Published in final edited form as:

Eval Program Plann. 2015 February ; 48: 100–116. doi:10.1016/j.evalprogplan.2014.09.007.

COMPARING RATING PARADIGMS FOR EVIDENCE-BASED PROGRAM REGISTERS IN BEHAVIORAL HEALTH: EVIDENTIARY CRITERIA AND IMPLICATIONS FOR ASSESSING PROGRAMS

Stephanie N. Means^a [Project Manager], Stephen Magura^b [Director], Jason T. Burkhardt^c [Project Manager], Daniela C. Schröter^d [Director of Research], and Chris L.S. Coryn^e [Director]

^aThe Evaluation Center at Western Michigan University, 1903 W Michigan Ave Kalamazoo, MI 49007 stephanie.means@wmich.edu

^bThe Evaluation Center at Western Michigan University, 1903 W Michigan Ave Kalamazoo, MI 49007 stephen.magura@wmich.edu

^cThe Evaluation Center at Western Michigan University, 1903 W Michigan Ave Kalamazoo, MI 49007

^dThe Evaluation Center at Western Michigan University, 1903 W Michigan Ave Kalamazoo, MI 49007 daniela.schroeter@wmich.edu

^eInterdisciplinary Ph.D. in Evaluation, Western Michigan University, 1903 W Michigan Ave Kalamazoo, MI 49007 chris.coryn@wmich.edu

Abstract

Decision makers need timely and credible information about the effectiveness of behavioral health interventions. Online evidence-based program registers (EBPRs) have been developed to address this need. However, the methods by which these registers determine programs and practices as being “evidence-based” has not been investigated in detail. This paper examines the evidentiary criteria EBPRs use to rate programs and the implications for how different registers rate the same programs. Although the registers tend to employ a standard Campbellian hierarchy of evidence to assess evaluation results, there is also considerable disagreement among the registers about what constitutes an adequate research design and sufficient data for designating a program as evidence-based. Additionally, differences exist in how registers report findings of “no effect,” which may deprive users of important information. Of all programs on the 15 registers that rate individual programs, 79% appear on only one register. Among a random sample of 100 programs rated by more than one register, 42% were inconsistently rated by the multiple registers to some degree.

© 2014 Elsevier Ltd. All rights reserved.

Corresponding Author 1903 W Michigan Ave Kalamazoo, MI 49007 jason.t.burkhardt@wmich.edu 269-387-5913.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Evidence-based programs; evidence-based practices; best practices; evidence-based program registers; standards of evidence; hierarchies of evidence; treatment effectiveness; behavioral health

Introduction

Since the 1960s, professional evaluation has aimed to apply scientific research methods to develop evidence-based practices or programs (EBPs). The term “evidence-based” gained traction in a number of primary disciplines ranging from healthcare to education to law enforcement. Most recently, federal, state, and local governments have begun to increase their mandates for the use of EBPs (Minnesota Department of Corrections, 2011; Reickmann, Kovas, Cassidy, & McCarty, 2011; Hawai'i State Center for Nursing, 2013; United States Office of Management and Budget, 2013; Office of the President of the United States, 2014a&b). Such mandates may require recipients of funding to spend a stated portion of their funds on EBPs (Office of Adolescent Health/Mathematica Policy Research, 2014). A policy memorandum from the United States Office of Management and Budget (OMB) directs federal agencies to use credible evidence in the formulation of budget proposals and performance plans. OMB further encourages funding of programs that are backed by strong evidence of effectiveness (United States Office of Management & Budget, 2013). However, exactly what constitutes credible evidence continues to be debated (Sackett et al., 1997; Donaldson, Christie, & Mark, 2009).

Discussions surrounding credible evidence have largely focused on the philosophical and ideological disagreements concerning taxonomies and hierarchies of methodological quality and rigor. Relatively well established hierarchies describing standards of evidence do exist, ranging from highly rigorous systematic reviews to randomized controlled trials to single-case designs. The preferred designs in those hierarchies are those that are likely to yield evidence that is subject to the least amount of bias (Rossi, Lipsey, & Freeman, 2004; Shadish, Cook, & Campbell, 2002). However, evidence that is derived outside of the sphere of experiments is often accepted as credible as well (Donaldson, Christie, & Mark, 2009). Furthermore, Gambrill points out that what constitutes a program as being evidence-based must not be limited to quality of evidence (i.e. what method was used), but must also include factors such as ethics and impartiality (2006). The evaluation field has come to recognize that the context of evidence is also important. Information beyond manifest outcomes must be considered when making decisions about program implementation (Shadish, Cook, & Leviton, 1991; Donaldson Christie and Mark, 2009; Scriven, 2014). Although considering multiple forms and sources of information of credible evidence sounds ideal in theory, it may cause problems for decision makers who need concrete and unambiguous rules to identify EBPs.

It is a daunting process for individuals and even organizations to access, review, and synthesize the large and rapidly growing body of research literature in many of the primary disciplines (Bastian, Glasziou, & Chalmers, 2010; Jette, et al., 2003; Shadish, Cook, & Leviton, 1991). Policymakers, administrators, clinicians, and practitioners need assistance in

efficiently collecting, aggregating, and interpreting what constitutes valid evidence of a program's effectiveness. Consumers need resources that will help them filter information in order to make sound decisions for program participation. Evidence-based program registers (EBPRs) are a relatively recent mechanism for assisting this in this process in the behavioral health field (United States Government Accountability Office, 2009). These registers were established to assess applied research and evaluation studies of programs/interventions according to evidentiary (evidence-based) standards, in order to help potential users decide which programs/interventions to support or select for implementation.

Several recent reviews of EBPRs, primarily focused on federally sponsored registers - such as the National Register of Evidence-Based Programs and Practices, the What Works Clearinghouse, and Social Programs that Work - found needs for more transparency in how such registers assess program effectiveness, in their use of evidence-based standards, and in how information is disseminated to the public (United States Government Accountability Office, 2009 & 2010; Hennessy & Green-Hennessy, 2011).

At the inception of the present study, it was unknown whether these problems were remedied. The authors' preliminary examination of relevant EBPRs found many different paradigms for rating programs. In particular, some registers listed programs as "evidence-based" (or not) with no further gradation in ratings, while others employed graded levels ("tiers") of evidence in support of a program. Among the latter type of register, it appeared that the top two tiers of evidence were indicative of what the field has termed "evidence-based," although the distinctions between the top two tiers varied among registers. In general there seemed to be many and often not obvious differences among registers in designating programs as evidence-based.

The purpose of this study is to describe the similarities and differences in evidentiary criteria of effectiveness and program rating schemes for EBPRs in behavioral health, with special attention to variation in evidentiary criteria between the top two tiers for registers with multiple rating tiers. This study also attempts to determine the extent to which such differences in evidentiary criteria affect ratings of the same program appearing in more than one register.

Methods

The present study was phase 2 of a sequential mixed methods study. In phase 1, the authors identified 20 active evidence-based practice registers that included interventions pertinent to behavior health. This was followed by analysis of the websites of those registers to identify their scope, purpose, key structural elements, funding sources, marketing and dissemination strategies, and challenges associated with maintaining them. Interviews with the register managers were also conducted in order to confirm the interpretation of documents and to provide the opportunity for the register managers to provide additional information (Burkhardt et al., this issue). The present second paper focuses on the variations in evidentiary criteria used by the registers, the ways in which those criteria are applied, and the implications of applying different criteria to rating the same programs.

Sample

The population of interest for the study was EBPRs in behavioral health. Behavioral health is defined as being “An umbrella term for care that addresses behavioral problems bearing on health, including patient activation and health behaviors, mental health conditions and substance use, and other behaviors that bear on health” (Peek et al, 2013).

Evidence-based behavioral health interventions are interventions whose effectiveness and appropriateness for implementation are supported by empirical data derived from systematic scientific inquiry (Agency for Healthcare Research and Quality, 2013; APA Presidential Task Force, 2006; Chambless & Hollon, 1998; Council for Training in Evidence-Based Behavioral Practice, 2008; Mattox & Kilburn, 2013; National Institute of Health - US Department of Health and Human Services, 2011; Southam-Gerow & Prinstein, 2014). EBPRs for behavioral health were defined in this study as: (1) online and actively managed for updates, (2) including behavioral health interventions, (3) including documentable criteria for including and excluding programs or interventions, and (4) presenting evaluative information that could support decision making. Online actively-managed registers were chosen for investigation because the authors believe that print listings or printed articles of evaluation results or syntheses no longer constitute the kind of timely information needed by decision makers, and that decision makers would likely primarily access web-based resources in preference to printed materials, which quickly become out of date. We did find several examples of printed evidence-based program lists that had been placed on-line, but these were not being updated or otherwise actively managed, and were thus excluded from study eligibility (U.S. Department of Education, 2001).

The authors began with a set of reports, resource lists, linking sites, practice manuals, and evidence-based program registers as defined by this study, conducted a current search of the web, and developed a final pool of 129 candidate EBPRs. The study inclusion/exclusion criteria listed above were applied, reducing the list to the final candidates. EBPRs that focused exclusively on medical treatment, physical health or general education programming (e.g., the Best Evidence Encyclopedia, Johns Hopkins University Center for Data Driven Reforms in Education, 2013) were then removed from the study, as were registers that solely relied on another register for their content. One example of the latter is FindYouthInfo.gov, which states that it uses a common database with Crimesolutions.gov. Another example is the OJJDP Model Programs Guide, which now shares a common dataset with Crimesolutions.gov. Additionally, although the large majority of the interventions reviewed by the Cochrane Collaboration are medically-related, many do address behavioral health, and given the Cochrane Collaboration's position in the evidence-based practice discussion, the Cochrane Collaboration was included in the study. The authors believe that the present study includes all existing EBPRs which include behavioral health interventions within their scope.

