to adopt methods for improving and assessing the external validity of randomized prevention trials (see Table 1). The reach, efficacy, adoption, implementation, mainintenance (RE-AIM) framework (Glasgow et al., 2003) highlights the need to measure and examine factors related to reach into the target population, efficacy or effectiveness of interventions, adoption by providers, implementation (e.g., fidelity vs. adaptation), and maintenance or sustainability of interventions and their effects. Each component has received extensive documentation for its relevance to the critical task of disseminating evidence-based prevention: reach (Spoth et al., 2013), efficacy or effectiveness (Glasgow et al., 2003), adoption (Rohrbach, Ringwalt, Ennett, & Vincus, 2005), implementation (Durlak & DuPre, 2008), and sustainability (Tibbits, Bumbarger, Kyler, & Perkins, 2010). The RE-AIM framework sets forth a

priority to understand better the contexts and conditions that support EBI scale-up and to do so from the outset of the intervention development and evaluation process. Methods for assessing the generalizability of RCT results also are becoming more widely available; for example, Stuart, Bradshaw, and Leaf (2015) describe a method for matching a trial sample with the population on relevant characteristics using propensity score techniques. EBI registries would benefit greatly if, guided by RE-AIM, they could integrate into their review criteria consideration of the external validity of RCT results.

The traditional evaluation research cycle (efficacy–effectiveness–dissemination) emphasizes the controlled testing of interventions in a linear fashion, giving attention to context and other "confounding" or "nuisance" factors only in later stages of the cycle. Instead, program evaluation might begin using an abductive strategy, beginning with applying phenomena detection (e.g., intervention outcome effects gleaned from reliable patterns found in multiple data sources under a variety of conditions), followed by theory construction (e.g., examining the plausibility of the intervention's logic model relative to alternative explanations for the pattern of findings). This suggests *Recommendation 5*, which is to use alternatives to the traditional program evaluation research cycle (see Table 1). Such alternatives include, for example, hybrid effectiveness–implementation trials that explicitly integrate design elements that allow an examination of intervention effects as well as processes related to effective dissemination and implementation (Curran, Bauer, Mittman, Pyne, & Stetler, 2012). As another example, Mason et al. (2014) proposed a framework for identifying promising programs successfully implemented in practice settings but understudied, with the goal of conducting evaluations to ensure the programs work as intended to promote expanded dissemination.

Abductive theory of method further suggests implementing a wider array of evaluation methods beyond the RCT (*Recommendation 6*; see Table 1). Already, we have discussed the potential value of well-conducted quasi-experimental studies. Additional designs include the multivariate longitudinal design (Tucker & Roth, 2006), which involves repeated assessments on a variety of factors and processes. Also promising in this area are developments in nonlinear analysis methods, such as network modeling and agent-based modeling (Gilbert, 2008; Heath, 2000). In contrast to the traditional linear analysis methods, these nonlinear analysis techniques more optimally model complex and dynamical processes, such as is commonly encountered in multilevel preventive interventions and the ecological systems in which they operate (e.g., communities). For example, social

network analysis can be used to test the extent to which school- or community-wide drug use preventive interventions alter friendship networks to reduce antisocial peer influences (Osgood et al., 2013). Mixed methods designs that combine quantitative and qualitative data can provide rich information about intervention contexts and processes (Zhang & Watanabe-Galloway, 2014). *Recommendation 6* applies also to systematic reviews that go beyond results from RCTs to synthesize the best available evidence in an area of inquiry to inform public health (Ogilvie, Egan, Hamilton, & Petticrew, 2005). To the extent that research designs are non-experimental, selection effects must be considered (Shadish et al., 2002), but the trade-off involves explicit consideration and modeling of individual and contextual factors that play a role in the uptake of interventions and in the generation and maintenance of their effects.

Recommendations for Addressing the Credibility Challenge

The principles of abductive reasoning suggest a final set of recommendations for addressing the credibility challenge in evidence-based prevention. Of particular relevance is the concept of IBE, which involves an evaluation of the explanatory goodness of rival theories generated through an abductive process. There are no assurances that this process will lead necessarily or assuredly to the "true" explanation of observed phenomena, only to the best available theory relative to others (given the data, under the observed conditions, at the particular moment in time, etc.). Traditional research practices in evidence-based prevention also operate under a "guiding ideal" of truth seeking (Haig, 2008), and those practices, even when conducted with rigor and transparency, likewise generate only approximations of reality. Thus, *Recommendation 7* is to acknowledge and embrace uncertainty in evaluation research results (see Table 1). In the words of Gopal and Schoor (2016), "To get better results, we must be willing to shake the intuition that *certainty* (emphasis in original) should be our highest priority."

