

Making the Grade

A Review and Comparison of Selected Evidence Grading Systems



March 22, 2005

Prepared for The National Research Council

By:

The Human Services Research Institute

H. Stephen Leff, Ph.D.

Jeremy Conley, B.A.

Shekinah Elmore, B.A.

This work presented in this paper was also supported by Grant No.: SM55236-02 from the Substance Abuse and Mental Health Services Administration Center for Mental Health Services for The Evaluation Center@HSRI.

Government Project Officer: Crystal Blyler, Ph.D.

Opinions and statements included in this paper are solely those of the individual authors, and are not necessarily adopted, endorsed, or verified as accurate by the National Academies, including the National Academy of Sciences, National Academy of Engineering, Institute of Medicine, or National Research Council.



Human Services Research Institute

2269 Massachusetts Avenue

Cambridge, MA 02140

www.tecathsri.org

Overview..... 3

What Are Evidence Grading Systems & Why Did They Develop? 4

Methodological Criteria Included in Evidence Grading Systems 10

The Convergent Evaluation Approach: Criteria for Different
Evidence Grading Systems Cross Walked 14

Other Considerations for Describing & Assessing EGSS:
Psychometric Properties & Feasibility of EGSS..... 20

Future Directions for the Development & Investigation
of Evidence Grading Systems..... 27

References..... 32

OVERVIEW

This paper was prepared for the “Standards of Evidence and the Quality of Behavioral and Social Science Research Strategic Planning Initiative” of the National Research Council. The paper draws on information collected and analyses conducted as part of the Project on Evidence Grading Systems of The Evaluation Center@HSRI. The Evaluation Center@HSRI is a technical assistance center for the evaluation of adult mental health system change supported by the Substance Abuse and Mental Health Services Administration.

The long range goals of the Strategic Planning Initiative (as expressed in an email attachment hereafter referred to as the SPI, 2005) are “to improve the quality of research in the behavioral and social sciences and education and strengthen the ties between behavioral and social science, public policy, and practice” (SPI, 2005). The goals of this paper are to consider some issues related to measuring the quality of evidence for certain types of research in the behavioral and social sciences and education and to compare several systems for grading evidence of interest to the SPI with respect to:

- Their identifying and background characteristics
- The methodological criteria they include
- The manner in which they select and categorize interventions
- The manner in which they measure effects and take into account the time periods for outcome studied and the extent to which replications have been implemented

Based on these comparisons we draw some tentative conclusions about the desired characteristics of and remaining challenges for evidence grading systems (EGSs).

We think it is important at the outset to distinguish between two different types of research or science. We refer to these as basic science and intervention science (Leff, 2005). In our conceptual framework, intervention science is a particular type of applied research. Although we agree with Stokes (1997) that the both types of research can be “use inspired” and contribute to scientific understanding, we believe that because their goals differ, the two types of science differ in how they assess the quality of evidence (Leff, 2005).

The quality of evidence has to be judged in relationship to its intended use. It is our view that “vigorous debates” (SPI, 2005) about what is quality evidence often occur because participants have different goals (and therefore types of science) in mind. The differences we have focused on are the differences between the science of making discoveries and testing theories versus the science of establishing the safety and effectiveness of interventions. The same evidence viewed from the perspective of these different goals can be assessed as having different quality. For suggesting a fruitful hypothesis a piece of qualitative evidence may be viewed as higher¹ in quality. For supporting a theory in a scholarly article, it may be viewed as of medium quality. For proving the safety and effectiveness of an intervention to be widely disseminated, it may be viewed as lower in quality. The SPI (2005) recognizes this when it states “even the strongest advocates for more randomized trial research do not believe it is the only legitimate research method for all types of questions and purposes of research”. This paper does not purport to review EGSs for all types of questions and goals. It focuses only on EGSs for assessing the causal evidence for the safety and effectiveness of interventions. Perhaps the most frequent reason, although not the only reason, public policy and practice call upon the social and behavioral sciences is for assistance in choosing interventions to implement or promote.

