Intervention Effectiveness Research in Adolescent Health Psychology: Methodological Issues and Strategies

Norman A. Constantine

Interventions to promote adolescent health have been widely implemented with a variety of goals, settings, populations, and approaches. Many of these interventions focus on preventing risky behaviors, promoting healthy behaviors, or more broadly promoting healthy development—all within the province of adolescent health psychology. Research evidence regarding effectiveness has been accumulating for some intervention approaches, yet the validity and integrity of this evidence and the way in which it is used require careful scrutiny. The issues and challenges in conducting, interpreting, appraising, and synthesizing this type of research are substantial.

This chapter examines the nature of intervention effectiveness evidence, together with the scientific foundations for effectiveness research and its use. The fundamental strategy of identifying and addressing plausible alternative explanations for research findings is emphasized, together with the importance of qualitative reasoning and well-justified argument. The essential roles of theory and demonstrated mechanisms of change, converging evidence, and research critique are discussed. Common threats to validity are reviewed, as are threats to research integrity potentially fueled by largely unintentional conflicts of interest and motivated reasoning. A case example critiquing research syntheses on the effectiveness of interventions to reduce adolescent sexual risk behaviors is used to illustrate frequently encountered issues and challenges.

Interventions in Adolescent Health Psychology

> Health is a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity.
>
> (World Health Organization, 1946)

In the context of adolescent health psychology, an intervention is a systematic effort to promote the physical, mental, and social well-being of adolescents. Interventions are typically intended to work at one or more of the levels of individuals, families, systems, and communities. Interventions can involve population-based efforts such as outreach, social marketing, community organizing, and policy advocacy, or person-based efforts such as health education, case management, mentoring, consultation, and counseling. Because a majority of adolescents attend school, schools are common settings for adolescent health interventions, but interventions for adolescents also take place in community-based organizations, religious institutions, and in the broader community.

N.A. Constantine, Ph.D. (✉)
Center for Research on Adolescent Health and Development, Public Health Institute, 555 12th Street, Oakland, CA 94607, USA

School of Public Health, University of California, Berkeley, CA 94720, USA
e-mail: nconstantine@berkeley.edu

W.T. O'Donohue et al. (eds.), *Handbook of Adolescent Health Psychology*,
DOI 10.1007/978-1-4614-6633-8_20,

Interventions are sometimes classified within a disease-prevention framework comprising primary, secondary, and tertiary prevention (Williams, Holmbeck, & Greenley, 2002). Primary prevention interventions focus on avoiding the development of new health problems. In adolescent health psychology, this generally involves attempts to prevent or reduce health risk behaviors, for example, tobacco use, unsafe sex, or sedentariness. Positive health promotion and healthy development interventions also are considered primary interventions. Secondary prevention interventions provide early identification and treatment of existing health problems or established harmful health behaviors. Tertiary prevention interventions focus on the management and treatment of chronic diseases and conditions and of diseases with long-lasting consequences. This chapter focuses on methodological issues and strategies relevant to research on primary prevention and health promotion interventions, with most examples drawn from school-based risk behavior prevention interventions. The issues and strategies addressed, however, generally apply across other types of adolescent health psychology interventions as well.

Intervention Effectiveness

Increasingly, interventions are expected to be backed by evidence of effectiveness, and many funding sources formally require interventions to be "science-based" or "evidence-based." Intuitively this makes sense, especially in times of decreasing funds and increasing need. But it also raises potentially perplexing questions and opportunities for misunderstanding about the nature of effectiveness evidence and about standards of scientific evidence.

The concept of effectiveness might appear simple and straightforward—does an intervention accomplish what it was designed for? But answering this question requires complex judgments and tradeoffs. Part of the complexity involves specifying what is meant by effectiveness. In its broadest sense, effectiveness refers to meeting one or more intervention goals.[1] Most interventions have multiple goals, and the question of relative priority among goals is important. In appraising effectiveness, it is generally advisable to specify just one or a small number of primary-intended outcomes tied to the intervention's primary goal. Yet, it is possible that an intervention might not achieve its primary goal but still achieve one or more secondary goals. And by specifying a large enough number of secondary goals and outcomes, most interventions can be expected to statistically demonstrate success on at least one or a few of these just by chance alone, leaving the overall question of effectiveness debatable.

A related issue is that of socio-demographic and other moderators. Moderators are factors that affect the relationship between an intervention and its intended outcomes, leading to differential effectiveness in different subpopulations (also referred to as interactions). So another important question that must be addressed asks for which subpopulations is the intervention effective? What if an intervention appears to achieve its primary goal for girls but not boys? And what if this same intervention does not achieve its primary goal in a combined sample of girls and boys? What if it achieves its goal for Latina girls but not for non-Latina girls or for boys of any ethnicity? This subgroup division process could be further continued, increasing the likelihood that through chance differences alone a finding of effectiveness would emerge for some demographic or other subgroup level, again leaving the overall question of effectiveness debatable. And even if a purported effect of the intervention limited to a

[1] A separate question related to the meaning of effectiveness is how it differs from the concept of efficacy. Efficacy is used to refer to an intervention's success under ideal and highly controlled conditions, whereas effectiveness refers to an intervention's success under more typical real world conditions. Especially in medical research, efficacy studies are often conducted prior to effectiveness studies. While the focus of this chapter is on intervention effectiveness research, much of the discussion applies to efficacy studies as well.

specific subgroup is real, rather than due to chance, does it make sense to label an intervention as effective when it might only be so for a subgroup that represents a small proportion of the population for which the intervention was developed, and for which it is being promoted?

Evidence of Effectiveness

Evidence provides the grounds for a belief or judgment. It is "the raw material from which judgments, both of probability and of fact, are made" (Shafer & Tversky, 1985, p. 337). Evidence of effectiveness in regard to interventions in adolescent health psychology usually refers to research evidence, with special credibility given to research evidence that is believed to be scientific. For example, The National Campaign to Prevent Teen and Unplanned Pregnancy (Suellentrop, 2010) publishes a series of research briefs on the effectiveness of teen-pregnancy-prevention interventions titled *Science Says*, and Advocates for Youth (2008) publishes a similar series titled *Science and Success*. This focus on purported science is reinforced by a front page headline in a recent issue of the American Public Health Association membership newspaper: "Ineffective abstinence-only lessons being replaced with science: Teen pregnancy prevention focusing on evidence" (Krisberg, 2010). To be clear, this last example is not about replacing abstinence-only lessons with lessons on biology or chemistry. Instead, the use of the word science is intended to convey some ultimate credibility for the particular evidence that is the focus of the headline.

It is hard to argue with the desire for scientific evidence in evaluating intervention effectiveness and informing intervention adoption and funding decisions that follow. But to use evidence appropriately and responsibly requires that some critical questions are first addressed. For example, what counts as evidence, and when is evidence compelling? What counts as science, and what makes evidence scientific? Are some methods of developing evidence fundamentally better than others? What role do values, biases, and potential conflicts of interest play in selecting and appraising evidence? These and other related questions about the nature and use of evidence in science are sometimes minimalized or overlooked. A publication by the Centers for Disease Control and Prevention (2008) illustrates this tendency, defining a "science-based" teen pregnancy prevention program merely as "a program that research has shown to be effective in changing at least one (specified behavior)" (p. 24).

When these types of questions about scientific evidence and its use are addressed, intense disagreement among experts can result. Consider, for example, a debate within the American Evaluation Association (AEA) over the US Department of Education's priority statement on "scientifically based evaluation methods." The heart of the issue was the Department's statement, with substantial implications for funding eligibility, that "evaluation methods using [a randomized] experimental design are best for determining project effectiveness" (Scientifically Based Evaluation Methods, 2003, p. 62446). AEA submitted a board-approved position statement to the Department, objecting to the blanket nature of this conclusion and discussing other options and contextual considerations to inform the selection of the best methods. Shortly thereafter, a group of prominent evaluation theorists and methodologists, including several former AEA presidents, submitted a competing statement endorsing the Department's priority and its conclusions regarding the superiority of randomized experiments. One of the consequences of this debate was the resignation from the organization of a prominent former president and leading evaluation textbook author, who publicly stated his view that "AEA now has the same relationship to the field of evaluation as the Flat Earth Society has to the field of geology" (Lipsey, 2004, p. 9). When it comes to standards of scientific evidence, reasonable minds can differ, sometimes strongly.

Principles of Scientific Inquiry

A common belief among some researchers, many policy influentials and practitioners, and much of the general public is that science is defined by its

use of the "scientific method." In the general sense, this is supposed to consist of a series of steps beginning with observation and progressing to prediction, hypothesis, experimentation, and finally conclusion. More specifically, in intervention research the methods often equated to science are the randomized experiment (involving random assignment of units, for example, persons, schools, clinics, or communities, to intervention versus no-intervention control conditions) and the meta-analysis of randomized experiments (involving statistical cumulation of measures of effects across multiple studies). But science in the real world tends to be quite a bit more complicated and less orderly and defies any simple definition or defining characteristic. Many methods are used in science, and decisions about appropriate methods depend first and foremost on the particular research question to be addressed. Yet, even when a research method is well matched to the research question at hand, the science is only beginning. Methods are a means for obtaining evidence, but evidence rarely speaks for itself. And making good sense of evidence can be quite challenging.

Philosopher of science Susan Haack (2003) described scientific evidence as "complex and ramifying, structured more like a crossword puzzle than a mathematical proof. A tightly interlocking mesh of reasons well-anchored in experience" (p. 58). Campbell (2009) similarly spoke of the extended networks of implications within which scientific evidence must be presented and evaluated, and he emphasized the essential roles of plausible rival hypotheses and critical examination of their ramifications (i.e., implications):

> The core of the scientific method is not experimentation per se but rather the strategy connoted by the phrase "plausible rival hypotheses."… This strategy includes making explicit other implications of the hypothesis for other available data and reporting how these fit. It also includes seeking out rival explanations of the focal evidence and examining their plausibility. The plausibility of these rivals is usually reduced by ramification extinction, that is, by looking at their other implications on other data sets and seeing how well these fit. (p. 7)

Weiss (1980) concluded that "researchers bring not so much discrete findings as their whole theoretical, conceptual, and empirical fund of knowledge into the decision making process" (p. 12). From her cognitive-developmental psychology work on scientific reasoning and evidence appraisal, Koslowski similarly emphasized the importance of one's network of evidentially relevant collateral information to thinking in general and to scientific explanation in particular (Koslowski, 1996; Koslowski & Thompson, 2002). A common theme across these and other analyses (e.g., Chinn & Brewer, 2001; Evans, 1989; Gigerenzer, 2009) of the nature of scientific research evidence and its use is that "neither theory nor data alone is sufficient to achieve scientific success; each must be evaluated in the context of, and constrained by, the other" (Koslowski, 1996, p. 252).

These views of the inherent complexity of scientific evidence and the essential role of theory are at odds with the apparent beliefs of many adolescent health promotion researchers and research users, as well as evidence-based policy and practice proponents more generally. As currently understood and widely implemented, evidence-based policy and practice involve the assumption that scientific research evidence can be validly classified into hierarchical levels of quality according to the type of research methods employed to generate the evidence. And when theory is invoked it is often in name only, or in the form of what Gigerenzer (1998, 2009) has called *theoretical minimalism*—the application of surrogate theories such as one-word explanations, circular restatements, lists of vague dichotomies, and data fitting:

> The problem is not that a majority of researchers would say that theory is irrelevant; the problem is that almost anything passes as a theory…What distinguishes these surrogates from genuine theory is that they are vague, imprecise, and/or practically unfalsifiable. (Gigerenzer, 1998, p. 195)

In spite of an understandable desire for simplicity among consumers of research on intervention effectiveness, adequate appraisal of this evidence requires more than consulting a hierarchy of design and analysis methods or a checklist

of basic research quality criteria. Most fundamentally, scientific research interpretation and appraisal requires scrupulous attention to theory-informed plausible rival hypotheses or plausible alternative explanations and their implications, by scientists as well as by research consumers. Doing this well calls for deep substantive knowledge of the subject matter and context, strong theoretical grounding, and rigorous critical thinking and reasoning (Abelson, 1995; Campbell, 1982; Freedman, 2010; Levy, 2010).