Drawing on a larger number of studies conducted under a wider range of conditions and using a variety of designs suggests considerable effort, expense, and time. Rather than improve upon the estimated 17 years it takes to move innovations from bench to bedside (Meffert et al., 2016), closing the science-to-practice gap could take longer. However, without certainty as the ultimate priority, *Recommendation 8* is to take action in applying evaluation research results before they are conclusive, that is, as soon as it can be judged that the interventions plausibly work as intended and do no harm

(see Table 1). Reiss (2015) takes a significant step in this direction by offering guidelines for accepting plausible intervention results (e.g., large observed effect sizes, predictable manner and timing of effects) and ruling out alternative explanations. One can then follow such action with close monitoring in the form of additional evidence gathering (i.e., replication) and refinements to conclusions, as needed.

Abandoning the notion of certainty may seem counterintuitive for addressing credibility concerns but not in light of our final recommendation. *Recommendation 9* is to maintain humility in the application of evidence-based prevention (see Table 1). Lest critics view our sense of certainty in the evidence base for prevention as immoderate, we would do well to avoid overpromising on the power of prevention in our efforts to influence policy (Shonkoff & Fisher, 2013). Lacking certainty, we should be honest with ourselves and others (policy makers, stakeholders, and consumers) by taking care not to overplay the strengths of evaluation research results when drawing implications.

Summary and Implications

In a relatively short period of time, evidence-based prevention has made significant strides; however, it should be clear that there is substantially more work down the road. Under current research practices, challenges have emerged for the field and gains may have reached a plateau. A new theory of evidence that extends the HD method with the principles of abductive reasoning could help address concerns related to reproducibility, generalizability, and credibility, thereby permitting further advancements. We have outlined a first step toward promoting those advancements by offering recommendations to researchers in the field for strengthening and expanding the standards of evidence-based prevention. This step necessarily omitted certain areas of relevant inquiry, in particular, Bayesian methodology (Mayo, 1997; Sprenger, 2013), which is regarded by some as an alternative to abduction. We also acknowledge the need for greater specificity when implementing the recommendations; for example, it is uncertain how to revise evidence-based standards to more fully incorporate program evaluation results from emerging fields, such as big data.

Of course, abductive reasoning is not without its criticisms and limitations (Lipton, 2004), many of which we have alluded to already (e.g., the challenge of developing criteria for selecting among rival theories), and it informs the logic of scientific inquiry only in conjunction with a consideration of the roles of inductive and deductive reasoning, as implied by the

hybrid method we have described. Our goal here has been to provide a fresh perspective on the methods, challenges, and opportunities of evidence-based prevention. It is our hope that this perspective sparks enthusiasm for the difficult work of building on the noteworthy achievements of prevention science to date. Further gains will be contingent on ensuring that findings generated by the field are reliable, valid, and credible. We contend that a sole commitment to the more rigorous and transparent conduct of current research practices, which rely heavily on the RCT, may only partially realize the gains needed to advance the science of prevention. Instead, extending current research practices to include those guided by abductive principles and considering the recommendations provided herein can achieve the necessary steps needed to move forward. Doing so may help the field increase its influence and impact, ultimately for public health benefit.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

W. Alex Mason http://orcid.org/0000-0002-8132-8045
Lawrence M. Scheier http://orcid.org/0000-0003-2254-0123

Notes

1. A review of the philosophy of science underlying modern social and behavioral science practice, including the hypothetico-deductive method, is beyond the scope of this article, but interested readers are referred to Fisher (1935), Hempel and Oppenheim (1945), Neyman (1957), Popper (1959, 1963), and Whewell (1840) as several important works that present a wide range of views on this topic.
2. This can be construed as a manipulation check or testing the active ingredients of a program. The minimum requirements include testing program effects on putative mediators (i.e., target skills). In turn, the path from the mediator to the outcomes (i.e., drug use) is also part of the mediation chain.
3. In several different scholarly venues, Gorman (2002, 2003, 2005) has raised these and other concerns but not without a concerted retort from the architects

of several drug prevention programs (e.g., Botvin & Griffin, 2005; Hawkins & Catalano, 2005).