Document review

The registers included in the study were reviewed and coded according to a scheme that was iteratively developed. The major coding dimensions of relevance to this paper were: the degree of overlap of register criteria with the Cochrane Collaboration Bias Reduction

Standards (defined below), the types of evaluative designs the registers considered when including or ranking programs, the types of outcomes required for inclusion or ranking, and additional factors which influenced ratings of the programs or modalities. Additionally, a listing of the programs included in each register was compiled for use in later comparisons.

Interviews

For each of the 20 registers included in the study, attempts were made to contact a representative of the register to schedule an interview. Seventeen out of the 20 registers (85%) participated in interviews. Three registers did not complete the scheduling process to conduct an interview. Interview results pertaining to the evidentiary criteria used by the registers to rate programs and interventions are presented in this paper.

Sampling of programs

In order to study the implications of variations in criteria used by the registers, the authors compiled a list of programs contained in the 15 registers that rated individual programs. This list included 1,654 unique programs, of which 1,299 (78.5%) appeared in only one register while 355 (21.5%) appeared in more than one register. From this list, two samples were selected. First, a sample of 100 programs was randomly selected from the 355 programs that appeared in more than one register, in order to investigate how different registers rate the same programs. A second random sample of 30 programs that appeared in only one register was selected in order to determine whether such programs were eligible for consideration by other registers, but were missed or perhaps excluded for some reason.

Results

Twenty EBPRs are included in the content analysis of evidentiary criteria (Table 1). Twelve of the registers rate only individual treatment, prevention or service programs (“programs”), five registers rate only broad modalities or types of intervention (“modalities”), and three rate both individual programs and modalities (“mixed”). Four of the registers that rate modalities do so using meta-analytic or systematic review procedures. One register of modalities uses a rating scheme similar to the ones used by the registers of individual programs.

The evidentiary criteria used by each the twenty EBPRs for rating programs or modalities are summarized in the Appendix.

Description of evidentiary rating schemas

Registers either classified programs by (a) simple inclusion/exclusion from the register or (b) ranking into tiers of evidence; i.e., degrees of supportive evidence. For comparison purposes, inclusion/exclusion registers are considered to be single-tiered, and that tier also is considered to be the “top tier” for the register. As shown in Table 1, 10 (50%) of the registers use multiple positive tiers of evidence, while the remaining 10 (50%) are single-tiered. Among those with multiple tiers, five have two tiers and five have three or more tiers, with a maximum of five. The inclusion/exclusion schema uses criteria which if met, certify the program or modality as being evidence-based. One such example is Evidence-based

Practices in Substance Abuse (University of Washington Alcohol and Drug Abuse Institute, 2013), which applies specific criteria to decide whether a program is deemed effective or not (See Appendix).

Registers that use a ranking system employ preliminary inclusion criteria (e.g., program must be studied using a randomized controlled trial [RCT] design), and then apply a system of ranking criteria to determine a “top-tier” and lower groups. These multi-tiered registers each contain different types of tiers such as positive tiers, null tiers, negative tiers, and insufficient evidence/not tested tiers. A positive tier is defined here as a specific level of positive scientific support for the effectiveness of a program, null tiers include programs that were tested but found to have no effect, negative tiers include programs that have negative or harmful effects, and insufficient evidence/not tested tiers include programs that have not been sufficiently assessed. (“No effects,” “insufficient evidence,” and negative evidence categories are not considered a “tier of positive evidence” in our method of classifying.)

CrimeSolutions is an example of a multi-tiered register; it ranks reviewed programs into four tiers of positive evidence - “effective-multiple studies”, “effective-one study”, “promising-multiple studies”, and “promising-one study.” It also includes a “no effects” category (See Appendix). Combined information on program outcomes (i.e., program “effectiveness”) and methodological quality is termed “strength of evidence” by the Cochrane Collaboration (Higgins and Green, 2008). This study found variation in how strength of evidence was defined by the registers. For example, HomVEE and TPP rated program outcomes and methodological quality separately. These registers first indicate whether a program does or does not meet DHHS criteria, then the registers indicate whether the studies supporting inclusion were of high or moderate methodological quality. NREPP has minimum criteria for program effectiveness required for inclusion, but rates the methodological quality of the evaluations and readiness for dissemination of the program separately. CrimeSolutions uses a procedure where a program's evaluations must meet certain minimal criteria of methodological quality in order to qualify for consideration, but the quality of individual evaluations may vary. One register (Washington) used a reputational standard related to designs; all designs are accepted, but require at least one peer-reviewed article or testimony of a reputable organization to support a program as being evidence-based. (See Appendix for these details.)

As shown in table 1, 13 registers list programs or modalities that have “no effects” or that have “insufficient evidence” to be considered evidence-based. Four of the registers that list individual programs only (“programs”) have a category for no effects/insufficient evidence and an additional four note whether a reviewed program has been shown to be ineffective in the overall summary for a program. Four of the eight registers that address modalities (“modalities” and “mixed”) are registers of meta-analyses that report null findings. Three of the modality registers have a no effect/insufficient evidence category. For each of the seven registers that lacked information about no effect or insufficient evidence, we inquired during the interview why the register lacked such a category. Typically, the register administrators stated that doing so would be exhaustive of resources, could potentially confuse the user, or that they preferred to limit information to only those programs with demonstrable effectiveness.

The strength of program outcomes was originally conceptualized as a combination of effect size of the program's impact and statistical significance (Cooper, Hedges, & Valentine, 2009; Higgins and Green, 2008). However, 19 of the 20 registers did not mention a specific effect size benchmark, but instead used terms such as “important effect” or “large effect.”¹ While the registers apparently do attach a numerical meaning to the terms, these were not explicitly described to the user. All of the multi-tiered registers required statistically significant effects (explicitly or implicitly $p < .05$ or the 95% confidence interval) in their top-tier requirements.

Nine of 20 registers require statistically significant effects for a program to be included; while seven do not require significance for inclusion. Four of the registers of modalities utilize statistical significance of effects in making determinations about program effectiveness, since they are based primarily on meta-analytic reviews. Programs that do not have significant effects would score into the “no effects” category and programs without enough research or enough high quality research would score into the “insufficient evidence” category.

Typically, the evaluation design portion of the strength of evidence focuses on experimental and quasi-experimental research designs that increase the likelihood of group equivalence. This is what has been termed control for “sequence generation” (Higgins and Green, 2008: 396). For the registers that assess individual programs ($n=16$), the evaluation designs required for inclusion in the highest tier (or the only tier) are randomized controlled trials only (RCT) ($n=3$), either RCTs or quasi-experimental designs (QED) ($n=6$, 30%), any design with comparison groups ($n=5$), or uncontrolled studies allowed ($n=2$). For the tiered registers ($n=10$), inclusion at the highest tier requires RCTs ($n=5$); RCTs and/or QEDs ($n=3$); or designs with features that align with QED principles (i.e. comparison groups, group equivalence, manipulation of treatment, etc.; Shadish, Cook, & Campbell, 2002), but that do not meet the full definition of standard QEDs ($n=2$). Designs less stringent than QEDs are excluded from all highest tiers.

Three of the registers of modalities are also registers that present the results of meta-analysis. In this case, the registers have protocols in place that guide the development of a systematic review, but in general, these registers themselves did not make design recommendations. One of the registers that also presents modalities, the Community guide, does note that they require that a study include a before and after concurrent comparison group if it is to be included in a review, but does not place a requirement on a specific design for decision making purposes, rather they classify that design instead as one piece of evidence in their final recommendation. Finally, the Effective Child therapy register allows non-experimental designs, but programs that are supported only by these designs would only score in the lower tiers of their rating system.

The study also determined whether the registers, in addition to requirements for evaluation design, utilized additional criteria for assessing the methodological quality of evaluations.

¹CEBC uses a specific effect size and the WWC includes effect sizes in its improvement indexes, but the WWC's use is more complex than a simple cutoff score.

The authors based the classification primarily on the “risk of bias reduction standards” in the Cochrane Handbook (Higgins and Green, 2008), but added three relevant categories – reliability of measures, validity of measures, and program implementation fidelity – which Cochrane does not explicitly include. The results are summarized in Table 2.

“Allocation concealment” refers to how well the study concealed the method of assigning participants to interventions. “Attrition” refers to whether the extent of participant attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), and reasons for attrition/exclusions. “Missing data” refers to the amount of missing outcomes for participants and any other data that is not available on participants. “Reliability of measures” refers to the degree of consistency and replicability in measuring participant outcomes and other key variables. “Validity of measures” refers to the degree to which the indicators actually measure what they are supposed to. “Implementation fidelity” represents the degree to which the specified procedures or components of the program are implemented as planned. “Blinding” refers whether participants, clinician and assessors are unaware of the participants’ intervention assignments. “Adequate statistical power” refers to the ability of a study to detect a program effect if it occurs; it is highly dependent on sample size. “Analytical adequacy” refers to whether appropriate statistical techniques were used to examine program effects, including the detailed description of those techniques. “Reporting bias” refers to determining whether the possibility of selective outcome reporting was examined and indicating what was found. “Other reactivity controls” refers to whether attention was paid to other factors that may have biased the evaluation results reported.