National Research Council Report

To help mediate the debate regarding the appropriate role of randomized experiments in educational research, the National Research Council (NRC) Committee on Scientific Principles for Educational Research (2002) discussed science as "competent inquiry that produces warranted assertions, and ultimately develops theory that is supported by pertinent evidence" (p. 54). Consistent with modern views of scientific evidence such as those discussed above as espoused by Haack (2003), Campbell (2009), Weiss (1980), and Gigerenzer (1998), six guiding principles for scientific research emerged from the committee's work (see Table 1). These principles "provide a framework for how valid inferences are supported, characterize the grounds on which scientists criticize one another's work, and with hindsight, describe what scientists do" (p. 54). Although developed in the context of educational research, they provide a solid frame of reference for intervention effectiveness research in adolescent health psychology (and many other fields) as well. The committee emphasized the following:

> Scientific research, whether in education, physics, anthropology, molecular biology, or economics, is a continual process of rigorous reasoning supported by a dynamic interplay among methods, theories, and findings. It builds understandings in the form of models or theories that can be tested. Advances in scientific knowledge are achieved by the self-regulating norms of the scientific community over time, not, as sometimes believed, by the mechanistic application of a particular scientific method to a static set of questions. (p. 2)

In discussing these principles, committee members made clear that they were specifically focusing on scientific research, yet not intending to minimize the importance of other types of scholarship such as humanistic, historic, and philosophical approaches (Feuer, Towne, & Shavelson, 2002; NRC Committee on Scientific Principles for Educational Research, 2002). A key point made throughout the committee's report and supporting materials was that particular research methods or designs do not make a study or program of research scientific:

> Judgments about scientific merit of a particular method can only be accomplished with respect to its ability to address a particular question at hand.…No method is good, bad, scientific, or unscientific in itself: Rather, it is the appropriate application of method to a particular problem that enables judgments about scientific quality. (Feuer et al., 2002, pp. 7–8)

The committee distinguished between three interrelated types of research questions: *description* (What's happening?), *cause* (Is there a systematic effect?), and *process or mechanism* (Why or how is this happening?). It discussed a variety of methods that have been successfully applied to each type of question, and it emphasized the importance of addressing all three types of questions in a program of research, together with the concurrent need for multiple methods (NRC Committee on Scientific Principles for Educational Research, 2002; Shavelson & Towne, 2004).

The committee's report was generally well received as articulating a responsible middle ground between the simplistic and extremist view that only randomized experiments can provide credible scientific evidence, and the equally simplistic and extremist view that science is hopelessly flawed and all research standards are arbitrary. For example, Berliner (2002) supported the committee's recommendations and commended its strong emphasis on science beyond randomized experiments. At the same time, he criticized the report for insufficiently addressing the unique complexity of educational research as compared with other fields of scientific research,

Table 1 Guiding principles of scientific research

Scientific principle 1 *Pose significant questions that can be investigated empirically* Moving from hunch to conceptualizing and specifying a worthwhile question is essential to scientific research. The questions, and the designs developed to address them, must reflect a solid understanding of the relevant theoretical, methodological, and empirical work that has come before.
Scientific principle 2 *Link research to relevant theory* It is the long-term goal of much of science to generate theories that can offer stable explanations for phenomena that generalize beyond the particular. Science generates cumulative knowledge by building on, refining, and occasionally replacing, theoretical understanding.
Scientific principle 3 *Use methods that permit direct investigation of the question* Methods can only be judged in terms of their appropriateness and effectiveness in addressing a particular research question. Moreover, scientific claims are significantly strengthened when they are subject to testing by multiple methods.
Scientific principle 4 *Provide a coherent and explicit chain of reasoning* Making scientific inferences is not accomplished by merely applying an algorithm for using accepted techniques in correct ways. Rather, it requires the development of a logical chain of reasoning from evidence to theory and back again that is coherent, shareable, and persuasive to the skeptical reader.
Scientific principle 5 *Replicate and generalize across studies* Scientific inquiry emphasizes checking and validating individual findings and results. Ultimately, scientific knowledge advances when findings are reproduced in a range of times and places and when findings are integrated and synthesized.
Scientific principle 6 *Disclose research to encourage professional scrutiny and critique* Scientific studies do not contribute to a larger body of knowledge until they are widely disseminated and subjected to professional scrutiny by peers. Indeed, the objectivity of science derives from publicly enforced norms of the professional community of scientists, rather than from the character traits of any individual person or design features of any study.

NRC Committee on Scientific Principles for Educational Research (2002, *pp. 3–5*)

especially in regard to the importance of personal, cultural, and educational contexts and to the ubiquity of interactions (differential effects in different subpopulations) in education research. Maxwell (2004) went further in his critique, arguing that the report inadequately addressed the importance of process, mechanism, and context in establishing and understanding intervention effects and other types of causation, and that it misrepresented the nature and potential value of qualitative research. According to Maxwell, qualitative methods should not be relegated to just descriptive and exploratory research questions but are important components of fully addressing questions of causation and mechanism as well. Despite these and other criticisms, the committee's report was a remarkable accomplishment and its primary messages still stand well. Its six principles of scientific inquiry provide a sound framework for designing, interpreting, and critically appraising intervention effectiveness research in adolescent health psychology.

Validity: How Might Research Conclusions Be Wrong?

Validity refers to the correctness of an inference or conclusion. "Validity is a property of inferences. It is not a property of designs or methods, as the same design may contribute to more or less valid inferences under different circumstances" (Shadish, Cook, & Campbell, 2002, p. 35). Validity is important to all types of scientific research. One of the best summaries of validity has been provided by a qualitative researcher:

> Validity is a goal rather than a product. It is never something that can be proven or taken for granted. Validity is also relative. It has to be assessed in relationship to the purposes and circumstances of the research, rather than being a context independent property of methods or conclusions. Validity threats are made implausible by evidence, not methods, methods are only a way of getting evidence that can help you rule out these threats. (Maxwell, 2005, p. 105)

A distinction between two primary types of validity that are especially relevant to quantitative research on intervention effectiveness was first articulated by Campbell (1953, 1957) and further developed by Campbell and Stanley (1963). *Internal validity* is the basic minimum for interpretation of an intervention study's findings, and it relates to the fundamental question of causation: Did the intervention contribute causally to a change in the outcome? *External validity* is concerned with generalizability: To which populations, settings, times, treatments, and outcomes can results be generalized? Subsequently (Cook & Campbell, 1979; Shadish et al., 2002), two additional types of validity were spun off from these original two and further developed. *Statistical conclusion validity* is a basic component of internal validity, regarding the magnitude of the association between an intervention and an outcome and the possibility that it might be due to chance, regardless of the question of causality. *Construct validity* like external validity involves questions of generalizability, but specifically in reference to the link between abstract constructs and operationalization of these constructs in the research: did we implement the intervention we intended to implement and did we measure the outcome we intended to measure?

Common Threats to Validity

The foundation of building a case for the validity of research inferences involves identifying and ruling out plausible rival hypotheses, or plausible alternative explanations, for research findings. For example, if adolescents who voluntarily sign a virginity pledge are found more likely to remain virgins, is this difference between pledgers and nonpledgers due to the pledging itself, or might it be due to a preexisting inclination to abstain from sex among those adolescents who voluntarily sign the pledge? Such plausible alternative explanations are also referred to as threats to validity. Campbell and Stanley (1963) originally discussed eight threats to internal validity and four threats to external validity in the context of intervention effectiveness research. These lists of threats have grown over time and with the further development of the validity typology. Shadish et al. (2002) discussed 36 specific threats, together with additional threats due to combinations of or interactions between the basic threats. Each of these threats represents a potential alternative explanation for a particular research finding that can challenge the conclusions and interpretations drawn by researchers and research users.

Threats to internal validity have received the most attention, and seven of the most prominent of these threats are listed in Table 2. Among these, selection threats can be especially daunting and often are insufficiently addressed in intervention research (Larzelere, Kuhn, & Johnson, 2004). Although selection threats can involve preexisting group differences from any nonrandomized selection mechanism, such as natural, administrative, and convenience selection, the most dangerous type of selection threat arises from motivated self-selection of individuals into intervention versus control conditions.

Threats to external validity involve the potential for unwarranted generalizations of intervention effectiveness inferences, including any interactions of the intervention's potential effectiveness with settings, populations, or outcomes. Threats to statistical conclusion validity include low statistical power due to small sample sizes or unreliable measures, as well as inflated probability of finding significant intervention effects due solely to chance (i.e., Type I error) through inappropriate use of statistical analysis methods. Two common practices that can substantially increase the probability of a Type I error are the conduct of large numbers significance tests without statistical adjustment for this multiple testing, and failure to adjust for the statistical clustering that results from assigning groups rather than individuals to

Table 2 Threats to internal validity

Selection. Preexisting differences between intervention and control groups, which can be especially serious when these differences are due to individuals' motivated self-selection
History. Extraneous events occurring during the intervention that could affect the outcome
Maturation. Naturally occurring changes in participants over time
Regression. Natural movement on subsequent measurements toward the overall group average—especially for groups composed on the basis of extreme scores
Attrition. Differential loss of participants between groups
Testing. Practice effects or other factors based on repeated exposure to the assessment instrument
Instrumentation. Changes in the function or meaning of the measures used over time or between groups

intervention and control conditions. Construct validity threats include inadequate implementation of an intervention (Dane & Schneider, 1998) and inadequate development of construct definitions and operationalizations, as well as situations in which program administrators or staff provide unplanned compensatory services to those not receiving the intervention (compensatory equalization) or when those not receiving the intervention are so resentful that they respond more negatively than they otherwise would have (resentful demoralization). Other threats to validity and in-depth discussions and examples of these threats can be found in Shadish et al. (2002).

Trochim and Donnelly (2007) discussed five general approaches to addressing threats to validity in quantitative research. First, a well-reasoned *argument* that explains why a potential threat is not likely can sometimes suffice. Second, systematic *measurement or observation* of plausible alternative explanations can provide evidence on whether a potential threat is occurring. Third, r*esearch design* is commonly used to rule out alternative explanations through strategies such as employing control groups that do not receive the intervention, or incorporating multiple waves of measurement to obtain data on existing trends in outcomes independent of the intervention. Fourth, *statistical analysis* can be used to test for suspected threats, such as differential attrition between the intervention and control groups, and under some special circumstances and strong assumptions, to reduce these threats through statistical adjustment. Finally, anticipated threats can sometimes be eliminated through *preventive action*, such as use of sample incentives to reduce attrition, or quality control procedures to identify and remediate data errors. These five approaches are not mutually exclusive, and in general it is preferable to use multiple methods to minimize threats to validity. In particular, argument development is always part of making a case for the validity of inferences or conclusions (Maxwell, 2005; Victora, Habicht, & Bryce, 2004).

Maxwell (2005) discussed validity issues in qualitative research from a similar perspective of identifying plausible alterative explanations and threats to the valid interpretation and understanding of research findings. These included biased selection or interpretation of data by the researcher (researcher bias) and any influence of the researcher on the setting or individuals studied (reactivity and reflexivity). These two threats can be relevant to quantitative research as well.

Research Designs

Research designs provide the blueprints from which research studies are built, and play a central role in addressing threats to validity, especially internal validity. Many threats to internal validity can be minimized or eliminated through the careful use of an appropriate randomized experimental or quasi-experimental design. Nonexperimental observational designs also can be used to address threats to validity, but generally on a more limited basis. Finally, qualitative designs have a unique and complementary role to play in addressing validity threats and enhancing intervention effectiveness research.