4. Three examples relevant to the current emphasis in prevention science are (1) National Registry of Evidence-based Prevention Programs (http://www.samhsa.gov/nrepp), (2) the California Evidence-Based Clearinghouse for Child Welfare (http://www.cebc4cw.org/), and (3) Blueprints for Healthy Youth Development (http://www.blueprintsprograms.com).

References

Andersen, H., & Hepburn, B. (2016). Scientific method. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2016/entries/scientific-method/

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*, 1–12.

Bechtel, W. (2009). Looking down, around and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*, 543–564.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science*, *1*, 31–49.

Botvin, G. J., & Griffin, K. W. (2005). Prevention science, drug abuse prevention, and life skills training: Comments on the state of the science. *Journal of Experimental Criminology*, *1*, 63–78.

Brook, J. S., Brook, D. W., Gordon, A. S., Whiteman, M., & Cohen, P. (1990). The psychological etiology of adolescent drug use: A family interactional approach. *Genetic, Social, and General Monographs*, *116*, 111–267. (Whole No.2).

Brown, C. H., & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiology paradigm. *American Journal of Community Psychology*, *27*, 673–708.

Brown, C. H., Wyman, P. A., Guo, J., & Peña, J. (2006). Dynamic wait-listed designs for randomized trials: New designs for prevention of youth suicide. *Clinical Trials*, *3*, 259–271.

Burks, A. W. (1946). Peirce's theory of abduction. *Philosophy of Science*, *13*, 301–306.

Capaldi, E. J., & Proctor, R. W. (2008). Are theories to be evaluated in isolation or relative to alternatives? An abductive view. *The American Journal of Psychology*, *121*, 617–641.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*, 11–20.

Cohen, B. H. (2017). Why the resistance to statistical innovations? A comment on sharpe (2013). *Psychological Methods*, *22*, 204–210.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Elbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., . . . Long, B. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist*, *48*, 1013–1022.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*, 612–613.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

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.

Derzon, J. H. (2014). Differences-in-differences in approaches. In G. Bruinsma, & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 1054–1064). New York, NY: Springer.

Douven, I. (2017). Abduction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/abduction/

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350.

Eisner, M. (2009). No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology*, *5*, 163–183.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empircal support from US States data. *PLoS One*, *5*, e1027.

Fergusson, D. M., Boden, J. M., & Horwood, L. J. (2009). Tests of causal links between alcohol abuse or dependence and major depression. *Archives of General Psychiatry*, *66*, 260–266.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*, 39–82.

Flay, B. R., Biglan, A., Boruch, R. F., Gonzalez Castro, F., Gottfredson, D. C., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Science*, *6*, 151–175.

Frantz, I., Stemmler, M., Hahlweg, K., Plück, J., & Heinrichs, N. (2015). Experience in disseminating evidence-based prevention programs in a real-world setting. *Prevention Science*, *16*, 787–800.

Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The devil is in the details: Examining the evidence for "proven"school-based drug abuse prevention programs. *Evaluation Review*, *31*, 43–74.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651.

Gilbert, G. N. (2008). *Agent-based models*. London, England: Sage.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Glasgow, R. E., & Emmons, K. M. (2007). How can we increase translation of research into practice? Types of evidence needed. *Annual Review of Public Health*, *28*, 413–433.

Glasgow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health*, *93*, 1261–1267.

Gopal, S., & Schoor, L. B. (2016). Getting "Moneyball" right in the social sector. *Stanford Social Innovation Review*. Retrieved from https://ssir.org/articles/entry/getting_moneyball_right_in_the_social_sector

Gorman, D. M. (2002). The "science" of drug and alcohol prevention: The case of the randomized trial of the Life Skills Training (LST) Program. *International Journal of Drug Policy*, *13*, 21–26.

Gorman, D. M. (2003). The best of practices, the worst of practices: The making of science-based primary prevention programs. *Psychiatric Services*, *54*, 1087–1089.

Gorman, D. M. (2005). Drug and violence prevention: Rediscovering the critical rational dimension of evaluation research. *Journal of Experimental Criminology*, *1*, 39–62.

Gorman, D. M. (2014). Is Project Towards No Drug Abuse (Project TND) an evidence-based drug and violence prevention program? A review and reappraisal of the evaluation studies. *Journal of Primary Prevention*, *35*, 217–232.

Gorman, D. M. (2017). The decline effect in evaluations of the impact of the strengthening families program for youth (SFP 10–14) on adolescent substance use. *Children and Youth Services Review*, *81*, 29–39.