Table 2 shows how many of the EBPRs use each of 11 bias reduction standards in assessing the methodological quality of evaluation studies. Two registers use all 11 bias reduction standards; two use 10 standards; two use 8 to 9 standards; six use 6 to 7 standards; and eight use less than 6 standards. The mean number of standards utilized was 6 and the range was 2-11. Although the registers tended to use numerous factors in making decisions about a program’s evidence-base of a program, the Appendix summarizes how the *primary* strength of evidence and methodological quality criteria are combined to produce a rating of program effectiveness or classification into a specific tier of effectiveness for each of the 20 registers.

Description of tiers of evidence for multi-tiered registers

Preliminary examination of the registers suggested that the highest two tiers of evidence defined what the field terms “evidence-based,” although the distinctions between the highest two tiers varied among registers. Consequently the study conducted an analysis the evidentiary criteria distinguishing the highest two tiers for the ten multi-tiered individual program registers. A summary of the evidentiary requirements for inclusion in these two tiers is displayed in Table 3. As shown, there are similarities and differences among the registers’ requirements for “top” and “second” tiers. Eight out of 10 registers explicitly require “RCT” for their top tier. A majority of the registers require formal and/or non-formal QED for inclusion into the second tier. (“Formal” QEDs are those identified by Shadish et al., 2002, sometimes registers use the term “QED” for designs that do not conform to the

formal definitions.) One-half (5/10) of these registers require at least one RCT for the second tier.

Moreover, major differences that exist among registers include sample size requirements and sustainability of effects. Although 4 out of 10 registers have top tier requirements for sample size, variation exists in acceptable levels. For example, two registers explicitly state the required sample size (i.e., 30 and 50), while others accept an “adequate equivalent group to detect effect” or simply “a clear statement of demographics.” Similar variations in sample size requirements are present in the second tier. Specific requirements for sustainability of effects ranged from 3 months to 1 year. Furthermore, similarities and differences were observed in the top two tiers in terms of outcome measures, methodology, analytical quality and reporting bias requirements.

Similarities and differences among registers on program ratings

First, a random sample of 30 programs that appear in only one register listing individual programs was reviewed to determine the number of additional such registers for which they would be eligible based on the content of their interventions (not their evaluation findings). These programs are potentially eligible for inclusion in an average of 5.6 additional registers (SD = 2.8) that list individual programs.

Second, a random sample of 100 from the total of 355 programs that appeared in more than one register was reviewed. Registers were defined as agreeing on a program's rating if the registers classified a program in an equivalent tier in each register. Single and multi-tiered registers are scored as agreeing when a program included in a single tier register is placed in the top tier of the multi-tiered register. “Disagreement on tier placement” exists when one register places the program in its top (or single) tier, while another register places the program below its top tier, or when registers disagree whether the program has been shown to have a positive effect versus no effect. These types of disagreement highlight differences in ratings for the same programs that may be of concern to users of these registers. For programs in three or more registers, agreement or disagreement on tier placement is based on 75% or more of the registers disagreeing or agreeing.

Table 4 displays the agreement and disagreement of program ratings between the registers. Overall, 42% (42/100) of the programs were classified as disagreeing on tier placement as defined above. Disagreement between registers on effect/no effect (or insufficient evidence) occurred for an additional 11% (11/100) of the programs. Thus substantial disagreement among registers was identified for more than half of the programs (53%) rated by more than one register.

Discussion

Decision-makers need independent and objective information about which programs have the strongest and most credible evidence of effectiveness. They may be faced with a large number of candidate programs and, given resource constraints, need to select the most promising for their circumstances. Or alternatively, they are attracted to a specific program or modality and need answers about whether it is likely to succeed in their setting and/or

whether it is likely to be an improvement over what is already being done. In order to ensure comparability of program ratings across registers, the registers would ideally use a consistent set of evidential criteria and standards when evaluating programs. The primary model of standards for judging the quality of evidence used by the EBPRs clearly follows a “Campbellian approach” (Gugiu, this issue), based on positivist principles of scientific inquiry including the concept of controlled experimentation (Shadish et al., 2002). Even the registers that do not require formal RCTs or QEDs nevertheless weigh the quality of evidence against a Campbellian standard of control for selection bias (“sequence generation”), which raises two issues.

First, it is rarely feasible even for large federal agencies to evaluate large scale or societal interventions by RCTs or even quasi-experimental designs. For instance, such designs could not be used to evaluate the Affordable Care Act (ACA). According to the Campbellian hierarchy of evidence, evaluations of the ACA would always be considered problematic or ambiguous; one would never be able to conclude whether the ACA was effective or not if an experimental design were the evidentiary standard. The same would hold true for any kind of changes in national legislation, policy or programming related to behavioral health care. Even randomization of institutions or cities to intervention conditions may not be feasible. Of course broad-based legislation or policies are being evaluated, but using evaluation paradigms that do not follow the orthodox Campbellian model, but rather an econometric approach, for instance. (Gu et al., 2014; Greene et al., 2014; Abraham, 2014; Wagner et al., 2014; Knudsen and Roman, 2014). The problem is that such non-experimental evaluations would either not be eligible for inclusion or would rank low in a registers’ current rating schemes.

Second, the registers do not explicitly use qualitative evidence of program impact. (Note that requirements such as “large effect” or “adequate sample size,” which sometimes appear, are not qualitative criteria but rather imprecise quantitative statements.) The lack of available protocols for assessing qualitative evidence in a systematic way may well be a deterrent to its utilization, which is unfortunate because many practical evaluations today include a qualitative component (6 & Bellamy, 2012²; Alkin, Vo, & Christie, 2012; Chelimsky, 2012; Shadish, Cook, & Campbell, 2002; Shipman, 2012). One such protocol to assess qualitative evidence has recently been proposed by Claes (this issue).

The great majority of the programs (79%) listed on the registers of individual programs appear on only one such register, but the study has shown that programs appearing only once are eligible for an average of 5.6 additional registers with relevant scopes. This in itself presents the user with a conundrum. If a program is indeed evidence-based, why is there no mention of it in other relevant registers?

A program within the scope of a register may not have been included in that register for several reasons. A register’s searches are typically conducted on varying schedules and/or limited to specific journals, so newer or relatively less well known programs may not ever be found by a register. Some registers depend on programs being nominated for inclusion

²Yes, the first author’s surname is legally “6.”

and any given program might not have been nominated. Some registers have minimal screening requirements, particularly that a program be studied using RCTs or that multiple evaluation studies have been conducted, both requirements being expensive, so only a limited number of higher profile or well-studied programs would appear on such registers. The registers also may not be basing their assessments on the same body of evaluation studies, depending on when the program reviews were conducted; this is especially difficult for users to determine. Some registers explicitly note the date of the review but others do not.

A basic difference between the registers is in whether they use an “inclusion/exclusion” protocol or rank programs in terms of degree of evidentiary support. Clearly a ranking system inherently yields finer distinctions regarding a program's evidence base, and thus may narrow down the number of programs involved in a choice. However, it also complicates the task of users in identifying suitable programs. If “evidence-based” is indeed a matter of degree, then what degree of evidence should be considered sufficient for what specific purposes? The registers do not provide guidelines for answering this question. For that standpoint, the inclusion/exclusion registers simplify the task of users, but perhaps at the expense of recommending many programs with very different degrees of evidence. Since consequently there are so many evidence based programs from which to choose, the task of choosing still needs to be faced. (Of course decisions about implementing programs involve considerations other than effectiveness, such as practicality, expense and values that can help to narrow choices.)

The study documented numerous differences among registers in their evidentiary requirements, including many dimensions of “strength of evidence.” These variations in evidentiary requirements may affect consumers’, researchers’, and policy makers’ decisions on what programs are considered evidence-based or best practice. The different rating protocols also make it difficult to compare different registers’ assessments of the same programs, since disagreements on the rating of a program are fairly common, as this study has shown. In particular, the study analyzed differences between multi-tiered registers in their criteria for top tier and second tier rankings. There was clearly overlap in the criteria for inclusion in top and second tiers. E.g., a program may qualify as top tier on register but only second tier on another. Programs labeled as second tier may or may not be considered acceptable for the user's purposes. Disagreements between register ratings that are based on differences in evidentiary criteria are especially difficult for users to reconcile. Unfortunately, users must investigate the reasons why registers disagree in rating programs to fully understand which program is “best” or which they should prefer for their particular purposes. The Appendix, which summarizes the evidentiary criteria employed by the 20 registers, can assist in that regard.

Although the evidentiary standards of the registers all derive from a Campbellian approach, as indicated above, the specifics of the standards vary considerably. Only nine registers (45%) require the supporting studies to be RCTs or QEDs. Cochrane-like bias control criteria of quality are not used in a consistent and comprehensive manner. This results in a situation where each register assesses and rates program in a particularistic fashion; common patterns are very difficult to ascertain (cf. Appendix). A user might well be edified when a

program of interest appears in more than one register, but deconstructing why a behavioral health program might be listed on one register but not another of similar scope is no easy task.

Most of the registers include or rate programs on overall strength of evidence which combines the concepts of effectiveness (outcomes) and methodological quality of the underlying studies. This has the potential to rate a program with a minimal (but statistically significant) effect size as “best” if the evaluations are of high quality. However, users may be surprised to learn that a program with a minimal effect can obtain the highest rating. Usually users are seeking programs with large potential effects; methodological quality of the evidence is a rather esoteric concept to lay people. Moreover, effect sizes that are significant in large published multisite trials might not be significant when the program is adopted in settings with fewer numbers of clients. Ninety-five percent of the registers do not specify a numerical effect size and few mention the magnitude of effects at all in including or rating programs. This relative inattention to effect size by the registers may lead to the implementation of programs whose impact on the target populations will disappoint users.