Randomized Experiments and Quasi-Experiments

Randomized experiments (sometimes referred to as randomized controlled trials [RCTs] or

randomized field trials) involve the random assignment of units such as persons, schools, clinics, or communities to intervention versus nonintervention control conditions. This is done to control or minimize potential threats to internal validity and can be especially powerful in reducing or eliminating selection effects. The putative power of randomization is its potential "to control an infinite number of rival hypotheses without specifying what any of them are" (Campbell, 2009, p. viii). Scriven (2008) cautioned, however, that most randomized experiments as actually implemented do not eliminate all plausible alternative explanations for purported effects, and that "the RCT banner in applied human sciences is in fact being flown over pseudo-RCT's" (p. 13).

This criticism derives from the understanding that randomization alone does not yield an RCT—other essential aspects of an RCT include, for example, a focus on a single or very few primary outcome measures that are specified prior to the start of data collection, and double-blinding of treatment and control group conditions so that neither the investigators nor the participants know the participants' treatment assignments (Meinert, 1986). These conditions are rarely met in field-based intervention research, and in fact, the ideal of double-blinding is commonly not met even in clinical research (Abel & Koch, 1999; Meinert, 1986).

Additional criticisms related to internal validity threats are based on other fundamental assumptions of an RCT. One of the most important of these assumptions is that participants accept and maintain their intervention assignments and that any refusal to participate (selection), loss of participants (attrition), or differential levels of participation that occur during the course of a study are not related to group assignment (West, 2009). RCTs have also been criticized for threats to external validity and construct validity (e.g., differences between the study protocol and routine practice) that are widely believed to be greater in randomized designs than in other types of designs (Rothwell, 2005). Cook (2002) provided a comprehensive review of criticisms of randomized experiments in school-based research. Although still supporting random assignment as the best and most credible mechanism for justifying causal conclusions about intervention effectiveness, he acknowledged that

> random assignment cannot be considered the "gold standard" for justifying causal inferences in school-based research. It creates only a probabilistic equivalence between the groups being contrasted, and then only at pretest. Moreover, treatment-correlated attrition is likely when treatments differ in intrinsic desirability. Also, treatments are not always independent of each other in practice like they are supposed to be in theory, and many of the ways used to increase internal validity can also reduce external validity. (p. 195)

None of these criticisms negate the potential power of a properly designed, implemented, maintained, and interpreted randomized experiment to yield strong evidence with regard to intervention effectiveness. Instead, they are reminders that randomization in itself does not necessarily eliminate important threats to validity (Abel & Koch, 1999; Scriven, 2008).

Quasi-experiments do not involve random assignment to intervention and control conditions but instead employ some combination of other design features to help rule out alternative explanations of observed effects. The quasi-experimental label is often applied to weak designs based on comparisons of preexisting groups composed of members who have self-selected into intervention and control conditions, and characterized by just one or two waves of data collection. Yet much more powerful and sophisticated quasi-experimental designs can be developed through the use of strategies such as matching or stratifying participants into intervention and control conditions, scheduling of multiple preintervention and postintervention measurements and time points, employing multiple treatment and comparison groups, and manipulating intervention timing. A variety of quasi-experimental designs involving these strategies has evolved over time (Campbell & Stanley, 1963; Cook & Campbell, 1979), and these designs have been discussed in depth by Shadish et al. (2002).

Both randomized experiments and quasi-experiments offer the potential to reduce the likelihood of plausible alternative explanations for a

purported effect. To be realized, this potential requires skillful application under the right circumstances and conditions. It is the researcher's responsibility to argue and sufficiently document the case that an appropriate design has been developed and skillfully applied to the research questions addressed.

Nonexperimental Observational Studies

Observational studies employ quantitative research methods to make inferences about causal risk factors or intervention effects in the absence of researcher control over most threats to internal validity. Intervention and control groups are based on existing memberships or conditions, and no controlled manipulation of intervention exposure occurs. These groups may or may not be based on self-selection, for example, an adolescent voluntarily choosing to make a virginity pledge. Observational studies often involve secondary analysis of existing population-based data sets. The National Longitudinal Study of Adolescent Health (Add Health) and the National Survey of Family Growth (NSFG) are two large national survey databases sometimes used in observational studies in adolescent health psychology. With observational studies that involve comparisons of respondents who experienced some type of intervention with those who did not, statistical analyses are commonly used to try to remove (i.e., statistically adjust or control for) preintervention group differences that could be the cause of group outcome differences. For example, using the Add Health dataset, Bearman and Bruckner (2001) attempted to show that virginity pledging delays initiation of sexual intercourse. Because pledgers and nonpledgers differed on many background variables (such as religiosity) that were associated with making the decision to pledge, the researchers statistically adjusted their data in an attempt to remove these prepledge differences. In this situation, however, it is hard to imagine how preexisting motivational inclinations among voluntary (self-selected) pledgers to abstain from sex could be meaningfully removed by statistical methods. As Anderson (1963) complained nearly a half century ago, in situations like this, "One may well wonder what exactly it means to ask what the data would look like were they not what they are" (p. 170). Along the same lines, Lord (1967) cautioned:

> With the data usually available for such studies, there is simply no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups. The researcher wants to know how the groups would have compared if there had not been preexisting uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of the available data. (p. 305)

Further critique and discussion of the virginity pledge example is provided in Constantine and Braverman (2004).

Modern statistical analysis methods provide an abundance of complex methodologies intended to better achieve the types of statistical adjustments that so perplexed Anderson and Lord. Yet, in most real world situations, these new methods are plainly unable to meet their hypothetical potential. Light, Singer, and Willett's (1990) admonition that "you can't fix by analysis what you bungled in design" (p. v) remains relevant. A lament by the editors of the International Journal of Epidemiology reinforces this point:

> Observational studies revealed strong apparently protective effects of beta-carotene, but long term RCTs found that, if anything, beta-carotene increased cardiovascular disease risk. There are now a series of similar examples: hormone replacement therapy, vitamin E and vitamin C intake in relation to cardiovascular disease, or fiber intake in relation to colon cancer among them. What these examples have in common is that the groups of people who were apparently receiving protection from these substances in the observational studies were very different from the groups not using them, on a whole host of characteristics of their lives. Belief that these differences could be summed up in measures of a few "potential confounders" and adequately adjusted for in statistical analyses, fails to recognize the complexity of the reasons why people differ with regard to particular and general characteristics of their lives. (Davey Smith & Ebrahim, 2001, p. 5)

Observational studies can provide evidence relevant to the understanding of the effectiveness of an adolescent health intervention. This generally occurs not through automatic use of complex statistics, but instead through a careful analysis and understanding of potential alternative explanations and threats to validity (Constantine, 2012). Evidence from the best of these observational studies can then be used for two primary purposes. First, this evidence can help justify the need for more controlled and expensive randomized or quasi-experimental studies. Second, as one component of a comprehensive evidence review to be combined with evidence from other studies, observational study results can be part of a critical review of the convergence of evidence across studies that experience different threats to validity and have complementary strengths and weaknesses.

For example, Kohler, Manhart, and Lafferty (2008) employed 2002 NSFG data to evaluate the effectiveness of sex education programs at the United States population level. They found that adolescents who received comprehensive sex education were significantly less likely to report teen pregnancies than were those who received either no sex education or abstinence-only sex education. These findings resulted from a strong design and analysis that statistically controlled for plausible alternative explanations based on preexisting group differences. The main reason that this was possible is that type of sex education received is much less likely due to purposeful self-selection than are such conditions as virginity pledging or dietary habits. Absent purposeful self-selection, preexisting group differences (e.g., family income) that might influence both sex education received and sexual behavior outcomes can be more amenable to meaningful statistical adjustment. Nevertheless, cautious interpretation and further study is warranted. One of the strengths of the findings from this study is that it provided convergent validity when combined with other types of available research evidence on the relative effectiveness of comprehensive versus abstinence-only sex education (Constantine, 2008a).

Qualitative Research

Unlike randomized experiments, quasi-experiments, and observational studies, all of which primarily employ the analysis of quantitative data, qualitative research involves the analysis of unstructured data such as interview transcripts, open-ended survey responses, behavior observations, and text materials. Typically, qualitative research focuses more on the why and how of behavior and other phenomena, whereas quantitative research focuses more on the what, whether, where, when, and how much.

Qualitative research is commonly regarded as a useful adjunct (or precursor) to experimental or quasi-experimental designs. Yet, Maxwell (2004) took issue with this hierarchical characterization, arguing that valid causal inference requires that qualitative research be given an equal place at the table. While acknowledging the important and more typically recognized exploratory value of qualitative research for hypothesis and theory development and its explanatory value in helping to elucidate quantitative findings, Maxwell saw a more fundamental role for qualitative research in supporting causal inferences about intervention effectiveness, arguing that the qualitative study of causal processes is indispensable for most causal inferences. This argument was supported by Freedman (2008): "Scientific inquiry is a long and tortuous process, with many false starts and blind alleys. Combining qualitative insights and quantitative analysis—and a healthy dose of skepticism—may provide the most secure results" (p. 313). Freedman (2008) further explicated the role of qualitative causal process observations in 10 of the major scientific discoveries from the histories of medicine and public health, such as the discovery of penicillin and the development of the smallpox vaccine, illustrating how

> progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. The examples show that qualitative methods can play a key role in all three tasks. (p. 312)

Several qualitative research frameworks for rigorous causal analysis have been developed, including Maxwell's (2005) interactive approach,

Miles and Huberman's (1994) cross-case analysis approach, and Yin's (2008) multiple case study approach. As with all types of methods and designs used in intervention effectiveness research, these approaches require diligence in recognizing and minimizing threats to validity and are best used as part of a well-integrated combination of complementary methods within a study or across a program of research.

Research Integrity: How Might Research Conclusions Be Biased?

> Bias is not a crime, is not necessarily intentional, and is not a sign of lack of [personal] integrity; rather, it is a natural human phenomenon . . . everyone is likely capable of rationalizing beliefs and denying influences that bias them.
>
> (Cain & Detsky, 2008)

Research integrity involves a commitment to intellectual honesty and to a range of practices that characterize responsible research conduct (National Research Council [NRC] Committee on Assessing Integrity in Research Environments, 2002). Although practices related to human subject protection, accurate representation of authorship roles, and research management are important aspects of research integrity, this section focuses specifically on those related to intellectual honesty in performing, interpreting, and using research. These issues apply not only to research scientists and their institutions, but also to advocates, journalists, bureaucrats, and other policy shapers who are part of the chain of research creation, communication, and use.

Conflicts of Interest and Motivated Reasoning

Issues in research integrity are often based in conflicts of interest, which occur when individuals' personal interests are in conflict with their professional judgment and obligations (Gorman & Conde, 2007; Kumar, 2008; Young, 2009). These competing personal interests can be directly or indirectly financial, or more broadly related to the goals of the individuals or their organization (Bachrach & Newcomer, 2002; Ioannidis, 2011; Smith, Feachem, Feachem, Koehlmoos, & Kinlaw, 2009; Young, 2009). MacCoun (2005) has placed conflicts of interest in public policy research "on a continuum from blatant pecuniary bias to more subtle ideological bias" (p. 233), whereas Chugh, Bazerman, and Banaji (2005) have distinguished among three types of conflicts of interest: the plainly visible, the visible yet dismissed through disclosure or denial, and the invisible.

For any type of conflict of interest, *bounded ethicality* can make it difficult to overcome or even to recognize one's own conflicts and biases. Bounded ethicality involves ethically limited judgment and decision making due to largely unconscious biases and ego protective mechanisms. This is enabled by "an ethical blind spot [that] emerges as decision makers view themselves as moral, competent, and deserving, and thus assume that conflicts of interest are nonissues" (Chugh et al., 2005, p. 80). Bounded ethicality has been well documented in studies of the psychological aspects of conflicts of interest and implausible denials in the field of financial auditing (Chugh, et al., 2005; Moore, Lowenstein, Tanlu, & Bazerman, 2003).