Gorman, D. M., Conde, E., & Huber, J. C. (2007). The creation of evidence in "evidence-based" drug prevention: A critique of strengthening families program plus life skills training evaluation. *Drug and Alcohol Review*, *26*, 585–593.

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*, 893–926.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*, 371–388.

Haig, B. D. (2008). On the permissiveness of the abductive theory of method. *Journal of Clinical Psychology*, *64*, 1037–1045.

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *American Journal of Psychology*, *122*, 219–234.

Hansen, W. B., & Collins, L. M. (1994). *Seven ways to increase power without increasing N*. Bethesda, MD: National Institute on Drug Abuse.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*, 447–457.

Hawkins, J. D., & Catalano, R. F. (2005). Doing prevention science: A response to D. M. Gorman and a brief history of the quasi-experimental study nested with the seattle social development project. *Journal of Experimental Criminology*, *1*, 79–86.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*, e1002106.

Heath, R. A. (2000). *Nonlinear dynamics: Techniques and applications in psychology*. Hillsdale, NJ: Erlbaum.

Hempel, C. G., & Oppenheim, P. (1945). A definition of degree of confirmation. *Philosophy of Science*, *12*, 98–115.

Henry, D., Tolan, P., Gorman-Smith, D., & Schoeny, M. (2017). Alternatives to randomized control trial designs for community-based prevention evaluation. *Prevention Science*, *18*, 671–680.

Hill, L. G., Goates, S. G., & Roseman, R. (2010). Detecting selection effects in community implementations of family-based substance abuse prevention programs. *American Journal of Public Health*, *100*, 623–630.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*, F236–F265.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Kellam, S. G., & Langevin, D. J. (2003). A framework for understanding "evidence" in prevention research and programs. *Prevention Science*, *4*, 137–153.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kraemer, H. C., & Kupfer, D. J. (2005). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*, 990–996.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*, 1372–1381.

Lee, M. R., Chassin, L., & MacKinnon, D. (2010). The effect of marriage on young adult heavy drinking and its mediators: Results from two methods of adjusting for selection into marriage. *Psychology of Addictive Behaviors*, *24*, 712–718.

Levine, T. R., Weber, R., Hullett, C. R., Park, H. S., & Lindsey, L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lipton, P. (2004). *Inference to the best explanation*. Milton Park, England: Taylor and Francis Group.

Mason, W. A., Fleming, C. B., Thompson, R. W., Haggerty, K. P., & Snyder, J. J. (2014). A framework for testing and promoting expanded dissemination of promising preventive interventions that are being implemented in community settings. *Prevention Science*, *15*, 674–683.

Mason, W. A., Russo, M. J., Chmelka, M. B., Herrenkohl, R. C., & Herrenkohl, T. I. (2017). Parent and peer pathways linking childhood experiences of abuse with marijuana use in adolescence and adulthood. *Addictive Behaviors*, *66*, 70–75.

Mayo, D. G. (1997). Duhem's problem, the bayesian way, and error statistics, or "What's belief got to do with it?" *Philosophy of Science*, *64*, 222–244.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meffert, S. M., Neylan, T. C., Chambers, D. A., & Verdeli, H. (2016). Novel implementation research designs for scaling up global mental health care: Overcoming translational challenges to address the world's leading cause of disability. *International Journal of Mental Health Systems*, *10*, 1–6.

Nation, M., Crusto, C., Wandersman, A., Kumpfer, K. L., Seybolt, D., Morriseey-Kane, E., & Davino, K. (2003). What works in prevention: Principles of effective prevention programs. *American Psychologist*, *58*, 449–456.

Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Revue d'Institute Internationale de Statistique* [In French], *25*, 7–22.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. (2005). Systematic reviews of health effects of social interventions: 2. Best available evidence: How low should you go? *Journal of Epidemiology and Community Health*, *59*, 886–892.

Osgood, D. W., Feinberg, M. E., Gest, S. D., Moody, J., Ragan, D. T., Spoth, R., . . . Redmond, C. (2013). Effects of PROSPER on the influence potential of prosocial versus antisocial youth in adolescent friendship networks. *Journal of Adolescent Health*, *53*, 174–179.

Osimani, B. (2013). Until RCT proven? On the asymmetry of evidence requirements for risk assessment. *Journal of Evaluation in Clinical Practice*, *19*, 454–462.

Pepper, S. C. (1942). *World hypotheses: A study of evidence*. Berkeley, CA: University of California Press.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, England: Routledge and Keagan Paul.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science*, *82*, 341–362.