The SPI recognizes another important point. Evidence for effectiveness (causality) is not evidence for utility or value. Evidence about the causal connection between an intervention and an outcome

¹We will tend to use the terms higher and lower in quality rather than high and low to reinforce the idea that our understanding of quality is more relative than absolute at this time.

does not speak to issues such as the affordability of the intervention, the trade-off between its positive and negative effects, and the ethics of using the intervention. These “nonmethodological” issues (West et al., 2002) are addressed in some EGSs, however, we believe the two types of assessments should be accomplished separately, by different types of stakeholders. This paper briefly describes, but does not present an in-depth discussion or comparison of nonmethodological criteria.

There is one more point to make about intervention science and methodological criteria. In their classic work on experimental and quasi-experimental designs for research, Campbell and Stanley (1963) state that they will examine the validity of research designs against 12 common *threats* [our italics] to valid inference. We find it instructive to use the word *risks* rather than threats. We believe “risk”, more than the term “threat”, communicates the idea that evidence in intervention science is about decision making under the condition of consequences – that is, deciding what our risk is of being wrong in making a decision to use, provide, purchase or promote an intervention given the possibilities of both desirable and undesirable effects. From this perspective, intervention science is about managing the risk of being wrong: higher quality evidence minimizes our risk; lower quality makes our risk greater. When we consider the quality of evidence in intervention science, the stakes are anything but “academic”. As Bernstein (1996) notes:

We must be constantly aware of the likelihood of malfunctions and errors. Without a command of probability theory and other instruments of risk management, engineers could never have designed the great bridges that span our widest rivers, homes would still be heated by fireplaces of parlor stoves, electric power utilities would not exist, polio would still be maiming children, no airplanes would fly, and space travel would be just a dream. (2)

The SPI lists 7 relevant issues at its end and notes that others are possible. This paper only addresses a portion of one: the way in which evidentiary standards are defined and implemented in some applications of intervention science.

The paper addresses four questions related to systems for grading the quality of the evidence for interventions. The questions addressed are:

1. What are EGSs and why did they develop?
2. What methodological criteria are and should be included in EGSs
3. What are other considerations for describing and assessing EGSs: psychometric properties and feasibility of EGSs
4. What are future directions for the development and testing of EGSs

WHAT ARE EVIDENCE GRADING SYSTEMS AND WHY DID THEY DEVELOP

Health care and other public policy decisions are increasingly being made on the basis of evidence from empirical studies rather than on expert opinion or clinical experience alone (SPI, 2005; West et al., 2002).

Higher quality studies should lead to more realistic and reproducible estimates of intervention effects and greater acceptance of these results within the public policy and practice communities. Evidence grading systems are efforts to distinguish higher and lower quality studies (Moher, et al., 1995).

EGSs measure the quality of a single or multiple intervention studies. Lohr and Carey (1999) have defined study quality as the “extent to which all aspects of a study’s design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error” (472). West et al. (2002) define quality as “the extent to which a study’s design, conduct, and analysis have minimized selection, measurement, and confounding biases” (p.1). EGSs have been and are being developed to minimize the bias that can enter into traditional opinion-based reviews of evidence

quality and intervention safety and effectiveness through study selection and un-systematic assessments of study quality (West et al., 2002).

It is important to note that EGSs are systems to rate the strength of evidence for interventions only. Put differently, their focus is on internal validity or avoiding Type I error. They are not systems for rating the degree to which studies are well planned or explaining why studies failed. For example, a power analysis is a desirable step in study planning. It tells us if a study has sufficient subjects to have a reasonable chance of confirming a hypothesis. However, EGSs are unlikely to have criteria about whether a power analysis was implemented or whether studies had enough subjects to detect an expected effect. The focus of an EGS will be on whether an effect was detected and if subjects were allocated and followed in a manner that protects against bias.

Similarly, certain criteria in EGSs are inconsistent with good practice in basic, as opposed to intervention science (Leff, 2005). For example, unplanned analyses of subgroup differences that explore whether an intervention might have worked for some persons and not others are good practice in basic science, but are frowned upon in intervention science because intervention scientists believe using planned analyses of hypotheses is the best way to avoid capitalizing on chance in evaluating interventions. Thus studies are penalized for this by some EGSs. The manner in which unplanned analyses are treated in EGSs is not meant to imply that exploratory analyses are not worth doing only that in higher quality intervention evaluations, all analyses should be planned prior to inspection of the data.