An implicit assumption of the EBPRs that list individual programs is that the effectiveness of a program is essentially context-free. That is, once shown to be effective with a certain population in a certain setting at a certain time; the results are generalizable to other populations, setting and times. This is, unfortunately, a problematic assumption. While the chances are certainly higher of succeeding with a program that has already shown a positive effect compared with one that has not, it remains necessary to evaluate the effectiveness of the new program implementation. That is, selecting and implementing a listed EBP does not absolve the user from the task of evaluation, particularly replication.

Program costs are of course important to the field. Six of these EBPRs directly address the question of program costs in their summaries. These registers provided information that ranged from direct program costs, to costs of curriculum materials, and in some cases, intangible resources that are also needed (i.e. staff time). In general, however, this question seems generally out of the scope of these registers as they are currently structured. One problem is that the source evaluation studies rarely address costs, so that cost information would usually need to be obtained as a separate data collection task. Providing program cost information would be an important way to enhance the utility of these registers.

The five registers limited to assessing intervention modalities, rather than individual programs, pose an additional question to the user. In meta-analyses and systematic reviews, not all specific programs may be shown to be effective, despite the fact that the overall clinical approach is effective. For instance, motivational interviewing as a counseling modality is generally found to be effective, but not every program incorporating motivational interviewing is effective or equally effective. Thus, users must peruse meta-analyses and systematic review carefully to identify the specific programs, their contexts and populations, for which the program has shown positive effects within the modality being assessed.

Conclusions

Several conclusions can be drawn about how EBPRs rate behavioral health interventions and the consequent implications for program selection. Because many programs have detectable positive effects, a procedure is needed by which to differentiate the best from the others, whether best, good, marginally effective, or ineffective. Although EBPRs are meant to be that mechanism, only registers that rank programs and include a “no effect” category potentially provide enough information to make those distinctions. Thus, registers that rank programs and identify best practices are more useful to decision-makers than those that only identify programs as evidence-based, or not. However, only 10 (50%) of the registers had more than one positive evidentiary tier.

Although greater consistency in the standards of evidence used by the registers would reduce uncertainty in program selection, it is not necessarily prudent to assume that evidence can be judged from a single framework. Randomized controlled trials may meet a certain concept of a “gold standard,” there are alternative paradigms for evaluating programs. Different *values* are at play in the world of evidence-based programs, so different users will value different types, qualities, and standards of evidence.

More work around clarifying consistent standards of evidence and incorporating qualitative evidence could potentially increase the adequacy of EBPRs in identifying the “best” interventions for given conditions beyond reasonable doubt. Also increased availability of resources to conduct thorough searches of the literature, to increase comprehensiveness and timeliness of the program reviews, and presenting findings of no effect, could enhance the utility of many of the EBPRs. Disagreement among registers in rating the same program is likely to be disconcerting for users, as is the meaning of finding a program listed in only one register. Insufficient attention to effect size in rating programs appears to ignore one of the most important criteria for program selection.

For the registers of modalities, guidance to users to help them select specific programs for implementation would be beneficial. Not all programs within a given class of programs may be effective or equally effective.

Lessons learned

Future research should more closely examine the implications of program or intervention scale on the types of evidence that are feasible and meaningful to collect. Large scale, multi-site studies have the resources to collect extensive outcome and implementation data, whereas most evaluations are more modest in capability. How this translates into assessment of which programs are evidence-based should be further studied.

Additionally, future research should examine the ways in which meta-analyses could be better used to facilitate decision making. The registers of modalities appear to be more difficult to use and interpret for decision-makers seeking the “best” program for their given context. Finally, extending these EBPRs’ evidentiary criteria to add qualitative data to the Campbellian model could provide additional relevant information on program effectiveness.

Acknowledgments

The authors would like to extend our appreciation to Dr. Robert Slavin, who has done much work related to evidence-based education reform. His contributions to this study helped us to make several critical leaps in our thinking, and although the Best Evidence Encyclopedia was ultimately not incorporated into our final analysis sample, we feel that his work is an important contribution to the field. <http://www.bestevidence.org/>

We would also like to acknowledge our expert review panel, which included Dr. Robert Boruch, Dr. Katrina Bledsoe, Dr. Christina Christie, Dr. Stuart Donaldson, Dr. Lois-Ellen Datta, Dr. Jennifer Greene, Dr. George Julnes, Dr. Mitchell Prinstein, Dr. M. Rebecca Kilburn, Dr. Frances Lawrenz, Dr. Michael Quinn Patton, Dr. Paul Roman, Dr. Roger Weissberg, and Dr. Brian Yates. The experience and knowledge these individuals brought to the project were invaluable in our understanding of evidence-based practice.

This article has been made possible through support from NIH NIDA (grant number 1R21DA032151).

APPENDIX

APPENDIX

Summary of Evidentiary Criteria for EBPRs in Behavioral Health

Blueprints for Healthy Youth Development (Blueprints)				
Model Programs		Promising Programs		
<ul style="list-style-type: none">• Program is clearly described• Two RCTs or one RCT and one QED evaluations• Significant positive change and no negative effects• Program is ready for dissemination		<ul style="list-style-type: none">• Program is clearly described• One RCT or two QED evaluations• Significant positive change and no negative effects• Program is ready for dissemination		
California Evidence Based Clearinghouse for Child Welfare (CEBC4W)				
Well Supported	Supported by research evidence	Promising research evidence	Evidence fails to demonstrate effect	Concerning practice
<ul style="list-style-type: none">• At least two independent RCTs that have been published in the peer reviewed literature• At least one RCT shows sustained effect of 12 months• Measures must be reliable and valid• No evidence of harm• The practice has a manual supporting it	<ul style="list-style-type: none">• At least one RCT that has been published in the peer reviewed literature• Sustained effect 6 months beyond end of treatment• Measures must be reliable and valid• No evidence of harm• The practice has a manual supporting it	<ul style="list-style-type: none">• At least one study using some form of control group• No evidence of harm• Measures must be reliable and valid	<ul style="list-style-type: none">• Two or more RCTs that have been published in the peer reviewed literature find no evidence of effect	<ul style="list-style-type: none">• The intervention has a negative effect, data suggests risk of harm• No data suggests risk of harm• The practice is generally accepted for use with children• The practice has a manual supporting it
Campbell Collaboration (Campbell)				
General. The Campbell Collaboration includes systematic reviews and meta-analysis of the evidence surrounding program effectiveness. The collaboration does not provide a single benchmark for outcomes or designs that must be included in their reviews. Instead, the review authors make decisions about what is included in individual reviews. However, there are standards which the collaboration requires in relation to review quality. These standards are not comparable to the standards used to include programs.				
Center for Reviews and Dissemination York (CRD York)				
The Center for Reviews and Dissemination, York includes systematic reviews and meta-analysis of the evidence surrounding program effectiveness. The center does not provide a single benchmark for outcomes or designs that must be included in their reviews. Instead, the review authors make decisions about what is included in individual reviews. However, there are standards which the center requires in relation to review quality. These standards are not comparable to the standards used to include programs.				
General:				
<ul style="list-style-type: none">• Are sufficient details about the individual included studies presented?• Was the quality of the included studies assessed?• Was the search adequate?• Were inclusion/exclusion criteria reported?				
Were the included studies synthesized?				

Child Trends	
Inclusion	No effect
<ul style="list-style-type: none"> • Only social interventions • Treatment and control group • Random assignment to conditions • Reports in English • Response rate >50% • P .05 significance level 	<ul style="list-style-type: none"> • Reviewed and does not meet effectiveness criteria.
Cochrane Collaboration (Cochrane)	
<p>General. The Cochrane Collaboration includes systematic reviews and meta-analysis of the evidence surrounding program effectiveness. The collaboration does not provide a single benchmark for outcomes or designs that must be included in their reviews. Instead, the review authors make decisions about what is included in individual reviews. However, there are standards which the collaboration requires in relation to review quality. These standards are not comparable to the standards used to include programs.</p>	
Compendium of Evidence-Based HIV Behavioral Interventions (Compendium/HIV) AND Effective Interventions: HIV Prevention that Works (Effective/HIV) (Both sponsored by the Centers for Disease Control)	
Pre-requisites	Best practice/evidence
<ul style="list-style-type: none"> • Analysis must be based on post-intervention levels or on pre-post changes in measures • Clear description of key aspects of the intervention • Comparison between intervention arm and an appropriate comparison arm • For an intervention with a replication evaluation, no significant negative intervention effects in the replication study • For pre-post changes used in analysis, measures must be identical, including identical recall period • No evidence that additional limitations resulted in a fatal flaw • No negative and statistically significant (p < .05) intervention effect for any relevant outcome (increase in HIV infection rate) • No other statistically significant harmful intervention effect • Positive and statistically significant (p .05) intervention effect for at least relevant outcome measure 	<ul style="list-style-type: none"> • Analysis based on an $\alpha = .05$ (or more stringent) and a 2-sided test • Analysis of participants regardless of the level of intervention exposure • Analysis of participants subjects in study arms as originally allocated regardless of contamination or logistic/implementation issues • Analytic sample 50 participants per study arm • Appropriate and concurrent comparison arm • At least a 70% retention rate at a single follow-up assessment for each study arm • Effect at the follow-up and based on the analyses that meet study implementation and analysis criteria • Follow-up assessment 3-months post completion of intervention for each study arm with recall not referring to pre-intervention period) • Prospective study design • Random or minimally biased assignment of subjects to study arms • Use of appropriate cluster-level analyses if assigned to study arms by cluster or group • With nonrandomized assignment, either no statistical differences in baseline levels of the outcome exist or baseline differences are controlled for in the analysis
Good practice/evidence	Good practice/evidence
<ul style="list-style-type: none"> • Analysis based on an $\alpha = .05$ (or more stringent) and a 2-sided test or one sided test if a-priori direction is hypothesized • Analysis of participants may be based on intervention exposure, where participants exposed to less than 50% of the entire intended intervention may be excluded • Analytic sample 40 participants per study arm • Appropriate and concurrent comparison arm or historical comparison • At least a 60% retention rate at a single follow-up assessment for each study arm • Follow-up assessment 1-months post completion of intervention for each study arm with recall not referring to pre-intervention period) • If participants excluded due to contamination or low exposure, retention rate must include these participant at each follow up they were assessed. • Prospective or quasi-prospective study design • Random or minimally biased assignment of subjects to study arms allowing for selection bias not related to HIV outcomes • Assignment may be based on pre-established groups or selection into something other than the intervention provided neither is directly related to HIV risk. • With nonrandomized assignment, either no statistical differences in baseline levels of the outcome exist or baseline differences are controlled for in the analysis. • If moderately biased assignment or historical comparison was used, differences in baseline demographics also must be controlled for in the analysis. 	