Ringwalt, C. L., Vincus, A. A., Handley, S., Ennett, S. T., Bowling, J. M., & Haws, S. (2011). The prevalence of evidence-based drug use prevention curricula in U. S. middle schools in 2008. *Prevention Science*, *12*, 63–69.

Robling, M., Bekkers, M.-J., Bell, K., Butler, C. C., Cannings-John, R., Channon, S., . . . Torgerson, D. (2016). Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (building blocks): A pragmatic randomized controlled trial. *The Lancet*, *387*, 146–155.

Rohrbach, L. A., Ringwalt, C. L., Ennett, S. T., & Vincus, A. A. (2005). Factors associated with adoption of evidence-based substance use prevention curricula in US school districts. *Health Education and Research*, *20*, 514–526.

Rotherham-Borus, M. J., & Duan, N. H. (2003). Next generation of preventive interventions. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 518–526.

Sandler, I., Wolchik, S. A., Cruden, G., Mahrer, N. E., Ahn, S., Brincks, A., & Brown, C. H. (2014). Overview of meta-analyses of the prevention of mental health, substance use, and conduct problems. *Annual Review of Clinical Psychology*, *10*, 243–273.

Scheier, L. M. (Ed.). (2010). *Handbook of adolescent drug use prevention: Research, intervention strategies, and practice*. Washington, DC: American Psychological Association.

Schulz, K. F., Altman, D., & Moher, D. (2010). Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*, 340.

Shadish, W. R., Cook, TD., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*, (2nd ed.). Boston, MA: Houghton Mifflin Company.

Shonkoff, J. P., & Fisher, P. A. (2013). Rethinking evidence-based practice and two-generation programs to create the future of early childhood policy. *Development and Psychopathology*, *25*, 1635–1653.

Sprenger, J. (2013). A synthesis of hempelian and hypothetico-deductive confirmation. *Erkenntnis*, *78*, 727–738.

Spoth, R., Rohrbach, L. A., Greenberg, M., Leaf, P., Brown, C. H., Fagan, A., … Hawkins, J. D. (2013). Addressing core challenges for the next generation of type 2 translational research and systems: The translation science to population impact (TSci Impact) framework. *Prevention Science*, *14*, 319–351.

Stanley, D. E., & Campos, D. G. (2013). The logic of medical diagnosis. *Perspectives in Biology and Medicine*, *56*, 300–315.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*, 497–507.

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*, 475–485.

Szapocznik, J., & Coatsworth, J. D. (1999). An eco developmental framework for organizing the influences of drug abuse: A developmental model of risk and protection. In M. Glantz, & C. R. Hartel (Eds.), *Drug abuse: Origins and interventions* (pp. 331–366). Washington, DC: American Psychological Association.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *Journal of Philosophy*, *75*, 76–92.

Tibbits, M. K., Bumbarger, B. K., Kyler, S. J., & Perkins, D. F. (2010). Sustaining evidence-based interventions under real-world conditions: Results from a large-scale diffusion project. *Prevention Science*, *11*, 252–262.

Tucker, J. A., & Roth, D. L. (2006). Extending the evidence hierarchy to enhance evidence-based practice for substance use disorders. *Addiction*, *101*, 918–932.

Upshur, R, & Fuller, J. (2016). Randomized controlled trials in the West African Ebola virus outbreak. *Clinical Trials*, *13*, 10–12.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*, 473–481.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., … Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

Vertue, F. M., & Haig, B. D. (2008). An abductive perspective on clinical reasoning and case formulation. *Journal of Clinical Psychology*, *64*, 1046–1068.

Whewell, W. (1840). *The Philosophy of the inductive sciences* (1967 ed.). London, England: Frank Cass & Co. Ltd.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., . . . Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, *98*, 1359–1366.

Williams, B. A. (2010). Perils of evidence-based medicine. *Perspectives on Biology and Medicine*, *53*, 106–120.

Wilson, S. J., & Tanner-Smith, E. E. (2014). Meta-analysis in prevention science. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science: Advances in prevention science* (pp. 431–452). New York, NY: Springer.

Worrall, J. (2007). Evidence in medicine and evidence-based medicine. *Philosophy Compass*, *2*, 981–1022.

Zhang, W., & Watanabe-Galloway, S. (2014). Using mixed methods effectively in prevention science: Designs, procedures, and examples. *Prevention Science*, *15*, 654–662.