If an expected effect is not detected, it is scientifically important to understand why this may have happened and how an intervention might be improved or refined. However, to avoid making EGSs too demanding in time and resources, we do not believe EGSs should be designed for these purposes. For example, we believe EGSs should not ask about the sufficiency of sample sizes, or if control conditions could have been contaminated in ways that reduced group differences. Nevertheless, as Table 2, below shows, some EGSs do contain criteria about aspects of studies, such as whether comparison conditions were “contaminated” by intervention being studied, designed to explain why interventions may have failed.

EGSs also differ in that some contain nonmethodological criteria (West et al., 2002) for assessing the utility of or recommending interventions. These criteria go beyond methodological issues and address issues such as cost, subpopulations covered, disseminability, and safety. We believe that the ratings for and analyses of these components of EGSs should be separated from those for methodological components. At least one EGS (The National Registry of Evidence-based Practices and Programs) has decided to have these ratings done by a diverse group of stakeholders including consumers, providers, purchasers and policy makers, and not just scientists.

Why are EGSs important

If findings in regard to the safety and effectiveness of interventions are significantly affected by study quality, then stakeholders deserve to know the quality of those studies in making decisions about promoting, purchasing, delivering and using interventions. (Moher et al, 1995)). Moreover, if researchers know the requirements of EGSs in advance, they are likely to better design, conduct, and document their studies leading to higher quality evidence (West et al., 2002).

Varieties of EGSs

There are different types of EGSs. It is important to identify and compare the various types. Hopefully, this can lead to resolving inconsistencies. As the Cochrane Collaboration GRADE Working Group has noted, “Inconsistencies among systems for grading the quality of evidence...reduce their potential to facilitate critical appraisal and improve communication...” (The GRADE Working Group, 2004, p.6).

At the broadest level, there are two different types of EGSs. Policy oriented EGSs are ones developed for policy makers, purchasers, providers and users. The best example of a policy oriented

EGS would be the multi-faceted one developed by the FDA. This EGS has evolved over the last one hundred years in the United States (Leff, 2005). Two important features of this type of EGS is that it assumes some sort of (1) application process in which investigators are expected to produce required documentation for (2) an organizational arbiter of value.

The second type of EGS is the basic science oriented EGS. This type of EGS is best illustrated by the Cochrane Collaboration. These systems work primarily from and are essentially “by-products” of published articles. These EGSs focus heavily on randomized clinical trials invented in the 1940s and depend heavily on analyses of threats to the validity of studies and meta-analysis articulated beginning in the 1960s (Leff, 2005).

Below, we review a number of EGSs. Most are basic science oriented. However, our review is from a policy oriented EGS perspective. Thus our comments assume systems in which there is some sort of application process to a recognized arbiter of value.

The recent study by the RTI-UNC Evidence-based Practice Center for AHRQ provides an example of a review of EGSs from a basic science perspective. This review identified and reviewed systems to rate the strength of evidence in the research literature (West et al., 2002). The two perspectives do not necessarily differ in what criteria are considered important for grading evidence. But a policy orientation assumes more direct access to information than a basic science one.

EGSs can also be distinguished in terms of their evidence source. West et al. (2002) assessed systems for grading the evidence in systematic reviews, articles on randomized clinical trials, observational studies, diagnostic studies, and “bodies of evidence – presumably consisting of multiple studies”. Our focus will be on EGSs for rating single or multiple randomized and observational studies as described in articles and in “bodies of evidence”. From our perspective, important lessons and future directions are suggested by comparing these EGSs. Like West et al. (2002) we found that comparing EGSs for systematic reviews and studies of diagnostic tests with those for the other sources of evidence risked confusion.

Additionally, we have identified three methodologic approaches to EGSs: guidance documents, checklists and scales. Guidance documents such as the Criteria for Evaluating Treatment Guidelines published by the American Psychological Association (2005) consist of general principles stated in narrative form for evaluating evidence. Checklists tend to ask binary (present/absent) questions about quality features of studies. Scales usually list multiple response options for each methodological criterion ordered from lower to higher quality. Guidance documents do not actually grade the evidence for interventions. Checklists and scales do. According to Moher et al. (1995), the first checklist was published in 1961 and nine had been published by 1993. The first scale was published in 1981, with an additional 24 published by 1993.

For this report, we did not attempt to inventory all checklists and scale produced since 1995. Instead, we drew on work done in The Evaluation Center Project on Evidence Grading Systems to identify scales and checklists that in our view were well-designed and/or likely to be influential and included EGSs specifically requested by the SPI.