Crime Solutions.gov (contributes results to FindYouth Info.gov, and overlaps with OJJDP Model Programs Guide)

Effective - Multiple studies	Effective - One study	Promising - Multiple studies	Promising - One study	No effects
<ul style="list-style-type: none"> Programs have been demonstrated to be effective by multiple studies 	<ul style="list-style-type: none"> Programs have been demonstrated to be effective in at least one study 	<ul style="list-style-type: none"> Programs have been demonstrated to be effective by multiple studies 	<ul style="list-style-type: none"> Programs have been demonstrated to be effective by at least one study 	<ul style="list-style-type: none"> Programs are either determined to have no effect or significant harmful effect
<ul style="list-style-type: none"> At least one study must be an RCT or QED Must have significant positive effects 	<ul style="list-style-type: none"> At least one study must be an RCT or QED Must have significant positive effects 	<ul style="list-style-type: none"> Non-experimental designs allowed Must have significant positive effects 	<ul style="list-style-type: none"> Non-experimental designs allowed Must have significant positive effects 	

Effective Child Therapy

LEVEL 1 – Well established treatments	LEVEL 2 – Probably efficacious treatments	LEVEL 3 – Possibly efficacious treatments	LEVEL 4 – Unknown/Untested	LEVEL 5 – Does not work
<ul style="list-style-type: none"> Two independent, large scale RCTs which demonstrate the intervention's superiority over other treatments 	<ul style="list-style-type: none"> Level 2 treatments will have strong research support, too, but they will lack the independent investigator criteria from Level 1. It is also possible to reach Level 2 with one study if that study demonstrates the treatment is superior (or statistically equivalent) to a Level 1 treatment. 	<ul style="list-style-type: none"> Level 3 treatments are what might be called "promising treatments." There may be one strong treatment study showing that the treatment is better than no treatment. Or they may be a number of smaller clinical studies without all of the appropriate methodological controls. In short, confidence that a Level 3 treatment works is not as strong as it could be. 	<ul style="list-style-type: none"> Level 4 treatments are ones that are being used but have not been studied carefully. Some would suggest that these treatments should be avoided until further evidence is available. However, for some child/adolescent problem areas for which there is limited or no evidence that compares different treatments, a treatment at this level could be worth considering as an option. 	<ul style="list-style-type: none"> Level 5 treatments have been tested in well-designed studies and have not yet demonstrated positive findings. A Level 5 treatment may either not be better than no-treatment (or other comparison treatments) or else the treatment actually made things worse. Whereas it is possible that new evidence will show that a Level 5 treatment works better in the future, a treatment listed as Level 5 would not be a good treatment option.

Evidence-Based Practices for Substance Use Disorders (Washington)

Inclusion

- Must have published research citing them as evidence based, or be recognized by reputable organizations as being evidence based.
- Preference is given to those programs with a manual.
- Has been subject to study using either RCT, QED, or in some cases a less rigorously controlled design.
- The results have been consistent in more than one study.
- The programs in studies have been implemented with fidelity.

Guide to Community Preventive Services (The Community Guide) Note: These are the general recommendation criteria. Individual study review guidelines are also available)

The Community Guide includes systematic reviews and meta-analysis of the evidence surrounding program effectiveness. The guide does not provide a single benchmark for outcomes or designs that must be included in their reviews. Instead, the review authors make decisions about what is included in individual reviews. However, there are standards which the guide requires in relation to review quality. These standards are not comparable to the standards used to include programs.

General

- Citations or discussion as to why the use of these measures is reliable
- Citations or discussion as to why the use of these measures is valid
- Clear definition of the outcome variable
- Considered consistency of coding scoring or categorization between observers (e.g., inter-rater reliability checks) or between different outcome measures

Guide to Community Preventive Services (The Community Guide) Note: These are the general recommendation criteria. Individual study review guidelines are also available)

- Considered how setting and sampling of study population might affect reliability
 - Did the authors conduct an appropriate analysis
 - Did the authors specify the screening criteria for study eligibility
 - Interpretation of results
 - Is the intervention well described
 - Measurement of the outcome in different ways
 - Measures of internal consistency
 - Were the exposure and outcome measures valid and reliable
-

Home Visiting Evidence of Effectiveness (HomVEE)

Meets DHHS Criteria. At least 1 high- or moderate quality impact study of the model finds favorable, statistically significant impacts in two or more of the eight outcome domains OR at least two high- or moderate-quality impact studies of the model using non-overlapping analytic study samples with one or more favorable, statistically significant impacts in the same domain.

High quality studies
Random assignment

- Meets WWC standards for acceptable rates of overall and differential attrition
- No reassignment; analysis must be based on original assignment to study arms
- No confounding factors; must have at least 2 participants in each study arm and no systematic differences in data collection methods
- Controls for selected measures if groups are different at baseline

Single case designs

- Timing of intervention is systematically manipulated
- Outcomes meet WWC standards for inter-assessor agreement
- At least three attempts to demonstrate an effect
- At least five data points in relevant phases
- Regression Discontinuity
- Integrity of forcing variable is maintained
- Meets WWC standards for low overall and differential attrition
- The relationship between the outcome and the forcing variable is continuous
- Meets WWC standards for functional form and bandwidth

Moderate quality studies
Random assignment

- RCT with reassignment **OR** unacceptable rates of overall or differential attrition
- RCT with Baseline equivalence established on selected measures
- RCT with no confounding factors; must have at least 2 participants in each study arm and no systematic differences in data collection methods

Matched comparison group

- Baseline equivalence established on selected measures and controls for baseline measures of outcomes, if applicable
- No confounding factors; must have at least 2 participants in each study arm and no systematic differences in data collection methods

Single case design

- Timing of intervention is systematically manipulated
 - Outcomes meet WWC standards for interrater agreement
 - At least three attempts to demonstrate an effect
 - At least three data points in relevant phases
 - Regression discontinuity
 - Integrity of forcing variable is maintained
 - Meets WWC standards for low attrition
 - Meets WWC standards for functional form and bandwidth
-

Does not meet DHHS criteria (Insufficient evidence). The program was reviewed and did not meet the criteria

National Registry of Evidence-based Programs and Practices (NREPP)

Inclusion

- Design must be experimental (randomized controlled trial) or quasi-experimental (minimally pre- and post-test assessments with a control or comparison group).
- At least one statistically significant behavioral outcome.
- At least one study.
- Appropriate analysis is necessary to make an inference that an intervention caused reported outcomes.
- Availability of quality assurance procedures.
- Availability of training and support resources.

National Registry of Evidence-based Programs and Practices (NREPP)			
<ul style="list-style-type: none">• Availability of treatment materials.• Often variables other than the intervention may account for the reported outcomes. The degree to which confounds are accounted for affects the strength of causal inference.• Outcome measures should have acceptable reliability to be interpretable. “Acceptable” here means reliability at a level that is conventionally accepted by experts in the field.• Outcome measures should have acceptable validity to be interpretable. “Acceptable” here means validity at a level that is conventionally accepted by experts in the field.• Study results can be biased by participant attrition and other forms of missing data. Statistical methods as supported by theory and research can be employed to control for missing data and attrition that would bias results, but studies with no attrition or missing data needing adjustment provide the strongest evidence that results are not biased.• The “experimental” intervention implemented in a study should have fidelity to the intervention proposed by the applicant. Instruments that have tested acceptable psychometric properties (e.g., inter-rater reliability, validity as shown by positive association with outcomes) provide the highest level of evidence.			
Office of Adolescent Health Teen Pregnancy Prevention Resource Center (Contributes results to FindYouthInfo.gov) (TPP)			
Meets DHHS Criteria for Evidence based program		Does not meet criteria (Insufficient Evidence) The program was reviewed, and did not meet DHHS criteria.	
<ul style="list-style-type: none">• Program demonstrates impact that meets DHHS criteria, in studies of high or moderate quality.			
High study quality			
<ul style="list-style-type: none">• Random or functionally random assignment• Meets WWC standards for overall and differential attrition• Control for statistically significant baseline differences• Analysis is based on original assignment• Must have at least two subjects or groups in each research group and no systematic differences in data collection methods			
Moderate study quality			
<ul style="list-style-type: none">• Quasi-experimental design with a comparison group; random assignment design with high attrition or reassignment• Must establish baseline equivalence of research groups and control for baseline outcome measures• Must have at least two subjects or groups in each research group and no systematic differences in data collection methods			
PracticeWise			
Best support	Good support	Moderate support	Minimal support
<ul style="list-style-type: none">• At least two randomized control trial demonstrating superiority to comparison group or placebo• Experiments must be conducted with treatment manuals• The intervention effects are demonstrated by at least two independent investigators	<ul style="list-style-type: none">• Two experiments showing the treatment is superior to wait list or no-treatment group• OR one between group design demonstrating efficacy (superior to placebo or equivalent to an established treatment)• Has program manuals• Comparison groups clearly specified	<ul style="list-style-type: none">• One between group design-experiment• Clear specification of groups• Has program manual• Equivalent to already established treatments	<ul style="list-style-type: none">• One experiment showing statistically significant superiority to a wait list or control group• This program has been tested in at least one study, but has failed to meet criteria for one of the support levels
Promising Practices Network (PPN)			
Proven programs		Promising programs	
<ul style="list-style-type: none">• At least one outcome is changed by 20%, 0.25 standard deviations, or more.		<ul style="list-style-type: none">• Change in outcome is more than 1%	
		Not listed on site	
		<ul style="list-style-type: none">• Distribution of assessment is restricted, for example only to the sponsor of the evaluation.	