Feinstein (1988) discussed several types of biases in the quest for scientific truth, especially distinguishing between deliberately planned fraud and inadvertent deception. Inadvertent deception was further divided into one-time distortions of evidence versus more robust delusions: "A distortion is usually produced by failure to recognize important distinctions in the complexity of nature, [whereas] a delusion usually arises from excessive zeal in the expectations, beliefs, or behavior of the investigators" (pp. 475–476). Each of these can contribute to deluded consensus among experts, or the *consensus syndrome*, which Feinstein argued is particularly detrimental to scientific progress. MacCoun (1998) similarly concluded that "under a wide variety of circumstances, collective decision making will significantly amplify individual bias, rather than attenuate it" (p. 278). Both Feinstein

and MacCoun emphasized the importance and prevalence of bias and deception that occur outside of the realm of deliberately planned fraud. Cain and Detsky (2008) concurred:

> Conflicts of interest are problematic not only because they are widespread but also because most people incorrectly think that succumbing to them is due to intentional corruption, a problem for only a few bad apples. . . . (S)uccumbing to a conflict of interest is more likely to result from unintentional bias, something common in everyone. (p. 2893)

Much conflict of interest involves *motivated reasoning*, the unintentionally biased appraisal of evidence to support one's goals through a set of preconscious cognitive processes. These processes include biased selection of evidence, biased access to background beliefs, and biased selection of statistical reasoning heuristics (Dawson, Gilovich, & Regan, 2002; Evans, 1989; Kunda, 1990). Heuristics are simple rules of thumb that are generally true in many but not all situations. For example, a common statistical reasoning heuristic is the belief that larger sample sizes lead to more reliable and valid results. Statistical reasoning heuristics are often applied differentially to research evidence that supports or challenges one's motivated beliefs:

> Heuristics that have judgmental implications congenial to perceivers' existing beliefs are especially likely to be used, whereas incongenial heuristics may be ignored or disparaged. . . . Information that is congruent with one's existing beliefs, such as research supporting one's position on abortion, will be judged more favorably than incongruent information . . . [while] incongruent information may be scrutinized in an effort to derogate its validity. (Chen & Chaiken, 1999, p. 45)

Level of motivation and type of motivation are both important determinants of the nature of biased cognitive processing that occurs (Chen & Chaiken, 1999). Motivated reasoning and its biases can affect research design and implementation, research interpretation, and research appraisal and synthesis.

In adolescent health psychology, common situations that might lead to real or apparent conflicts of interest, bounded ethicality, and motivated reasoning include effectiveness research conducted by intervention program developers or publishers, and research reviews or syntheses conducted by researchers involved in some of the reviewed research. Moskowitz (1993) argued that "much of the drug abuse (prevention) research conducted to date suffers from real or apparent conflicts of interest" (p. 7), and discussed a variety of motivations and pressures for these conflicts, primarily arising from investigators evaluating programs that they or their institutions developed. Weiss and colleagues (Gandhi, Murphy-Graham, Petrosino, Chrismer, & Weiss, 2007; Weiss, Murphy-Graham, Petrosino, & Gandhi, 2008) raised similar concerns regarding conflict of interest in drug prevention intervention research and research use. Gorman and Conde (2007) quantified this phenomenon in a study of the 34 model school-based interventions for drug and violence prevention in the Substance Abuse and Mental Health Services Administration National Registry of Effective Programs. Of the 246 published evaluation reports located for these interventions, 78 % included the intervention developer as an author of the evaluation report, and for another 11 %, the developer had some other association such as working in the same organization as one of the evaluation report authors. Only 11 % showed no identifiable association between the evaluation authors and the program developer.

Threats to Research Integrity

Threats to validity have been well studied and publicized, and a variety of strategies for dealing with these threats has been developed. Threats to research integrity can be just as damaging or even more so. Growing bodies of research on unconscious conflicts of interest, bounded rationality, unintentional biases, and motivated reasoning in evidence selection and appraisal help explain the etiology of these threats and some of the cognitive and affective mechanisms behind them. It is also useful to consider the common methodological mechanisms that comprise these threats.

1. *Multiple significance testing* (*fishing for significance*, *data dredging*). Sometimes referred to simply as *multiplicity*, this involves testing large numbers of potential

outcomes for statistical significance, and capitalizing on the increased likelihood of finding spurious effects due to chance as more outcomes are tested (Feinstein, 1988; Howel & Bhopal, 1994; Mills, 1993).

2. *Within-study selective reporting* (*data suppression*, *cherry picking*). This type of threat builds on multiple significance testing but goes a step further, involving the selective reporting or combining of results across multiple outcomes, subgroup analyses, and other multiplicities, such that results that support the researcher's hypotheses are more likely to be reported than are those that do not (Chan, Hrobjartsson, Haahr, Gotzsche, & Altman, 2004; Dwan et al., 2008; Hahn, Williamson, & Hutton, 2002; Ioannidis, 2005; Kumar, 2008; Mills, 1993; Simmons, Nelson, & Simonsohn, 2011).
3. *Exploiting ambiguities* (*researcher degrees of freedom*). Simmons and colleagues (2011) discussed a variety of ambiguities that researchers commonly exploit to increase the likelihood of a positive result. In addition to multiple significance testing, these include decisions about deleting outliers (suspicious extreme values in the data), choosing sample size, using covariates, and reporting subsets of treatment conditions. Testing several of these through computer simulations of experimental data, they reported a 61 % false positive rate, i.e., "A researcher is more likely than not to falsely detect a significant effect by using these four common researcher degrees of freedom." (p. 1361).
4. *Biased misreporting of statistical results*. Errors in reporting of statistical results have been found widely prevalent in peer reviewed articles published in natural science and medicine (Garcia-Berthou & Alcarz, 2004), psychiatry (Berle & Starcevic, 2007), and psychology (Bakker & Wicherts, 2011). Bakker and Wicherts found in their sample of articles that these errors fell overwhelmingly (92 % for congruence errors and 100 % for rounding errors) in the direction to lend support for the researcher's hypotheses and expectations. Friedlander (1964) commented nearly a half century ago on this particular mechanism of Type I bias, which he attributed to the natural tendency of researchers, himself included, to more readily investigate and verify results that do not support their expectations.
5. *Hypothesizing after the results are known* (*HARKing*, *data-driven hypothesizing*). HARKing involves presenting a post hoc hypothesis developed from a study's results as if it were an a priori hypothesis confirmed by these results (Kerr, 1998; Kumar, 2008).
6. *Methodological impenetrability* (*statisticization*). This involves the use of unnecessarily complex analysis methods and impenetrable descriptions of these methods to discourage critical appraisal by others. "If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as to theories." (Wilkinson & APA Task Force on Statistical Inference, 1999).
7. *Selective publication* (*publication bias*). Selective publication of manuscripts based on the direction and magnitude of results has been well documented. See especially a systematic review of eleven studies investigating publication bias in health-care intervention research by Dwan and colleagues (2008). In particular, research with statistically significant positive results is more likely to be submitted for publication, to be published, and to be published more quickly than research with negative or null results (Constantine, 2008c; Dwan et al., 2008).
8. *Redundant publication*. This involves publication of the same results multiple times as if they were independent replications (Constantine, 2008c; Huston & Moher, 1996; Kassirer & Angell, 1995; Rennie, 1999). In addition, *data augmentation* occurs when after publishing results, additional data are collected and combined with the originally published data and then published as a new study (Kumar, 2008).
9. *Biased peer-review*. The influence of a reviewer's personal biases on the results and recommendations of their peer reviews has

been widely demonstrated (Altman, 2002; Ioannidis, Tatsioni, & Karassa, 2010; Mahoney, 1977; Shatz, 2001; Young, Ioannidis, & Al-Ubaydi, 2009). *Confirmation bias* in peer review involves the general tendency to less critically evaluate evidence that is consistent with one's existing beliefs. *Ideological bias* occurs when a reviewer's values-based views for or against an author's position unduly influence a review. *Ad hominem* and *affiliation biases* are found when a review is influenced by knowledge of the author's identity or affiliation (Constantine, 2008b).

10. *Postpublication peer review limitations*. Postpublication peer review includes letters to the editor as well as full articles critiquing a published work. As Altman (2002) cautioned, "many readers seem to assume that articles published in peer-reviewed journals are scientifically sound, despite much evidence to the contrary. It is important, therefore, that misleading work be identified after publication" (p. 2766). Authors sometimes choose to ignore a published critique or respond minimally to peripheral issues in place of the specific criticisms made. Even when serious errors are detailed in a critique, retractions or corrections are the exception. PsycINFO and other databases rarely link postpublication critiques to the original article, and narrative reviews and other research syntheses that cite a criticized work frequently ignore the critique (Altman, 2002; Rennie, 1998). Another aspect of this threat is *selective data sharing*—researchers' reluctance to share raw data for reanalysis and external verification (Wolins, 1962; Wicherts, Borsboom, Kats, & Molenaar, 2006), and the inverse relationship between this reluctance and strength of evidence and reporting quality (Wicherts, Bakker, & Molenaar, 2011).
11. *Motivated communication of results*. This involves selective emphasis of a study's supportive versus limiting conclusions by researchers, funders, media, or advocates (Constantine & Braverman, 2004; Scher, Lin, & Constantine, 2009), and more generally, the minimization of study limitations by researchers, reviewers, and research users (Ioannidis, 2007). Cronbach's (1982, p. 108) caution that "validity depends not only on data collection and analysis but also on the way a conclusion is stated and communicated," applies to research integrity as well as research validity.
12. *Biased research synthesis*. As with individual studies, research syntheses can be affected by selective inclusion of studies or outcomes based on the direction of their results (Dwan et al., 2008; Hahn, Williamson, Hutton, Garner, & Flynn, 2000; Ioannidis & Karassa, 2010). In addition, biases in the included individual studies can carry over to the research synthesis, especially when these threats appear in multiple studies.

Issues in Consolidating Research Evidence

Rarely is a research question settled by a single study. To better address questions of intervention effectiveness, research evidence from multiple sources must be located, appraised, and consolidated. This activity is referred to as research synthesis, comprising a set of processes through which multiple research studies are reviewed and assessed with the objective of summarizing the evidence relating to a particular question. The most common types of research synthesis in adolescent health psychology are narrative reviews, programs-that-work lists, and systematic reviews and meta-analyses.

Narrative Reviews

The *narrative review* designation is used in a variety of ways, sometimes to indicate a review that does not meet standards of rigor expected of a systematic review. It also is sometimes used synonymously with the term literature review. Narrative reviews range from primarily descriptive to primarily critical. A *descriptive narrative*

review attempts to summarize research results relating to a particular question, whereas a *critical narrative review*, sometimes referred to as an integrative literature review, "presents a logically argued case founded on a comprehensive understanding of the current state of knowledge about a topic of study" (Machi & McEvoy, 2009, p. 4).

Descriptive narrative reviews frequently include a *box score* presentation of empirical results. This takes the form of a table of intervention studies and outcomes tested, with check marks or other indicators to denote whether the study reported a statistically significant result in the expected direction on each tested outcome. Descriptive narrative reviews have been widely criticized as especially susceptible to reviewer bias and publication bias due to insufficiently objective and transparent criteria for selection of studies and appraisal of results. And the use of box score approaches in narrative reviews has been criticized as an inappropriate overuse of statistical significance tests without regard for magnitudes of the reported effects (Egger & Davey Smith, 1997; Shadish et al., 2002; Slavin, 1995).