Current EGSs

In 1995 Moher et al. (1995) reported 25 scales and 9 checklists for assessing the quality of randomized clinical trials published before January, 1993. In 2002 West et al. (2002) reported assessing 49 evidence grading systems for articles on RCTs and 19 for articles on observational studies. They also reported reviewing 40 systems for grading the evidence in “bodies of evidence”. Seven of these systems fully addressed all the criteria that West et al. (2002) deemed crucial to assessing bodies of evidence. The earliest EGS for a body of evidence was published in 1994 and the remaining ones were published in 1999 and 2000. However, West et al. (2002) do not include the FDA’s EGS, which has been evolving for the last century, probably because strictly speaking, it is not a scale (Leff, 2005).

Below we describe and compare a number of EGSs we discovered by locating studies cited in previous materials, and contacting researchers, guideline developers, and policy officials. We also did several more systematic literature searches to identify EGSs. However, our conclusions are similar to those made by West et al. (2002).

“In the end, our formal literature searches were the least productive source of systems...Our literature search was most problematic for identifying systems to grade the strength of a body of evidence...For those involved in evidence-based practice research, we caution that they may not find it productive simply to search for quality rating or evidence grading schemes through standard (systematic) literature searches...investigators wishing to build on our efforts might well consider tactics involving citation analysis and extensive contact with researchers and guideline developers to identify the rating systems they are presently using” (p.6).

Table 1 presents a number of identifying and background characteristics for the EGSs we reviewed. In this table, we note such things as who developed the EGS and whether it is a guidance document, checklist or scale. EGSs are listed alphabetically. We categorized 5 EGSs as guidance documents, 6 as checklists, 2 as scales and 1 as both a guidance document and a checklist.

Table 1. Identifying and Background Characteristics for Evidence Grading Systems Reviewed

Evidence Grading System	Developer/Sponsor	Acronym/Abbrev.	System Focus	Format	Staff Reviewers	Audience	Program Categorization
<i>Academic Achievement Programs and Youth Development: A Synthesis</i>	Child Trends	Child Trends	Programs with Academic Achievement Component	Guidance Document	Y	Any Interested in Academic Achievement Programs	Y
<i>The Blueprints for Violence Prevention Initiative</i>	The Center for the Study and Prevention of Violence & Office of Juvenile Justice and Delinquency Prevention	Blueprints	Violence Reduction & Intervention Programs	Checklist	Y	Communities seeking Violence Reduction Programs	Y
<i>Cochrane Reviewers' Handbook</i>	Cochrane Collaboration	Cochrane	All Healthcare	Guidance Document	Y	General Public	N
<i>Criteria for Evaluating Treatment Guidelines</i>	The American Psychological Association	APA	Mental Health Treatment Guidelines; All Healthcare	Checklist; Guidance Document	N	Providers	N
<i>Guide to Clinical Preventive Services</i>	U.S. Preventive Services Task Force	USPSTF	Screening Tests & Prevention Services	Checklist	Y	Provider Organizations, Quality Review Groups	Y
<i>The Guide to Community Preventive Services</i>	The Centers for Disease Control and Prevention	Community Guide	Population-Based & Public Health Interventions	Scale	Y	Public Health Practitioners & Policy Makers	Y
<i>Guidelines for the Preparation of Review Protocols</i>	Campbell Collaboration	Campbell	Social and Behavioral Interventions	Guidance Document	Y	General Public	N

Table 1. Identifying and Background Characteristics for Evidence Grading Systems Reviewed (Continued)