Promising Practices Network (PPN)

- At least one outcome with a substantial effect size is statistically significant at the 5% level.
- Program must directly impact one of the indicators used on the site
- Publicly available
- Sample size of evaluation exceeds 30 in both the treatment and comparison groups.
- Study design uses a convincing comparison group to identify program impacts, including randomized-control trial (experimental design) or some quasi-experimental designs.
- Outcome change is significant at the 10% level (marginally significant)
- Program may impact an intermediary outcome for which there is evidence that it is associated with one of the PPN indicators.
- Program materials must be publicly available.
- Sample size of evaluation exceeds 10 in both the treatment and comparison groups.
- Study has a comparison group, but it may exhibit some weaknesses, e.g., the groups lack comparability on pre-existing variables or the analysis does not employ appropriate statistical controls.
- No outcome change is significant at less than the 10% level.
- No outcome is changed more than 1%
- Program impacts an outcome that is not related to children or their families, or for which there is little or no evidence that it is related to a PPN indicators (such as the number of applications for teaching positions).
- Sample size of evaluation includes less than 10 in the treatment or comparison group.
- Study does not use a convincing comparison group. For example, the use of before and after comparisons for the treatment group only.

Resource Center for Adolescent Pregnancy Prevention (ReCAPP)

Inclusion

- The intervention is based upon approaches that have been demonstrated to be effective
- The interventions content areas include at least one of the following: Pregnancy Prevention, HIV/AIDS, and STI prevention
- The intervention provides skill building opportunities
- Portrays sexuality as a healthy part of life
- Population is school ages youth
- Must measure risk or behavior outcomes, not just knowledge or attitudes
- Research must be of sufficient sample size
- Research includes an intervention group and a control or comparison group
- Follow up measurement takes place four weeks after the intervention
- A report of the study has been published (or submitted for publication in a peer reviewed journal)

Social Programs that Work

Top tier

- Few or no control group members received the intervention group equivalence
- If the study asked sample members to consent to study participation, they provided such consent before learning whether they were assigned to the intervention versus control group
- If the study claims that the intervention has an effect on outcomes, then it reports the policy or practical significance of the effect
- If the study claims that the intervention has an effect on outcomes, then it reports the size of the effect
- If the study claims that the intervention has an effect on outcomes, then it reports the statistical significance of the effect intervention
- Demonstrated in more than one site using RCT
- Intervention was evaluated in real-world settings
- No countervailing evidence

Near top tier

- Study has nearly all the features of the top tier, but needs an additional step in order to reach the top tier (i.e. an additional replication)

Insufficient Evidence

- Does not yet meet criteria for Top Tier or Near Top Tier

Social Programs that Work

- Preferably, the study measured whether the intervention's effects lasted long enough to constitute meaningful improvement in participants' lives (e.g., a year, but preferably longer)
- Random assignment
- Report on all outcomes
- The study collected outcome data in the same way, and at the same time, from intervention and control group members
- The study measured outcomes that are of policy or practical importance – not just intermediate outcomes that may or may not predict important outcomes
- The study obtained outcome data for a high proportion of the sample members originally randomized (i.e., the study had low sample “attrition”)
- The study used “valid” outcome measures – i.e., outcome measures that are highly correlated with the true outcomes that the intervention seeks to affect.
- The study, in estimating the effects of the intervention, kept sample members in the original group to which they were randomly assigned
- Where appropriate, the members of the study team who collected outcome data were “blinded” – i.e., kept unaware of who was in the intervention and control groups

What Works Clearinghouse (WWC)

Pre-requisites

- The study uses an RCT or quasi-experimental design
- The study must account for confounding factors
- In what range do the overall and differential attrition rates fall?
- Is there evidence of baseline non-equivalence?
- Outcome data must be collected in the same way for both the treatment and experimental groups
- Outcome must demonstrate face validity or reliability
- Outcome must not be over-aligned with the intervention
- The study includes an outcome within a domain specified in the protocol
- The study is a primary analysis of the effect of an intervention
- The study uses a sample that is aligned with its protocol
- The study was published within the relevant time frame

Positive	Potentially positive	Mixed	No discernible	Potentially negative	Negative
<ul style="list-style-type: none"> • Strong evidence that intervention had a positive effect on outcomes 	<ul style="list-style-type: none"> • Evidence that intervention had a positive effect on outcomes with no overriding contrary evidence 	<ul style="list-style-type: none"> • Evidence that intervention's effect on outcomes is inconsistent 	<ul style="list-style-type: none"> • No evidence that intervention had an effect on outcomes 	<ul style="list-style-type: none"> • Evidence that intervention had a negative effect on outcomes with no overriding contrary evidence 	<ul style="list-style-type: none"> • Strong evidence that intervention had a negative effect on outcomes

Biography

Stephanie N. Means

Stephanie N. Means is a project manager at WMU's Evaluation Center and a Doctoral Candidate in the Interdisciplinary Doctoral Program in Evaluation (IDPE). She is currently working on several projects, including the Urban Core Collective, Professional Development in Science Teaching, and an evaluation of Literacy Center of West Michigan's adult literacy programs. Her past grant work includes projects with National Institute on Health, National Science Foundation, W.K. Kellogg Foundation and several other state and local level evaluations.

Stephen Magura

Stephen Magura is the Director of the Evaluation Center at Western Michigan University and former Deputy and Acting Executive Director at National Development and Research Institutes. He has led clinical efficacy trials of behavioral, psychosocial and medication treatments as well as evaluations of social services programs. He has been principal investigator of 19 research grants from the National Institutes of Health and is the current Editor-in-Chief of the journal *Substance Use and Misuse*. He has also contributed three evidence-based programs to the National Registry of Evidence-based Programs and Practices (NREPP).

Jason T Burkhardt

Jason Burkhardt is a project manager at WMU's Evaluation Center, and a Doctoral Candidate in the Interdisciplinary Doctoral Program in Evaluation (IDPE). He is currently working on several projects, including the Evaluation Resource Center for the National Science Foundation's Advanced Technical Education Program and an evaluation of WMU's Upward Bound Program. His past grant work includes projects with the National Institute on Drug Abuse, the American Red Cross, the American Council on Education, and several state and local level evaluations. His dissertation work focuses on "Users of Evidence-Based Practice Registers".

Daniela C. Schröter

Daniela Schröter is the Director of Research in WMU's Evaluation Center (EC) and an Associate Faculty member in the Interdisciplinary Ph.D. in Evaluation program (IDPE). She leads evaluation studies, conducts research on evaluation, provides professional development in evaluation, serves on the EC's leadership team, and works with a diverse group of doctoral students. Schröter received the 2013 AEA Guttentag Outstanding New Evaluator Award and represents The Evaluation Center on the Joint Committee on Standards for Education Evaluation.

Chris L.S. Coryn

Dr. Coryn is the Director of the Interdisciplinary Ph.D. in Evaluation (IDPE) program and an Associate Professor in the Evaluation, Measurement, and Research (EMR) program at

Western Michigan University (WMU). He has been involved in and led numerous research studies and evaluations, funded by the Department of Justice, National Science Foundation, National Institutes for Health, and others, across several substantive domains, including research and evaluation in education, science and technology, health and medicine, community and international development, and social and human services. He has given numerous lectures, speeches, and workshops, both nationally and internationally.