Whereas descriptive narrative reviews tend to focus on empirical evidence alone, critical narrative reviews generally make more extensive use of theory to integrate empirical evidence. The latter consider both supportive and challenging evidence, with special attention to plausible alternative explanations and their implications. Good examples of a theory-focused critical narrative reviews can be found in such journals as *Psychological Bulletin*, *Perspectives on Psychological Science*, and *School Psychology Review*. Compared with other forms of research synthesis, critical narrative reviews tend to involve more complex forms of argumentation and justification, and more nuanced answers to research questions. Accordingly, they can be more difficult to develop, and more difficult to translate into black and white research-based policy and practice decisions. This might explain why they have not been more commonly used in synthesizing intervention research.

Programs-That-Work Lists

Programs-that-work lists are a second type of research synthesis, comprising lists of interventions that meet prespecified criteria of effectiveness findings. They are sometimes referred to as evidence-based or science-based program lists, or best practice lists. Lists of this nature are often used to determine program eligibility for federal or other types of funding, and they have been prevalent at least since the introduction in 1992 of the Centers for Disease Control and Prevention now-defunct lists of effective health education programs (Collins et al., 2002). There is now a proliferation of lists in many areas of adolescent health psychology.

Although they vary widely in the criteria used, most programs-that-work lists allow the inclusion of a program based on just one study with one statistically significant result, regardless of the number of outcomes tested within a study or the number of studies conducted on the same intervention. In other words, an intervention program for which just 1 of 20 relevant tested outcomes is found to be statistically significant could earn a place on a programs-that-work list as an evidence-based program. In fact, exactly this did happen with the Second Step violence prevention curriculum—this intervention received an "exemplary program" certification by a U.S. Department of Education Safe, Disciplined, and Drug-Free Schools Expert Panel based on a randomized study yielding just 1 statistically significant outcome out of 20 relevant tests conducted (Grossman et al., 1997; critiqued in Constantine & Braverman, 2004 and Gorman, 2002). Although this example exhibits statistical irony in its precise details (1 in 20 statistically significant outcome tests is exactly what is expected by chance when an intervention has no effect and the tests are conducted with the usual significance level criterion of 0.05), the problem reflected is not at all unique.

Weiss and colleagues (Gandhi et al., 2007; Petrosino, 2003; Weiss et al., 2008) investigated seven prominent programs-that-work lists of school-based drug prevention interventions and the five programs appearing most frequently across the seven lists, concluding that "when we

look at all of the evaluations cited across the lists, we are disturbed by the frailty of evidence for some of the 'proven' programs" (Gandhi et al., 2007, p. 65). Several factors were identified to explain how so many questionably effective programs were ending up on these lists. These included the following: (a) the insufficient standard of requiring just one or two evaluations to designate a program as effective, (b) the common practice of conducting multiple significance tests of outcomes (with an example provided of a program that was listed based on two statistically significant outcomes out of 100 tests conducted), (c) the failure to adjust for clustering when interventions were assigned to groups rather than individual persons, and (d) the potential for conflicts of interest and biases due to the common practice of program developers' evaluating their own programs (Gandhi et al., 2007). This last point was extended to the review process itself: "The [program review] procedures used, even by a prestigious group of outside experts, seem to reflect a degree of bias and favoritism. Experts, it seems, may be as subject to human frailties as the rest of us" (Weiss et al., 2008, p. 43). A similar set of issues has been raised by Gorman (2002; Gorman & Conde, 2007), who concluded that "with regard to the entry criterion of one effect from one evaluation, this is far too low a standard by which to designate Exemplary status" (Gorman, 2002, p. 301).

An unusually rigorous programs-that-work system is the What Works Clearinghouse (2008), which is focused on educational programs and strategies. Among other evidence standards employed, What Works Clearinghouse requires that study results be adjusted for biases that arise from assigning interventions at the group (e.g., school or classroom) level rather than at the individual student level, and for biases due to conducting multiple significance tests. Studies that have been published without properly adjusting for these biases are retroactively adjusted as part of the review and synthesis process. This system also goes further than most programs-that-work lists in attempting to consider the full body of relevant program effectiveness study results in supporting a judgment of positive effects. Yet it still suffers from one of the fundamental validity and integrity threats that all such systems experience—the ease with which a program can meet the stated criteria of having positive effects based on chance findings alone. For example, consider a program that had been evaluated ten times with each evaluation testing ten outcomes, for a total of 100 tests of statistical significance. According to the What Works Clearinghouse rules, this program could qualify as a positive effects program (the systems highest rating) if just one outcome in each of two studies, one of which was judged to have a strong design, were found to be statistically significant—even with all other 98 tested outcomes not yielding statistically significant results (What Works Clearinghouse, p. 22).

Systematic Reviews and Meta-Analyses

The *systematic review* label is generally used to describe a quantitative review based on a standardized protocol intended to protect the process from bias. A *meta-analysis* is a type of systematic review employing quantitative procedures for averaging effect sizes across multiple studies. In meta-analyses of intervention studies, one commonly used effect size is the standardized mean difference between intervention and control groups (McCartney & Rosenthal, 2000). Systemic reviews are generally found in the same journals that publish critical narrative reviews, such as *Psychological Bulletin*, *Perspectives on Psychological Science*, and *School Psychology Review*. In addition, two international organizations sponsor, monitor, and maintain systematic reviews and meta-analyses of interventions in different areas—the Cochrane Collaboration for health-care interventions and the Campbell Collaboration for social interventions. Both collaborations include reviews related to adolescent health psychology interventions.

Quantitative systematic reviews of multiple randomized controlled trials are generally considered to occupy the top rung of the hierarchy of sources of effectiveness evidence. Yet, just as with individual research studies, inferences resulting from systematic reviews and meta-analyses are

subject to a variety of validity and integrity threats. Any of the threats experienced by individual studies, such as selection or attrition threats to internal validity, can be carried over to the systematic review. Additional threats are related to the nature of the research syntheses itself. These include publication bias and biased sampling of studies, biased selection of outcomes, lack of statistical independence among multiple effect sizes used, study rater biases and rating instability, and many others (Dwan et al., 2008; Hahn et al., 2000; Ioannidis & Karassa, 2010; Matt & Cook, 2009; Shadish et al., 2002). Briggs (2005) has gone so far as to argue that "researcher subjectivity is no less problematic in the context of a meta-analysis than in a narrative review" (p. 87).

Case Example: Interventions to Reduce Adolescent Sexual Risk Behaviors

An extensive body of research and research syntheses on the effectiveness of interventions to reduce adolescent sexual risk behavior provides for a compelling case example. Hundreds of individual studies exist, along with numerous narrative reviews, programs-that-work lists, and meta-analyses. This case example involves a brief methodological critique focusing on the most influential research syntheses in this area.

Scher, Maynard, and Stagner (2006) identified 14 descriptive narrative reviews of effectiveness studies of sexual risk behavior interventions for adolescents. Now in its third incarnation, *Emerging Answers* (Kirby, 2007) is the most extensive and influential of such reviews. One of its main components is a descriptive summary of risk and protective factors that purportedly affect teens' sexual behavior. The author concluded that more than 500 specific factors affect one or more adolescent sexual risk behaviors and their outcomes. A box score table was provided for the 71 factors deemed most important, based on a large collection of primarily observational studies reviewed, and employing statistical criteria such as three or more studies reporting statistically significant associations for the specific risk or protective factor (pp. 54–61). A fundamental limitation in this review was not sufficiently distinguishing between factors that are merely associated with and occur before the sexual behaviors (*risk or protective markers*) and those for which evidence of causality has been found, for example when a manipulation of the factor has been shown to contribute to a change in the outcome (*causal risk or protective factors*). Causal claims were made repeatedly, as in referring to this group of factors as "affecting teen sexual behavior and its outcomes" (p. 54) and "influential on teen's sexual behavior" (p. 63), and in arguing that each factor "exerted an effect," (p. 54). Yet evidence of causality over and above mere association was weak or completely absent for many or most of the factors listed, for example, hours of paid work and peer substance use (risk markers), and taking a virginity pledge and peer condom-use support (protective markers). This not uncommon failure to adequately distinguish between association and causation, referred to by Rosenthal (1994) as the problem of *causism*, has important negative implications for the development and evaluation of interventions. Kraemer and colleagues (Kraemer et al., 1997; Kraemer, Lowe, & Kupfer, 2005) provide in-depth discussions on the differentiation between risk markers and causal risk factors and the fundamental importance of recognizing these distinctions.

The primary focus of *Emerging Answers* was a review of the evidence of effectiveness across a large number of adolescent sexual behavior and other related outcomes for sexual risk behavior prevention interventions. Studies were selected for review based on criteria such as program goals and measured outcomes, and having follow-up data collected at least 3 months after intervention completion, as well as several vague methodological criteria:

> Include a reasonably strong experimental or quasi-experimental design, have reasonably well matched intervention and comparison groups, collect data both before and after implementation of the program, have a sample size of at least 100 persons [and] employ appropriate statistical analyses. (Kirby, 2007, p. 83)

No further specification was provided of the criteria used to identify studies that met the inclusion standards of "a reasonably strong" design or "appropriate statistical analysis." While several issues of design and statistical analysis were discussed, these were not used as a basis for exclusion of weaker studies. For example, the author disclosed that "almost one-third of the studies in this review are biased in favor of more significant results because they did not adjust statistically for clustering" (p. 93), and that "99 of the 115 studies conducted multiple tests of significance, but only seven studies adjusted for them" (p. 96), yet no remedial adjustments were made for these biases as part of this synthesis.

Emerging Answers concluded with a programs-that-work list of 15 programs characterized as having "strong evidence of positive impact on sexual behavior or pregnancy or STD rates" (Kirby, 2007, pp. 190–191). This list included seven curriculum-based interventions, four of which are published by the author's employer, ETR Associates. Many of the 15 listed programs are characterized by questionable evidence of effectiveness, involving such issues as unadjusted multiple significance testing, selective reporting, differential attrition, and failure to adjust for clustering (for example, see Constantine and Braverman's (2004) critique of the effectiveness evidence and its use for ETR Associate's *Reducing the Risk* curriculum). In addition, programs for which the preponderance of reported outcomes showed no statistically significant effects were nevertheless included on the list. In fact, of the seven curriculum-based interventions listed in *Emerging Answers* as having strong evidence of effectiveness, six were subsequently judged by the Coalition for Evidence-Based Policy (2010) as not having strong evidence of effectiveness for pregnancy or STD prevention (the seventh was not addressed by the Coalition).

Various other programs-that-work lists have been developed and promoted for adolescent sexual risk behavior interventions. Most recently, the federal Personal Responsibility Education Program requests for applications required that grantees "replicate evidence-based effective program models or substantially incorporate elements of effective programs that have been proven on the basis of rigorous scientific research to change [sexual] behavior" (US Department of Health and Human Services, Office of Adolescent Health, 2010). Partly but incompletely modeled after What Works Clearinghouse (2008) principles, this list of programs eligible for funding is more rigorous than the lists in this area typically developed by advocates and publishers. Nevertheless, it suffers from many of the same validity problems. For example, no adjustments were made by the reviewers for biases due to clustering or multiple significance testing in studies that had neglected to do so. Programs designated as evidence-based were initially classified into eight levels of evidence strength, but ultimately all were collapsed into one list of "evidence-based programs" preapproved for federal funding eligibility. Upon release of the request for applications employing this list, the independent Coalition for Evidence-Based Policy (2010) commented that "HHS's evidence-based teen pregnancy prevention program is an excellent first step, but only 2 of 28 approved models have strong evidence of effectiveness" (p. 1).