Evidence Grading System	Developer/Sponsor	Acronym/ Abbrev.	System Focus	Format	Staff Reviewers	Audience	Program Categorization
<i>HIV/AIDS Prevention Research Synthesis Project</i>	The Centers for Disease Control and Prevention	PRS	HIV Prevention Programs	Checklist	Y	Prevention Service Providers, Planners, & Others	Y
<i>International Conference on Harmonisation</i>	EU, EFPIA, MHLW, JPMA, FDA and PhRMA	ICH	Pharmaceuticals	Guidance Document	N	General Public, International Regulatory Organizations	
<i>National Registry of Evidence-based Programs and Practices</i>	Substance Abuse and Mental Health Services Administration	NREPP	Mental Health and Substance Abuse Interventions	Scale	Y	Any Interested in Mental Health/Substance Abuse Prevention & Treatment	Y
<i>Preventing Drug Use Among Children and Adolescents</i>	National Institute on Drug Abuse	NIDA	Drug Abuse Prevention Programs	Guidance Document	Y	Communities seeking to adopt or evaluate Drug Abuse Prevention Programs	Y
<i>The Safe and Drug-Free Schools Program</i>	U.S. Department of Education	SDFS	Educational Programs	Checklist	Y	Teachers, Administrators, Policy Makers, & Parents	Y
<i>Standards of Evidence</i>	Society for Prevention Research	SPR	Prevention Programs	Checklist	N	Any Interested in Prevention	N
<i>What Works Clearinghouse</i>	U.S. Department of Education	WWC	Educational Interventions	Checklist	Y	Educators, Policy Makers, Researchers	Y

METHODOLOGICAL CRITERIA INCLUDED IN EVIDENCE GRADING SYSTEMS

It is usual to assume that the evidence from randomized trials along with other methodological features common to such trials such as double-blinding are the reference standard to which estimates from studies without these features are compared. However, as with other gold standards, this standard is based on a framework of assumptions, in this case, ones developed about intervention research by intervention scientists (Leff, 2005). It is not possible to validate this gold standard outside this framework (West et al., 2002). Thus we must recognize that we do not know “true effects” in some absolute sense. It is more appropriate to say that within an intervention science framework, we assume well-conducted studies involving randomization and other methodological features probably give us more accurate estimates of treatment effects than other methods. Note that from this framework randomization is one of a number of methodological requirements and that from this perspective even randomized trials can have unrecognized flaws (Kunz et al., 1998).

The assumptions of intervention science as spelled out by Cochrane (Alderson et al., 2005), Campbell and Stanley (1966) and others are a set of methodological requirements, commonly referred to as criteria, which if met, contribute to our certainty that interventions under study are safe and effective. These criteria can be met to varying degrees and studies can be scored according to the extent to which they meet the various criteria. Given the above, as much as is possible, “it is important to question assumptions about research methods, and to test the validity of these assumptions empirically,” (Kunz et al., 1998, p.1185)

Table 2, below is a list of criteria drawn from the EGSs reviewed for this paper. This table shows that there are a large number of possible criteria and immediately raises questions such as are all these criteria necessary and how should they be weighted. Addressing any criteria in a study design or in a report of a study will take valuable time and resources. Ideally, any EGS should ask only about important criteria. We will comment on, but not resolve these issues below.

It should be noted that these are methodological criteria. They are not nonmethodological criteria for assessing the “utility of” or recommending interventions (Lohr and Carey, 1999), although some EGSs include such criteria. We discuss this aspect of EGSs in our section on other properties of EGSs.

The criteria are named in terms of the desired states to achieve higher quality evidence. In the interest of brevity, we have tried to make the criteria names self-explanatory. We do not always use the most common language for describing the criteria. This is because sometimes this language refers to an ideal method, when some EGSs recognize other, sometimes less than ideal, but still better than nothing, methods. For example, “control for subject awareness of assigned intervention” refers to controlling for subject expectations that might come from knowing one was assigned to a new intervention. This is ideally accomplished by keeping subjects ignorant of their assignment and referred to as “blinding”. However, this is impossible for some interventions and some EGSs recognize that there are less ideal methods of controlling for subject expectations such as measuring and statistically controlling for these; hence our less specific version of the criterion.

Table 2. Representative Methodological Criteria Included in EGSs

Pre-specification of primary outcomes	Reliability and validity of exposure measures
Pre-specification of all analyses	Control for contamination and co-intervention
Pre-specification of all measures	Comparison condition fidelity
Control for assignment/selection bias	Reliability of outcome measures
Appropriate comparison condition	Validity of outcome measures
Control for subject awareness of assigned intervention	Adherence to standards for data collection
Control for provider awareness of assigned intervention	Adjustment for differential attrition
Control for data collector awareness of assigned intervention	Adjustment for overall loss to follow-up
Assurances to participants to elicit disclosure	Adjustment for missing data
Intervention fidelity/Measurement of exposure	Analysis meets statistical assumptions
	Analysis consistent with study theory
	Adjustment for multiple measures
	Absence of or explanation for anomalous findings