References

- 6, P.; Bellamy, C. Principles of Methodology: Research Design in Social Science. Sage Publications, Ltd; London: 2012.
- Abraham J. How Might the Affordable Care Act's Coverage Expansion Provisions Influence Demand for Medical Care? *The Millbank Quarterly*. 2014; 92(1):63–87.
- Agency for Healthcare Research and Quality. Effective Health Care Program: Helping you make better treatment choices. 2013. <http://effectivehealthcare.ahrq.gov/index.cfm/glossary-of-terms/?pageaction=showterm&termid=24> [from Agency for Healthcare Research and Quality:] [October 27, 2013]
- Alkin MC, Vo AT, Christie CA, Julnes G, Mathison S. The Evaluator's Role In Valuing: Who and With Whom. *New Directions for Evaluation*. 2012; (133):29–41.
- APA Presidential Task Force. Evidence-Based Practice in Psychology. *American Psychologist*. May-Jun;2006 61(4):271–285. [PubMed: 16719673]
- Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up. *PLOS Medicine*. 2010; 7(9):1–6.
- Campbell, DT.; Stanley, JC. Experimental and Quasi-Experimental Designs for Research. Rand McNally; Chicago, Illinois: 1966.
- Center for the Study and Prevention of Violence - Institute of Behavioral Science, University of Colorado Boulder. Blueprints for Healthy Youth Development. 2013. Retrieved 2012-2013, <http://www.colorado.edu/cspv/blueprints/index.html> [from Blueprints for Health Youth Development:]
- The California Evidence-Based Clearinghouse for Child Welfare. Chadwick Center for Children and Families - Rady Children's Hospital; San Diego: 2006-2013. Retrieved 2012-2013, <http://www.cebc4cw.org> [from The California Evidence-Based Clearinghouse for Child Welfare:]
- Chambless DL, Hollon SD. Defining Empirically Supported Therapies. *Journal of Consulting and Clinical Psychology*. 1998; 66(1):7–18. [PubMed: 9489259]
- Chelmsky, E. Valuing, Evaluation Methods, and the Politicization of the Evaluation Process.. In: Julnes, G.; Mathison, S., editors. *New Directions for Evaluation*. Spring; 2012. p. 77-83.
- Child Trends. ChildTrends/Links Databank. 2013. Retrieved 2012-2013, <http://www.childtrends.org/> [from ChildTrends/Links Databank:]
- Coalition for Evidence-Based Policy. Social Programs that Work/Top Tier Evidence. 2012. Retrieved 2012-2013, <http://www.evidencebasedprograms.org> [from Social Programs that Work/Top Tier Evidence:]
- Cochrane Collaboration. Cochrane Collaboration. 2013. Retrieved 2012-2013, <http://www.cochrane.org> [from The Cochrane Collaboration:]
- Cooper, H.; Hedges, LV.; Valentine, JC., editors. *The Handbook of Research Synthesis and Meta-Analysis*. 2nd ed.. Russel Sage Foundation; New York, NY: 2009.
- Council for Training in Evidence-Based Behavioral Practice. Definition and Competencies for Evidence-Based Behavioral Practice (EBBP). National Institute of Health; 2008.
- Danya International/Center for Disease Control. Effective Interventions: HIV Prevention that Works. Retrieved 2012-2013, from Effective Interventions: HIV Prevention that Works. 2012. <https://www.effectiveinterventions.org/en/Home.aspx>
- Donaldson, SI.; Christie, CA.; Mark, MM. What Counts as Credible Evidence in Applied Research and Evaluation Practice?. SAGE Publications; Thousand Oaks, CA: 2009.

- ETR Associates. ReCAPP - Resource Center for Adolescent Pregnancy Prevention. Retrieved 2012-2013, from ReCAPP - Resource Center for Adolescent Pregnancy Prevention. 2007-2009. <http://www.recapp.etr.org/recapp>
- Gambrill E. Evidence-Based Practice and Policy: Choices Ahead. *Research on Social Work Practice*. 2006; 16(3):338–357.
- Greene J, Sacks RM, McMenamin SB. The Impact of Tobacco Dependence Treatment Coverage and Copayments in Medicaid. *American Journal of Preventive Medicine*. Apr; 2014 46(4):331–336. [PubMed: 24650835]
- Gu Q, Koenig L, Faerberg J, Steinberg R, Vaz C, Wheatley MP. The Medicare Hospital Readmissions Reduction Program: Potential Unintended Consequences for Hospitals Serving Vulnerable Populations. *Health Services Research*. Jun; 2014 49(3):818–837. [PubMed: 24417309]
- Hawai'i State Center for Nursing. [November 2, 2013] Hawai'i State Center for Nursing. 2013. <http://www.hinursing.org/best-practice-quality.htm> [from Hawai'i State Center for Nursing:]
- Hennessy KD, Green-Hennessy S. A Review of Mental Health Interventions in SAMHSA's National Registry of Evidence-Based Programs and Practices. *Psychiatric Services*. 2011:303–305. [PubMed: 21363903]
- Higgins, J.; Green, S. *Cochrane Handbook for Systematic Reviews of Interventions* version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011.
- Institute of Education Sciences. What Works Clearinghouse. 2013. Retrieved 2012-2013, <http://www.ies.ed.gov/ncee/wwc> [from What Works Clearinghouse:]
- Jette UD, Bacon K, Batty C, Carlson M, Ferland A, Hemmingway RD, Volk D. Evidence-Based Practice: Beliefs, Attitudes, Knowledge, and Behaviors of Physical Therapist. *Physical Therapy*. 2003; (83):786–805. [PubMed: 12940766]
- Johns Hopkins University Center for Data Driven Reforms in Education. Best Evidence Encyclopedia. 2013. Retrieved 2012-2013, <http://www.bestevidence.org/> [from Best Evidence Encyclopedia:]
- Knudsen HK, Roman PM. The Transition to Medication Adoption in Publicly Funded Substance Use Disorder Treatment Programs: Organizational Structure, Culture, and Resources. *Journal of Studies on Alcohol and Drugs*. May; 2014 75(3):476–485. [PubMed: 24766760]
- Mattox, T.; Kilburn, M. [October 27, 2013] What is Evidence Based Practice?. 2013. http://www.promisingpractices.net/briefs/briefs_evidence_based_practices.asp [from Promising Practices Network:]
- Minnesota Department of Corrections. Study of Evidence-Based Practices in Minnesota: 2011 Report to the Legislature. St. Paul; MN.: 2011. from <http://www.doc.state.mn.us/publications/legislative-reports/documents/12-10EBPreport.pdf> [August 25, 2013]
- National Institute of Health - US Department of Health and Human Services. [1 23, 2012] Behavioral and Social Sciences Definition. 2011. 10 25 http://obssr.od.nih.gov/about_obssr/BSSR_CC/BSSR_definition/definition.aspx [from Offices of Behavioral and Social Sciences Research:]
- Office of Adolescent Health/Mathematica Policy Research. Evidence-Based Programs. Retrieved 2012-2014. Feb 26, 2014 http://www.hhs.gov/ash/oah/oah-initiatives/teen_pregnancy/db/tpp-searchable.html [from Teen Pregnancy Prevention Resource Center:]
- Office of the President of the United States. [May 29, 2014] The President's Budget FY2015 - Opportunity for All: Strengthening Partnerships with Nonprofits, Foundation, Faith-Based Institutions, and Community-Based Organizations. 2014. http://www.whitehouse.gov/sites/default/files/omb/budget/fy2015/assets/fact_sheets/strengthening-and-supporting-the-social-sector-to-grow-the-middle-class.pdf [from whitehouse.gov:]
- Office of the President of the United States of America. [May 29, 2014] The President's Budget FY2015 - Opportunity for All: Building and Using Evidence to Improve Results. 2014. http://www.whitehouse.gov/sites/default/files/omb/budget/fy2015/assets/fact_sheets/building-and-using-evidence-to-strengthen-results-in-government.pdf [from whitehouse.gov:]
- Patton, MQ. Best Practice.. In: Mathison, S., editor. *Encyclopedia of Evaluation*. SAGE Publications, Inc.; Thousand Oaks, CA: 2005. p. 31-32.
- Peek, CA. *Lexicon for Behavioral Health and Primary Care Integration: Concepts and Definitions Developed by Expert Consensus*. Agency for Healthcare Research and Quality; Rockville, MD: 2013. from <http://integrationacademy.ahrq.gov/sites/default/files/Lexicon.pdf> [May 2014]

- PracticeWise LLC. Practice Wise. 2013. (<http://www.practicewise.com>). Retrieved 2012-2013, <http://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/Mental-Health/Documents/CRPSychosocialInterventions.pdf#> [from PracticeWise:]
- RAND Corporation. Promising Practices Network on children, families, and communities. Retrieved 2012-2013, from Promising Practices Network on children, families, and communities. 2013. <http://www.promisingpractices.net/>
- Reickmann TR, Kovas AE, Cassidy EF, McCarty D. Employing Policy and Purchasing Levers to Increase the Use of Evidence-Based Practices in Community Based Substance Abuse Treatment Settings. *Evaluation and Program Planning*. 2011; 34(4):366–374. [PubMed: 21371753]
- Rossi, PH.; Lipsey, MW.; Freeman, HE. *Evaluation: A Systematic Approach*. 7th ed.. SAGE Publications, Inc.; Thousand Oaks, CA: 2004.
- Sackett DL, Rosenberg WM, Gray J, Haynes R, Richardson W. Evidence-based medicine: What it is and what it isn't. *British Medical Journal*. Jan; 1996 312(13):72–72.
- Scriven, M. *Evaluation Specific Methodology*. Kalamazoo; MI.: 2014. 4 29
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and Quasi Experimental Designs for Causal Inference*. SAGE Publications, Inc.; Thousand Oaks, CA: 2002.
- Shadish, WR.; Cook, TD.; Leviton, LA. *Foundations of Program Evaluation*. SAGE Publications; Thousand Oaks, CA: 1991.
- Shipman, S. The Role of Context in Valuing Federal Programs.. In: Julnes, G.; Mathison, S., editors. *New Directions for Evaluation*. Spring; 2012. p. 53-63.
- Southam-Gerow, MA.; Prinstein, MJ. Evidence Base Updates: The Evaluation of Psychological Treatments for Children and Adolescents.; *Journal of Clinical Child and Adolescent Psychology*. 2014. p. 1-6.doi:<http://dx.doi.org/10.1080/15374416.2013.855128>
- Substance Abuse and Mental Health Services Administration. National Registry of Evidence- Based Programs and Practices. Retrieved 2012-2013. 2013. <http://www.nrepp.samhsa.gov> [from National Registry of Evidence-Based Programs and Practices:]
- The American Psychological Association. Effective Child Therapy. Retrieved 2012-2013, from Effective Child Therapy, Evidence-based mental health treatment for children and adolescents. 2012. <http://www.effectivechildtherapy.com/>
- The Campbell Collaboration. The Campbell Library. 2013. Retrieved 2012-2013, <http://www.campbellCollaboration.org> [from The Campbell Collaboration:]
- The Centre for Reviews and Dissemination - University of York. Prospero - The Register of Prospective Systematic Reviews. 2013. Retrieved 2012-2013, <http://www.crd.york.ac.uk/prospero/> [from Prospero - The Register of Prospective Systematic Reviews:]
- The Centre for Reviews and Dissemination - University of York. The Centre for Reviews and Dissemination. 2013. Retrieved 2012-2013, <http://www.york.ac.uk/inst/crd> [from Centre for Reviews and Dissemination:]
- The Community Preventive Services Task Force. The Community Guide. 2013. 9 13 Retrieved 2012-2013, <http://www.thecommunityguide.org> [from The Guide to Community Preventive Services: What works to promote health:]
- United States Center for Disease Control. HIV/AIDS Prevention Research Synthesis Project. 2013. Retrieved 2012-2013, <http://www.cdc.gov/hiv/dhap/prb/prs/index.html> [from HIV/AIDS Prevention Research Synthesis Project:]
- United States Department of Education. Exemplary & Promising Safe, Disciplined, and Drug-Free Schools. United States Department of Education; Jessup, MD: 2001.
- United States Government Accountability Office. A Variety of Rigorous Methods Can Help Identify Effective Interventions. United States Government Accountability Office; Washington DC: 2009.
- United States Government Accountability Office. Improved Dissemination and Timely Product Release Would Enhance the Usefulness of the What Works Clearinghouse. United States Government Accountability Office; Washington DC: 2010.
- United States Interagency Working Group on Youth Programs. FindYouthInfo.gov. Retrieved 2012-2013. 2013. <http://www.findyouthinfo.gov> [from FindYouthInfo.gov:]
- United States Office of Justice Programs. Crimesolutions.gov. 2013. Retrieved 2012-2013, <https://www.crimesolutions.gov/> [from Crimesolutions.gov:]