Systematic reviews and meta-analyses in this area have been almost as prolific as narrative reviews and programs-that-work lists. In their Campbell Collaboration review of interventions intended to reduce pregnancy-related outcomes among adolescents, Scher and colleagues (2006) identified six previously published meta-analyses on adolescent sexual risk behavior interventions. Subsequently, Oringanje et al. (2010) published a Cochrane Collaboration review on interventions for preventing unintended pregnancies among adolescents, and Johnson, Scott-Sheldon, Huedo-Medina, and Carey (2011) updated their original 2003 meta-analysis on adolescent HIV prevention interventions. These reviews vary in program type, intervention focus, eligible research designs, outcomes considered, and number of studies analyzed and in patterns of strengths and weaknesses exhibited. They also vary in findings and conclusions. Four of these reviews that are arguably the most influential, either because of their Cochrane Collaboration or Campbell Collaboration sponsorship or as evidenced by a relatively large number of citations are worth considering.

Oringanje et al.'s (2010) Cochrane Collaboration systematic review was based on analyses of 15 of 41 eligible randomized studies for which appropriate data were available. It reported no significant effects for any type of intervention on any type of sexual behavior or pregnancy-related outcome based on full sample analyses. Among the many subgroup analyses included, one significant effect was reported for "gender mixed or not specified" (p. 65) subgroups' initiation of sexual intercourse in interventions that combine education with contraception promotion. Inexplicably, one nonsignificant result that "approached significance" (p. 14) and was based on just two studies led the authors to erroneously conclude in their abstract, without qualification, that this type of combined education with contraception promotion "lowered the rate of unintended pregnancy among adolescents" (p. 2).

Scher and colleagues (2006), in their Campbell Collaboration review, analyzed 19 studies of school-based sex education programs. These were selected based on explicit criteria, such as employing a randomized design, reporting at least one of three prespecified outcomes (sexual experience, unprotected sexual activity, and pregnancy rates), and meeting defined sample retention standards. For sex education programs with an abstinence focus, the authors found "limited evidence" of a negative effect involving *higher* pregnancy rates among intervention groups (p. 3). For sex education programs with a comprehensive focus "no consistent evidence" was found that these programs "altered the likelihood that youth would initiate sex, would risk pregnancy, or would become [or get someone] pregnant" (p. 3). "Promising results" based on six randomized studies were reported for intensive multicomponent youth development programs serving higher risk adolescents (p. 3). The authors noted that these results did not show the programs to be ineffective, but rather, were most likely a reflection of the dearth of high quality research evidence available in this field.

In one highly cited systematic review, DiCenso, Guyatt, Willan, and Griffith (2002) analyzed 26 randomized studies of interventions developed to reduce unintended adolescent pregnancies and found no effects on initiation of sexual intercourse, use of birth control, or number of pregnancies. Methodological quality scores were calculated for each included study but used only descriptively to illustrate the poor methodological quality of most of the studies analyzed. Most recently, Johnson et al. (2011) updated their highly cited 2003 review, analyzing 67 adolescent HIV prevention intervention studies selected based on criteria that included a single methodological standard, "use [of] a randomized trial or a quasi-experimental design with rigorous controls" (p. 78), without further elaboration on how rigor was evaluated. Again, methodological quality scores were computed, but this was done subsequent to study inclusion and they were used only descriptively. In this review, intervention effects in the desired directions were found across the 67 studies for condom use, incidence of sexually transmitted infections, reducing or delaying sex, and negotiation skills.

Although hundreds of individual studies have been conducted, many of the synthesis authors have commented on the dearth of high quality studies. The largest and most rigorous individual study in this area to date was a 5-year randomized trial of the U.S. Title V, Section 510 Abstinence Education Program, which tested four of the most promising abstinence-only interventions. The results indicated no significant differences between individually randomized trial participants and control students on any of the primary outcomes (Trenholm et al., 2007, 2008). This study provides compelling evidence for the limited potential of abstinence-only education approaches, especially when considered in light of converging evidence derived from other studies based on complementary research designs, such as the previously discussed Kohler and colleagues' (2008) NSFG observational study. Yet, no study of the scope and rigor of this Trenholm and colleagues trial has ever been conducted of abstinence-plus interventions. Together with differences in the criteria used to select research studies for inclusion in a synthesis, this dearth of high quality research might help explain the wide disparity of conclusions across

the many narrative reviews, programs-that-work lists, and meta-analyses.

At the same time, these research syntheses as a group are characterized by neglect of some of the most serious and common threats to validity and integrity found in the individual studies analyzed, such as failure to adjust for clustering bias when intervention assignments are made at the group rather than individual level and failure to account for multiple significance testing biases. Additional threats introduced at the research-synthesis level in some reviews include the potential for conflict of interest due to close reviewer connections to the programs and studies reviewed, and insufficiently systematic and transparent criteria for study selection. And for the narrative reviews and programs-that-work lists, the same one study/one outcome criterion for effectiveness that has been widely criticized in other areas of intervention research continues to be perhaps the most serious threat of all.

A further weakness in this and many other areas of research synthesis and evidence-based policy has been insufficient attention to additional relevant and important sources of evidence, especially evidence from basic science research (Hirsch, 2002; Lochman, 2000; Westen & Bradley, 2005). This includes evidence from established and emerging programs of research in social, cognitive, developmental, and educational psychology and neuropsychology. For example, programs and curricula focused on prevention of adolescent sexual risk behaviors tend to view adolescents as rational, deliberative decision makers motivated to maximize positive outcomes. Yet basic research in developmental, cognitive, and social psychology has for some time demonstrated how judgment and decision making are much more complex in general (Gigerenzer & Selten, 2002; Schneider & Shanteau, 2003), and specifically regarding health behavior (Wiers, et al., 2010), adolescents (Jacobs & Klaczynski, 2005; Moshman, 2011; Reyna & Rivers, 2008), and adolescent sexual health behavior (Goldfarb & Constantine, 2011). Evidence from this type of basic science research is essential to appraising and understanding intervention effectiveness and its contexts and practices, yet it is routinely ignored.

This brief methodological critique of research syntheses on the effectiveness of interventions to reduce adolescent sexual risk behaviors illustrates several of the commonly encountered validity and integrity threats discussed in the chapter. It demonstrates how the etiological roots of some of the most insidious threats can be found in the quest for black and white answers within an area characterized by varied shades of gray. These roots are nurtured by the abundance of poor quality research in the field.

In spite of these weaknesses, several broad conclusions are supported by the full body of evidence in this area:

(a) Abstinence-only interventions as typically conceived and implemented have limited potential.
(b) Abstinence-plus interventions that directly focus on promoting behavioral change and include instruction on condoms and contraception methods have better potential, and evidence is accumulating of some modest positive effects overall.
(c) With few exceptions, the effectiveness of specific individual abstinence-plus interventions, programs, or curricula is not well supported by the available evidence. This does not necessarily mean that these programs are ineffective, just that the nature and quality of the available research is woefully inadequate to answering questions at this level.
(d) The full potential to enhance adolescents' sexual health and development through primary prevention and health promotion interventions is not being realized by the currently popular abstinence-only and abstinence-plus intervention models.

Concluding Comments and Recommendations

> Failed certainties in social science litter the landscape like so many elephant bones bleaching in the African sun. Honest hard scientists never claim final answers; good social science shouldn't either.
>
> (Carter, 2004)

> In considering the ways in which research data are typically interpreted, I became convinced that there is a strong cult of naive and overconfident empiricism in psychology and the social sciences with an excessive faith in data as the direct source of scientific truth and an inadequate appreciation of how misleading data can be. I concluded that the commonly held belief that research progress requires only that we "let the data speak" is sadly erroneous. If data are allowed to speak for themselves, they will typically lie to you.
>
> (Schmidt, 2010)

There exists an important need for interventions based in adolescent health psychology to reduce adolescent risk behaviors and to promote adolescent health and development more broadly. Appropriately, questions of intervention effectiveness have been and are continuing to receive steadfast attention. Funders, policy shapers, practitioners, and other stakeholders understandably seek direct and straightforward answers, especially to the one deceptively decisive question of highest perceived importance—does an intervention achieve its intended effect, yes or no? But reality rarely yields to such desired simplicity, nor does principled scientific inquiry enable it. Oversimplification of research, its appraisal, and its use in the service of this single yes or no question opens the door to unchecked threats to research validity and research integrity, neglect of valuable types of relevant evidence, and ultimately to misleading research conclusions, misinformed policy and funding decisions, and unfulfilled potential.

Consistent with the National Research Council (2002) principles of scientific inquiry and other modern views on the nature of science, there are a number of ways in which intervention effectiveness research and its use in adolescent health psychology could be improved to better support the development, evaluation, and dissemination of effective interventions. A good start would be to move beyond the widely embraced myth that method determines validity and its corollary fiction that methodological hierarchies and methodological quality checklists can substitute for genuine critical appraisal. This would be supported by a better understanding and acceptance of the need for and importance of qualitative reasoning (Brady & Collier, 2010; Freedman, 2008; Maxwell, 2005) and carefully reasoned argument (Abelson, 1995; Campbell, 1982; Victora, Habicht, & Bryce, 2004) in all types of research. As Abelson has noted, "the purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric" (1995, p. xiii). And Lancet editor Richard Horton's (1998) advice to physicians should resonate with adolescent health psychology researchers and practitioners as well: "The argument is the fundamental unit of all medical thought" (p. 249).

A principal tool for putting this into practice would be the critical narrative review, characterized by deeper and more meaningful attention to theory and mechanisms. This would embrace basic research evidence from relevant fields such as social, cognitive, developmental, and educational psychology and neuropsychology. It would focus on cumulative evidence and theoretical replications, together with theoretically expected convergence of evidence across multiple studies, research groups, methods, and contexts. And its essence would involve the spirited consideration of plausible alternative explanations for all results and potential conclusions, with attention to the implications of competing explanations on multiple data sets and to the fit between these implications and actual data (Campbell, 2009).

At a more fundamental level, research programs based on theory-driven model-building approaches have the potential to strengthen intervention development and evaluation. A model-building approach has been described as "iterative within a program of research, cycling though the following phases: theory, field observations, construct definition, measurement development, construct analysis, model testing, experimental field trials, and model revision" (Dishion & Patterson, 1999). Such an approach includes the probing of theory-based moderators and mediators to increase the understanding of relevant processes and mechanisms of change (Cook, 2002; Hinshaw, 2002; Kazdin, 1997; Kotchick, Shaffer, & Forehand, 2001; Lochman, 2006; Weersing & Weisz, 2002), consistent with the understanding that

> The "gold standard" studies in intervention research are those that not only demonstrate efficacy but also demonstrate that the postulated change mechanisms . . . do indeed carry the weight of improvement on (intervention) outcomes. (Hughes, 2000, p. 307)

Although programs-that-work lists have been characterized by what might appear to be insurmountable problems, the quest for such straightforward direction by funders and program administrators is not surprising, and these types of lists are unlikely to disappear anytime soon. One strategy to address this challenge would be the development of a new generation of evidence-based program lists that are grounded upon more genuinely scientific criteria of effectiveness. In place of the currently popular practice of trolling through individual research studies for any possible signs of effectiveness, this would involve critically appraising intervention content, approach, and intended populations for consistency with more inclusive theory-grounded evidence and principles derived from comprehensive and integrative critical narrative reviews. This process would not be easy, but like principled scientific inquiry more generally, principled scientific research synthesis rarely is easy.

Finally, recognizing that biases associated with conflict of interests are pervasive and generally not intentional or even within one's conscious awareness, a greater separation among program developers, researchers, and research reviewers is needed.

> We must move beyond mere disclosure of conflicts of interest toward developing additional regulatory mechanisms aimed at minimizing their pervasive influence. Like George Washington admitting that he chopped down his father's cherry tree, our willingness to disclose conflicts of interest does not absolve us of further responsibility. (Abi-Jaoude & Gorman, 2010, p. 1546)

The issues and strategies discussed in this chapter are intended to address the need for fundamental improvements in the conduct and use of research and research synthesis on intervention effectiveness in adolescent health psychology. Through better understanding and application of principled scientific inquiry, better attention to common threats to research validity and research integrity, and better use of theory, evidence, and reasoned argument, the field of adolescent health psychology should be able to make further progress toward reaching its full potential.