Evaluating EGS criteria

Are all these criteria necessary? Are they all equally important or should they be weighted? Several methods for addressing these types of questions seem possible:

- The literature approach is to check sources such as textbooks, published articles, and reports to see if their authors endorse the importance of these criteria. As part of its Project on Evidence Grading Systems, The Evaluation Center has searched the literature in this way and established a database of articles on evidence grading that can be sorted by criterion. We have found one or more materials for each of the criteria we have identified. (We hope to make this Evidence Grading Library available on our website once copyright issues have been addressed.)
- The evidence-based approach is to compare studies that meet these criteria to varying degrees to see if these differences are associated with differences in estimated effects. This has been done to a limited degree and we report on this below.
- The convergent evaluation approach is to compare different EGSs to see if they contain the same or related criteria. We also report on this for the EGSs studied for this paper, below.
- The predictive validity approach would be to make specific predictions about what will occur if study quality is manipulated and implement studies to test these predictions.

In the next section we discuss and present findings relevant to the evidence-based and convergent evaluation approaches for EGSs reviewed for this paper. In the last section of this paper we discuss the predictive validity approach.

The evidence-based approach for evaluating EGS criteria

Assumptions about optimal research methods are operationally defined to varying degrees in the criteria included in EGSs. However, as Table 2, above, shows, many criteria could be and are included in EGSs. For reasons of efficiency and economy it would be nice to know which criteria really are important and which, if any, can be disregarded. One way to attempt to answer this question is to explore the empirically demonstrated associations between criteria and the estimated effects of interventions. Below we describe some findings about the associations between criteria assumed to be related to quality and estimates of effect.

Previous searches of PubMed and PsychInfo using key words such as methodological quality have yielded between 100 and 200 citations on articles related to methodological quality. The Cochrane Collaboration has an Empirical Methodological Studies Methods Group. In June of 2000, this group planned to set up a database of methodological reviews to be published in *The Cochrane Library*. This group has also set up The GRADE Working Group to address issues relevant to

evidence grading systems. This Working Group has published several documents links to which can be found in the publications page on its website: www.gradeworkinggroup.org/publications/index.htm. (These documents, which address more what we would call the psychometric properties and feasibility of EGSs are discussed below.) Nevertheless, in 1998, Kunz and Oxman wrote it proved difficult to develop efficient search strategies for locating empirical studies comparing investigations differing in overall quality or the quality with which methodological issues were addressed. Although there are many potentially useful citations, inconsistent reporting conventions and use of language require a significant amount of “hand searching” of materials available.

It is beyond the scope of this paper to review the 100-200 empirical studies of methodological quality we have identified. Instead, in Table 3, below, we present a summary of more frequently cited, readily available, and interpretable meta-analytic studies on the association between various criteria for quality and estimates of effect. This table gives the “flavor” of what such studies find.

Table 3 suggests that in some analyses some criteria did differ in their empirical association with effect sizes. In these instances, higher effects were associated with studies that did not implement concealed randomization, and with studies that did not “blind” outcome assessors, while studies with no or inadequate double-blinding were associated with estimates that varied from higher quality studies in both directions. Studies with no or inadequate concealed randomization also were associated with more varied effects. However, results varied. Table 3 also shows that overall quality measures in some analyses was associated with estimated effect and that quality weighting reduced the heterogeneity of intervention effects.

Table 3. The evidence base for EGS criteria

Criterion	Estimate of Treatment Effect: Low Quality vs. High Scoring
Unmasked (unblinded) rating of quality	19% lower (p<.05) (Moher et al., 1995)
One-group pre-post designs vs. control/comparison group designs	61% larger (Lipsey & Wilson, 1993)
Randomization not used or inadequately concealed vs. randomization adequately concealed	37% greater (p<.05) (Moher, et al., 1998) 41% greater (p<.05) (Schulz et al., 1995) 75% of comparisons greater (Kunz et al., 1998) No effect (Jüni et a., 1999) Heterogeneity greater (Schulz et al., 1995; Kunz et al., 1998; Lipsey & Wilson, 1993)
Inadequate or no double-blinding vs. adequate	11% lower (NS) (Moher, et al., 1998) 17% higher (Schulz et al., 1995)
Open assessment (lack of blinding) of outcome assessor	35% higher (Jüni et a., 1999)
Inadequate generation of random numbers vs. adequate	11% higher (NS) (Moher et al., 1995) 5% higher (NS) (Schulz et al., 1995)
Attrition/Exclusion (yes)	7% lower (NS) (Schulz et al., 1995) No effect (Jüni et a., 1999)
Total scale scores (categorized as low vs. high)	34% greater (p<.05) (Moher, et al., 1998). On average, larger effect with heterogeneity (Kunz et al., 1998) No difference in average effects (Emerson et al., 1990) Effect of quality differed for different scales (Emerson et al., 1990) Use of a quality weight produced the least statistical heterogeneity (Moher, et al., 1998). Estimates differed in both directions including reversal of effect (Kunz et al., 1998) No difference in heterogeneity (Emerson et al., 1990) Variation by scale suggests heterogeneity (Jüni et a., 1999) Heterogeneity greater (Lipsey & Wilson, 1993)