- United States Office of Management and Budget. Memorandum M-13-17. United States Executive Branch; Washington, DC: Jul 26. 2013
- University of Washington Alcohol and Drug Abuse Institute. Evidence Based Practices for Substance Use Disorders. 2013. Retrieved 2012-2013, <http://lib.adai.washington.edu/EBPsearch.htm#> [from Evidence Based Practices for Substance Use Disorders:]
- US Department of Health & Human Services Administration for Children and Families/Mathematica Policy Research. Home Visiting Evidence of Effectiveness. 2014. Retrieved 2012-2014, <http://homvee.acf.hhs.gov/> [from Home Visiting Evidence of Effectiveness:]
- Wagner, Z.; Wu, Y.; Sood, N. The Affordable Care Act May Increase the Number of People Getting Tested for HIV by Nearly 500,000 by 2017. Jul. 2014 Retrieved 2014, <http://content.healthaffairs.org/content/33/3/378.abstract> [from Health Affairs:]

Highlights

- Registers tend to recognize the "traditional" Campbellian hierarchy of evidence
- Registers sometimes disagree on the criteria that define program effectiveness and research quality
- Registers use statistical significance as evidence of effectiveness, but usually ignore program effect size
- Registers avoid assessing qualitative evidence
- Using meta-analyses to guide program-level decision-making can be problematic

TABLE 1

Registers in the Study

Register Name	Short Name	Type of register	Number of tiers	Lists “no effects” or “insufficient evidence”
Blueprints for Healthy Youth Development	Blueprints	Programs	2	No
California Evidence-Based Clearinghouse for Child Welfare	CEBC	Programs	5	Yes
Campbell Collaboration	Campbell	Modalities ^a	1	Yes
Center for Reviews and Dissemination, York	CRD York	Modalities ^a	1	Yes
Child Trends	ChildTrends	Programs	1	Yes
Cochrane Collaboration	Cochrane	Modalities ^a	1	Yes
CrimeSolutions.gov	CrimeSolutions	Mixed	4	Yes
Effective Child Therapy	Effective Child Therapy	Modalities ^a	3	Yes
Effective Interventions: HIV Prevention that Works	Effective/HIV	Mixed	2	No
Evidence-Based Practices for Substance Use Disorders	Washington	Programs	1	No
Guide to Community Preventive Services	Community Guide	Modalities ^a	1	Yes
Home Visiting Evidence of Effectiveness	HomVee	Programs	1	Yes
National Registry of Evidence-based Programs and Practices	NREPP	Programs	1	No
Office of Adolescent Health Teen Pregnancy Prevention Resource Center	TPP	Programs	1	Yes
Practicewise	PracticeWise	Mixed	4	Yes
Promising Practices Network	PPN	Programs	2	No
Resource Center for Adolescent Pregnancy Prevention	ReCapp	Programs	1	No
Social Programs that Work	Social Programs That Work	Programs	2	Yes
The Compendium of Evidence-Based HIV Behavioral Interventions	Compendium/HIV	Programs	2	No
What Works Clearinghouse	WWC	Programs	5	Yes

^a Register consisting of meta-analyses/systematic reviews

TABLE 2

“Risk of Bias” Reduction Standards used by the Registers^{*}

Bias Reduction Standard	Uses the Standard	Does not use the Standard
Allocation Concealment	7 (35%)	13 (65%)
Attrition	14 (70%)	6 (30%)
Missing data	11 (55%)	9 (45%)
Reliability of measures	13 (65%)	7 (35%)
Validity of measures	13 (65%)	7 (35%)
Implementation fidelity	15 (75%)	5 (25%)
Blinding	5 (25%)	15 (75%)
Adequate statistical power	10 (50%)	10 (50%)
Other reactivity controls	5 (25%)	15 (75%)
Analytical Adequacy	18 (90%)	2 (10%)
Detailed analytic method	14 (70%)	
Specific analytic method required	4 (20%)	
Reporting Bias	17 (85%)	3 (15%)
Report all findings	14 (70%)	
Report only significant effects	3 (15%)	

* “Sequence generation” is discussed in the text.

TABLE 3

Evidentiary Criteria for Top Two Tiers of Multi-tier EBPRs*

Dimension	Top Tier	Second Tier (or “near top tier”)
Research Design	8 require at least one RCT 2 require functionally random assignment 1 requires “at least 2 RCT and/or 1 RCT and 1 QED 1 requires a “convincing comparison group” 1 requires at least a QED	5 require at least one RCT 2 require formal QED 4 require designs that have the essential features of a QED, but do not require formal QED 2 require minimally biased assignment to conditions
Sample Size	2 require minimum sample sizes (n=30 and n=50) 1 requires “adequate sample size/group equivalence” to detect effects 1 requires a “clear statement of demographic characteristics” in the supporting studies	2 require minimum sample sizes (n=10 and n=40). 1 requires an “adequate sample size to detect effects”
Outcome	1 requires a change of 20% or more	1 requires a change of 1% or more
Measures	1 considered “the standard measure of effect size, the standardized mean difference,” implying a cutoff for its categories Most required that the outcomes measured be specifically related to the domain of interest in the register	1 requires that outcomes are clearly defined and align with one of their required outcomes
Other Methodological Quality Criteria	2 consider attrition for their top tier rating (require a 70% subject retention rate) 3 have reliability and validity requirements 1 required blinding to some extent 1 required that interventions be evaluated in a real world setting 1 required that program evaluations are published in peer reviewed journals	1 requires “assignment at the appropriate level” 1 requires that control groups receive little or no amount of the intervention 2 have sample attrition constraints (require a 60% subject retention rate)
Analytic Methods	4 place constraints on analytic methods for their top tier 2 specifically mentioned power analyses	4 state requirements for analytic methods 1 requires “appropriate statistical analysis based on an intent to treat model” 2 state that analysis of participants may be based on intervention exposure, and must be based on a two sided test at $\alpha=.05$
Reporting Bias	1 explicitly required that supporting studies report all effects, although most registers implicitly assume this criterion	1 explicitly required that supporting studies report all effects, although most registers implicitly assume this criterion
Sustainability of Effects	3 mention specific outcomes that must be met 2 require programs to demonstrate effects for one year after treatment 2 require a three month follow-up 1 requires independent replication	1 required that program “effects last long enough to be meaningful” 2 required that programs demonstrate effects for at least 12 months after the initial intervention 1 requires that effects be sustained for at least six months after the intervention 2 specifically mention that the intervention must have impact on specific outcomes 2 require that outcomes be observable for three months post-intervention
Additional Requirements	Studies be publicly available; identification of risk and protective factors impacted by the program; the existence of a program logic model; specification of financial and human resources needed for implementation; and statements about other populations to which results may generalize	Studies be publicly available; identification of risk and protective factors impacted by the program; the existence of a program logic model; specification of financial and human resources needed for implementation; and statements about other populations to which results may generalize

* Programs may appear in more than one response category

TABLE 4

Consistency of Program Ratings among Registers (n=100 programs)

	Agreement		Disagreement		Total n
	Top Tier	Lower than top tier	Tier placement	Effect vs. no effect	
Two Registers	15 (33%)	6 (13%)	20 (44%)	4 (9%)	45
Three Registers	8 (33%)	0 (0%)	12 (50%)	4 (17%)	24
Four Registers	13 (68%)	1 (5%)	3 (16%)	2 (11%)	19
Five or More	4 (33%)	0 (0%)	7 (58%)	1 (8%)	12
Total	40	7	42	11	100