Acknowledgments The author thanks Eva Goldfarb, Petra Jerman, Wendy Constantine, and Jessica Lin for critical review and suggestions. Preparation of this chapter was facilitated by grants from the Ford Foundation and the William and Flora Hewlett Foundation.

References

Abel, U., & Koch, A. (1999). The role of randomization in clinical studies: Myths and beliefs. *Journal of Clinical Epidemiology, 52*, 487–497. doi:10.1016/S0895-4356(99)00041-4.

Abelson, R. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Abi-Jaoude, E., & Gorman, D. A. (2010). Disclosure: Only a first step. *Canadian Medical Association Journal, 182*, 1546. doi:10.1503/cmaj.110-2109.

Advocates for Youth. (2008). *Science and success: Sex education and other programs that work to prevent teen pregnancy, HIV & sexually transmitted infections* (2nd ed.). Retrieved June 6, 2011, from http://www.advocatesforyouth.org/storage/advfy/documents/sciencesuccess.pdf.

Altman, D. G. (2002). Poor quality medical research: What can journals do? *Journal of the American Medical Association, 287*, 2765–2767. doi:10.1001/jama.287.21.2765.

Anderson, N. H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Review, 70*, 162–179. doi:10.1037/h0044858.

Bachrach, C., & Newcomer, S. F. (2002). Addressing bias in intervention research: Summary of a workshop. *Journal of Adolescent Health, 31*, 311–321. doi:10.1016/S1054-139X(02)00395-6.

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

Bearman, P. S., & Bruckner, H. (2001). Promising the future: Virginity pledges and first intercourse. *American Journal of Sociology, 106*, 859–912. doi:10.1086/320295.

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. doi:10.1002/mpr.225.

Berliner, D. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher, 31*(8), 18–20. doi:10.3102/0013189X031008018.

Brady, H. E., & Collier, D. (Eds.). (2010). *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed.). Lanham, MD: Rowman and Littlefield.

Briggs, D. C. (2005). Meta-analysis: A case study. *Evaluation Review, 29*, 87–127. doi:10.1177/0193841X04272555.

Cain, D. M., & Detsky, A. S. (2008). Everyone's a little bit biased (even physicians). *Journal of the American Medical Association, 299*, 2893–2895. doi:10.1001/jama.299.24.2893.

Campbell, D. T. (1953). *Designs for social science experiments*. Evanston, IL: Northwestern University.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312. doi:10.1037/h0040950.

Campbell, D. T. (1982). Experiments as arguments. In E. R. House, S. Mathison, J. A. Pearsol, & H. Preskill (Eds.), *Evaluation studies review annual* (Vol. 7, pp. 117–127). Beverly Hills, CA: Sage.

Campbell, D. T. (2009). Forward. In R. K. Yin (Ed.), *Case study research: Design and methods* (4th ed., pp. vii–viii). Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton-Mifflin.

Carter, H. (2004). Presidential perspective. In M. T. Braverman, N. A. Constantine, & J. K. Slater (Eds.), *Foundations and evaluation: Contexts and practices for effective philanthropy*. San Francisco: Jossey-Bass.

Centers for Disease Control and Prevention. (2008). *10 steps to promoting science-based approaches to teen pregnancy prevention using Getting To Outcomes: A summary*. Retrieved from http://www.cdc.gov/reproductivehealth/adolescentreprohealth/PDF/LittlePSBA-GTO.pdf.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 291*, 2457–2465. doi:10.1001/jama.291.20.2457.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*, 323–393. Retrieved from http://www.jstor.org/stable/3233918.

Chugh, D., Banaji, M., & Bazerman, M. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In D. Moore, D. Cain, G. Loewenstein, & M. Bazerman (Eds.), *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy*. New York: Cambridge University Press.

Coalition for Evidence-Based Policy. (2010). *HHS's evidence-based teen pregnancy prevention program*. Retrieved June 6, 2011, from http://coalition4evidence.org/wordpress/wp-content/uploads/Coalition-comments-HHS-Teen-Pregnancy-Prevention-May-2010.pdf.

Collins, J., Robin, L., Wooley, S., Fenley, D., Hunt, P., Taylor, J., et al. (2002). Programs-that-work: CDC's guide to effective programs that reduce health-risk behavior of youth. *Journal of School Health, 72*, 93–99. doi:10.1111/j.1746-1561.2002.tb06523.x.

Constantine, N. A. (2008a). Converging evidence leaves policy behind: Effectiveness of and support for school-based sex education programs [Editorial]. *Journal of Adolescent Health, 42*, 324–326. doi:10.1016/j.jadohealth.2008.01.004.

Constantine, N. A. (2008b). The peer review process. In S. Boslaugh (Ed.), *Encyclopedia of epidemiology* (Vol. 1, pp. 794–796). Thousand Oaks, CA: Sage.

Constantine, N. A. (2008c). Publication bias. In S. Boslaugh (Ed.), *Encyclopedia of epidemiology* (Vol. 1, pp. 853–854). Thousand Oaks, CA: Sage.

Constantine, N. A. (2012). Regression analysis and causal inference: Cause for concern? *Perspectives on Sexual and Reproductive Health, 44*, 134–137. doi:10.1363/4413412.

Constantine, N. A., & Braverman, M. T. (2004). Appraising evidence on program effectiveness. In M. T. Braverman, N. A. Constantine, & J. K. Slater (Eds.), *Foundations and evaluation: Contexts and practices for effective philanthropy* (pp. 236–258). San Francisco: Jossey-Bass.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*, 175–199. doi:10.3102/01623737024003175.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Dane, A. U., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23–45. doi:10.1016/S0272-7358(97)00043-3.

Davey Smith, G., & Ebrahim, S. (2001). Epidemiology—is it time to call it a day? *International Journal of Epidemiology, 30*, 1–11. doi:10.1093/ije/30.1.1.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin, 28*, 1379–1387. doi:10.1177/014616702236869.

DiCenso, A., Guyatt, G., Willan, A., & Griffith, L. (2002). Interventions to reduce unintended pregnancies among adolescents: Systematic review of randomized controlled trials. *British Medical Journal, 324*, 1426–1434. doi:10.1136/bmj.324.7351.1426.

Dishion, T. J., & Patterson, G. R. (1999). Model building in developmental psychopathology: A pragmatic approach to understanding and intervention. *Journal of Clinical Child Psychology, 28*, 502–512. doi:10.1207/S15374424JCCP2804_10.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *Public Library of Science (PLoS) ONE, 3*, e3081. doi:10.1371/journal.pone.0003081.

Egger, M., & Davey Smith, G. (1997). Meta-analysis: Potentials and promise. *British Medical Journal, 315*, 1371–1374.

Evans, J. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum.

Feinstein, A. R. (1988). Fraud, distortion, delusion, and consensus: The problems of human and natural deception in epidemiologic science. *American Journal of Medicine, 84*(3, Pt. 1), 475–478. doi:10.1016/0002-9343(88)90268-9.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4–14. doi:10.3102/0013189X031008004.

Freedman, D. A. (2008). On types of scientific inquiry: The role of qualitative reasoning. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 300–318). New York: Oxford University Press.

Freedman, D. A. (2010). *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge, UK: Cambridge University Press.

Friedlander, F. (1964). Type I and Type II bias. *American Psychologist, 19*, 198–199. doi:10.1037/h0038977.

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. doi:10.1177/0193841X06287188.

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

Gigerenzer, G. (1998). Surrogates for theories. *Theories & Psychology, 8*, 195–204. doi:10.1177/0959354398082006.

Gigerenzer, G. (2009). Surrogates for theory. *Association for Psychological Science Observer, 22*, 21–23.

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: The MIT Press.

Goldfarb, E. S., & Constantine, N. A. (2011). Sexuality education. In B. B. Brown & M. Prinstein (Eds.), *Encyclopedia of adolescence*. New York: Academic.

Gorman, D. M. (2002). Defining and operationalizing 'research-based' prevention: A critique (with case studies) of the US Department of Education's Safe, Disciplined and Drug-Free Schools exemplary programs. *Evaluation and Program Planning, 25*, 295–302.

Gorman, D. M., & Conde, E. (2007). Conflict of interest in the evaluation and dissemination of "model" school-based drug and violence prevention programs. *Evaluation and Program Planning, 30*, 422–429. doi:10.1016/j.evalprogplan.2007.06.004.

Grossman, D. C., Neckerman, H. J., Koespell, T. D., Liu, P. Y., Asher, K. N., Beland, K., et al. (1997). Effectiveness of a violence prevention curriculum among children in elementary school: A randomized controlled trial. *Journal of the American Medical Association, 277*, 1605–1611. doi:10.1001/jama.1997.03540440039030.

Haack, S. (2003). *Defending science—within reason: Between scientism and cynicism*. New York: Prometheus.

Hahn, S., Williamson, P. R., & Hutton, J. L. (2002). Investigation of within-study selective reporting in clinical research: Follow-up of applications submitted to a local research ethics committee. *Journal of Evaluation in Clinical Practice, 8*, 353–359. doi:10.1046/j.1365-2753.2002.00314.x.

Hahn, S., Williamson, P. R., Hutton, J. L., Garner, P., & Flynn, V. (2000). Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in Medicine, 19*, 3325–3336. doi:10.1002/1097-0258(20001230)19:24<3325::AID-SIM827>3.0.CO;2-D.

Hinshaw, S. P. (2002). Intervention research, theoretical mechanisms, and causal processes related to externalizing behavior patterns. *Development and Psychopathology, 14*, 789–818. doi:10.1017/S0954579402004078.

Hirsch, E. D. (2002). Classroom research and cargo cults. *Policy Review, 115*, 51–69.

Horton, R. (1998). The grammar of interpretive medicine. *Canadian Medical Association Journal, 159*, 245–249.

Howel, D., & Bhopal, R. (1994). Assessing cause and effect from trials: A cautionary note. *Controlled Clinical Trials, 15*, 331–334. doi:10.1016/0197-2456(94)90030-2.

Hughes, J. N. (2000). The essential role of theory in the science of treating children: Beyond empirically supported treatments. *Journal of School Psychology, 38*, 301–330. doi:10.1016/S0022-4405(00)00042-X.

Huston, P., & Moher, D. (1996). Redundancy, disaggregation, and the integrity of medical research. *The Lancet, 347*, 1024–1026. doi:10.1016/S0140-6736(96)90153-1.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Library of Science (PLoS) Medicine, 2*, 696–701. doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology, 60*, 324–329. doi:10.1016/j.jclinepi.2006.09.011.

Ioannidis, J. P. A. (2011). An epidemic of false claims: Competition and conflicts of interest distort too many medical findings. *Scientific American, 304*, 16.

Retrieved from http://www.scientificamerican.com/article.cfm?id=an-epidemic-of-false-claims.

Ioannidis, J. P. A., & Karassa, F. B. (2010). The need to consider the wider agenda in systematic reviews and meta-analyses: Breadth, timing, and depth of the evidence. *British Medical Journal, 341*, c4875. doi:10.1136/bmj.c4875.

Ioannidis, J. P. A., Tatsioni, A., & Karassa, F. B. (2010). Who is afraid of reviewers' comments? Or, why anything can be published and anything can be cited. *European Journal of Clinical Investigation, 40*, 285–287. doi:10.1111/j.1365-2362.2010.02272.x.

Jacobs, J. J., & Klaczynski, P. A. (2005). *The development of judgment and decision making in children and adolescents*. Mahwah, NJ: LEA.

Johnson, B. T., Scott-Sheldon, L. A. J., Huedo-Medina, T. B., & Carey, M. P. (2011). Interventions to reduce sexual risk for human immunodeficiency virus in adolescents: A meta-analysis of trials, 1985–2008. *Archives of Pediatrics & Adolescent Medicine, 165*, 77–84. doi:10.1001/archpediatrics.2010.251.

Kassirer, J. P., & Angell, M. (1995). Redundant publication: A reminder [Editorial]. *New England Journal of Medicine, 333*, 449–450. Retrieved from http://www.nejm.org/doi/pdf/10.1056/NEJM199508173330709.

Kazdin, A. E. (1997). A model for developing effective treatments: Progression and interplay of theory, research, and practice. *Journal of Clinical Child Psychology, 26*, 114–129.

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

Kirby, D. (2007). *Emerging answers 2007: Research findings on programs to reduce teen pregnancy and sexually transmitted diseases*. Washington, DC: National Campaign to Prevent Teen and Unplanned Pregnancy.

Kohler, P. K., Manhart, L. E., & Lafferty, W. E. (2008). Abstinence-only and comprehensive sex education and the initiation of sexual activity and teen pregnancy. *Journal of Adolescent Health, 42*, 344–351. doi:10.1016/j.jadohealth.2007.08.026.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press/Bradford Books.

Koslowski, B., & Thompson, S. (2002). Theorizing is important, and collateral information constrains how well it is done. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 171–192). Cambridge, UK: Cambridge University Press.

Kotchick, B. A., Shaffer, A., & Forehand, R. (2001). Adolescent sexual risk behavior: A multi-system perspective. *Clinical Psychology Review, 21*, 493–519. doi:10.1016/S0272-7358(99)00070-7.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to terms with the terms of risk. *Archives of General Psychiatry, 54*, 337–343.

Kraemer, H. C., Lowe, K. K., & Kupfer, D. J. (2005). *To your health: How to understand what research tells us about risk*. Oxford, UK: Oxford University Press.

Krisberg, K. (2010). Teen pregnancy prevention focusing on evidence: Ineffective abstinence-only lessons being replaced with science. *The Nation's Health, 40*, 1–14. Retrieved from http://thenationshealth.aphapublications.org/content/40/3/1.1.full.

Kumar, M. N. (2008). A review of the types of scientific misconduct in biomedical research. *Journal of Academic Ethics, 6*, 211–228. doi:10.1007/s10805-008-9068-6.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498. doi:10.1037/0033-2909.108.3.480.

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An underrecognized confound in intervention research. *Psychological Bulletin, 130*, 289–303. doi:10.1037/0033-2909.130.2.289.

Levy, D. A. (2010). *Tools of critical thinking: Metathoughts for psychology* (2nd ed.). Long Grove, IL: Waveland.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.

Lipsey, M. W. (2004). *The 2004 Claremont debate: Lipsey vs. Scriven. Determining causality in program evaluation & applied research: Should experimental evidence be the gold standard?* Retrieved October 31, 2010, from http://www.cgu.edu/include/SBOS_2004_Debate.pdf.

Lochman, J. E. (2000). Theory and empiricism in intervention research: A dialectic to be avoided. *Journal of School Psychology, 38*, 359–338. doi:10.1016/S0022-4405(00)00038-8.

Lochman, J. E. (2006). Translation of research into interventions. *International Journal of Behavioral Development, 30*, 31–38. doi:10.1177/0165025406059971.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–305.

MacCoun, R. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology, 49*, 259–287. doi:10.1146/annurev.psych.49.1.259.

MacCoun, R. (2005). Conflicts of interest in public policy research. In D. A. Moore, D. M. Cain, G. Lowenstein, & M. H. Bazerman (Eds.), *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy* (pp. 233–262). Cambridge, UK: Cambridge University Press.

Machi, L. A., & McEvoy, B. T. (2009). *The literature review*. Thousand Oaks, CA: Corwin.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer-review system. *Cognitive Therapy and Research, 1*, 161–175. doi:10.1007/BF01173636.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537–556). New York: Russell Sage Foundation.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11. doi:10.3102/0013189X033002003.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Thousand Oaks, CA: Sage.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development, 71*, 173–180. doi:10.1111/1467-8624.00131.

Meinert, C. L. (1986). *Clinical trials: Design, conduct, and analysis*. Oxford, UK: Oxford University Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Mills, J. L. (1993). Data torturing. *New England Journal of Medicine, 329*, 1196–1199.

Moore, D. A., Loewenstein, G., Tanlu, L., & Bazerman, M. H. (2003). *Auditor independence, conflict of interest, and the unconscious intrusion of bias*. Harvard Business School Working Paper #03-116. Retrieved October 7, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.2829&rep=rep1&type=pdf.

Moshman, D. (2011). *Adolescent rationality and development: Cognition, morality, and identity*. New York, NY: Psychology.

Moskowitz, J. M. (1993). Why reports of outcome evaluations are often biased or uninterpretable: Examples from evaluations of drug abuse prevention programs. *Evaluation and Program Planning, 16*, 1–9. doi:10.1016/0149-7189(93)90032-4.

National Research Council Committee on Assessing Integrity in Research Environments. (2002). *Integrity in scientific research: Creating an environment that promotes responsible conduct*. Washington, DC: National Academies Press.

National Research Council Committee on Scientific Principles for Educational Research. (2002). *Scientific research in education*. Washington, DC: National Academies Press.

Oringanje, C., Meremikwu, M. M., Eko, H., Esu, E., Meremikwu, A., & Ehiri, J. E. (2010). Interventions for preventing unintended pregnancies among adolescents. *Cochrane Database of Systematic Reviews, 4*. doi:10.1002/14651858.CD005215.pub2.

Petrosino, A. (2003). Standards for evidence and evidence for standards: The case of school-based drug prevention. *Annals of the American Academy of Political and Social Science, 587*, 180–207. doi:10.1177/0002716203251218.

Rennie, D. (1998). Freedom and responsibility in medical publication: Setting the balance right. *Journal of the American Medical Association, 280*, 300–302. doi:10.1001/jama.280.3.300.

Rennie, D. (1999). Fair conduct and fair reporting of clinical trials. *Journal of the American Medical Association, 282*, 1766–1768. doi:10.1001/jama.282.18.1766.

Reyna, V. F., & Rivers, S. E. (2008). Current theories of risk and rational decision making. *Developmental Review, 28*, 1–11. doi:10.1016/j.dr.2008.01.002.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5*, 127–134. doi:10.1111/j.1467-9280.1994.tb00646.x.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?". *The Lancet, 365*, 82–93. doi:10.1016/S0140-6736(04)17670-8.

Scher, S., Lin, J., & Constantine, N. A. (2009). *Motivated translation of ambiguous scientific research findings: A case study from the sex education debates. Paper presented at the International Conference on Science in Society*. United Kingdom: University of Cambridge.

Scher, L. S., Maynard, R. A., & Stagner, M. (2006). Interventions intended to reduce pregnancy- related outcomes among adolescents. *Campbell Collaboration Systematic Reviews, 12*. Campbell Collaboration. doi: 10.4073/csr.2006.12.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*, 233–242. doi:10.1177/1745691610369339.

Schneider, S. L., & Shanteau, J. (2003). *Emerging perspectives on judgment and decision making*. Cambridge, UK: Cambridge University Press.

Scientifically Based Evaluation Methods, 68 Fed. Reg. 62,445 (October 29, 2003).

Scriven, M. (2008). A summative evaluation of RCT methodology and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5*, 11–24.

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

Shafer, G., & Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Science, 9*, 309–339. doi:10.1207/s15516709cog0903_2.

Shatz, D. (2001). *Peer review: A critical inquiry*. New York: Rowman and Littlefield.

Shavelson, R., & Towne, L. (2004). What drives scientific research in education? Questions, not methods, should drive the enterprise. *American Psychological Society Observer, 17*(4), 27–30. Retrieved from http://www.psychologicalscience.org/observer/getArticle.cfm?id=1557.

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. doi:10.1177/0956797611417632.

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology, 48*, 9–18. doi:10.1016/0895-4356(94)00097-A.

Smith, R., Feachem, R., Feachem, N. S., Koehlmoos, T. P., & Kinlaw, H. (2009). The fallacy of impartiality: Competing interest bias in academic publications. *Journal of the Royal Society of Medicine, 102*(2), 44–45. doi:10.1258/jrsm.2009.080400.

Suellentrop, K. (2010). *Effective and promising teen pregnancy prevention programs for Latino youth. Science Says, 43*. Retrieved June 6, 2011, from http://www.thenationalcampaign.org/resources/pdf/SS/SS43_TPPProgramsLatinos.pdf.

Trenholm, C., Devaney, B., Fortson, K., Clark, M., Quay, L., & Wheeler, J. (2008). Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. *Journal of Policy Analysis and Management, 27*, 255–276. doi:10.1002/pam.20324.

Trenholm, C., Devaney, B., Fortson, K., Quay, L., Wheeler, J., & Clark, M. (2007). *Impacts of four Title V Section 510 abstinence education programs: Final report*. Retrieved January 12, 2011, from http://www.mathematica-mpr.com/publications/pdfs/impactabstinence.pdf.

Trochim, M. K., & Donnelly, J. P. (2007). *The research methods knowledge base* (3rd ed.). Mason, OH: Thomson.

U.S. Department of Health and Human Services, Office of Adolescent Health. (2010). *Overview of the teen pregnancy prevention research evidence review.* Retrieved May 16, 2011, from http://www.hhs.gov/ash/oah/prevention/research/index.html.

Victora, C. G., Habicht, J., & Bryce, J. (2004). Evidence-based public health: Moving beyond randomized trials. *American Journal of Public Health, 94*, 400–405.

Weersing, V. R., & Weisz, J. R. (2002). Mechanisms of action in youth psychotherapy. *Journal of Child Psychology and Psychiatry, 43*, 3–29. doi:10.1111/1469-7610.00002.

Weiss, C. H. (1980). *Social science research and decision-making*. New York: Columbia University Press.

Weiss, C. H., Murphy-Graham, E., Petrosino, A., & Gandhi, A. G. (2008). The fairy godmother—and her warts: Making the dream of evidence-based policy come true. *American Journal of Evaluation, 29*, 29–47. doi:10.1177/1098214007313742.

West, S. G. (2009). Alternatives to randomized experiments. *Current Directions in Psychological Science, 18*, 299–304. doi:10.1111/j.1467-8721.2009.01656.x.

Westen, D., & Bradley, R. (2005). Empirically supported complexity: Rethinking evidence-based practice in psychotherapy. *Current Directions in Psychological Science, 14*, 266–271. doi:10.1111/j.0963-7214.2005.00378.x.

What Works Clearinghouse. (2008). *Procedures and standards handbook (Version 2.0)*. Retrieved June 6, 2011, from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *Public Library of Science (PLoS) ONE, 6*, e26828. doi:10.1371/journal.pone.0026828.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726–728.

Wiers, R. W., Houben, K., Roefs, A., de Jong, P., Hofmann, W., & Stacy, A. W. (2010). Implicit cognition in health psychology: why common sense goes out of the window. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition* (pp. 463–488). New York: Guilford.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594.

Williams, P. G., Holmbeck, G. M., & Greenley, R. N. (2002). Adolescent health psychology. *Journal of Consulting and Clinical Psychology, 70*, 828–842. doi:10.1037/0022-006X.70.3.828.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist, 17*, 657–658. doi:10.1037/h0038819.

World Health Organization. (1946). *Constitution of the World Health Organization*. Retrieved July 28, 2011, from http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf.

Yin, R. K. (2008). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage.

Young, S. N. (2009). Bias in the research literature and conflict of interest: An issue for publishers, editors, reviewers and authors, and it is not just about the money [Editorial]. *Journal of Psychiatry & Neuroscience, 34*, 412–417.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydi, O. (2009). Why current publication practices may distort science. *Public Library of Science (PLoS) Medicine, 5*, 1418–1422. doi:10.1371/journal.pmed.0050201.