As the sparse and inconsistent data in Table 3 suggest, it is too early to make judgments about a definitive list of evidence-based methodological criteria. As Jüni et al. (1999) point out, EGSs differ in the criteria they include and are applied in areas where the directions, strengths, and opportunities for intrusion of potential biases differ. This view is consistent with the view of Martin Orne, (1959), Robert Rosenthal (2002), and Boruch (1997) that experiments are social psychological situations in which subject and experimenter expectations play significant and varying roles. Given this, it seems possible that different contexts will require different criteria and approaches to scoring. No current EGS has been sufficiently applied or analyzed to make this determination.

Additionally, Schultz et al. (1995) and Jüni et al. (1999) note that certain unexpected findings such as their null findings for exclusions and concealment respectively may have little value because of reporting problems. Further, as these authors note, certain criteria may be surrogate measures for the quality of other aspects of studies (6-p?; 9-p?). These potential inter-relationships also suggest caution in using results available at this time to disregard criteria at this time.

Emerging principles about evaluating EGS criteria based on empirical evidence

Although we did not systematically review all available articles, we are left with the impression that several methodological principles are emerging in this area. These are:

- The criteria in EGSs should be about the quality of evidence in studies, not the quality of reports of studies. However, studies and their reports are not completely separable and reports of studies should be improved to include all information about studies needed to assess evidence quality.
- The criteria in EGSs should be applied separately to each important outcome (The GRADE Working Group, 2004).
- Analyses of the relationships between evidence quality and estimated effects should compare the associations for comparable interventions (Moher et al., 1995; Emerson et al., 1990).
- Investigators should expect effect estimates for lower quality evidence to differ from higher ones in both directions leading to greater heterogeneity (all articles cited).
- Analyses should investigate component criteria of quality identified a priori and overall quality (Moher et al., 1995).
- Based on current analyses, EGS criteria should include at a minimum method of allocation (including concealment), double blinding or blinding of outcome assessment, handling of attrition (Jüni et al., 1999), and concordance of evidence with other evidence (Boruch, 1997).

THE CONVERGENT EVALUATION APPROACH: CRITERIA FOR DIFFERENT EVIDENCE GRADING SYSTEMS CROSS-WALKED

Tables 4a, 4b, and 4c present a crosswalk of criteria from several important EGSs or guidelines for EGSs. These EGSs were chosen because they are currently influential or of interest to policy making organizations like the National Academy of Science. The major purpose of this table is to show which criteria are common to different EGSs.

This crosswalk was prepared by two raters (JC and SE) who worked from a manual for identifying and categorizing criteria in different EGSs. The descriptions of these criteria varied widely across EGSs and the distinct EGS formats (checklist, scale, and guidance document). To address this variation we created a manual on how to identify and categorize criteria. We then conducted a pilot study to investigate whether we could agree on identifying and categorizing criteria. This study showed high interrater agreement and in the cases in which we did not agree, we were easily able to reach consensus. Following our pilot test, one of us (JC) identified and categorized the criteria in the remaining EGSs. All extracted criteria were reviewed in conference by both reviewers to ensure that consistency was maintained throughout the identification and categorization process.

Criteria in the crosswalk are listed by the methodological domain that they address. These domains are “meta-criteria” developed to address the variation in the ways in which related criteria are addressed in different EGSs. They provide a format for investigating what broad topics have received the most attention in rating systems, and which have received less. The methodological domains are:

Pre-specification of Measures and Analyses: Criteria address a priori identification of outcomes, measures, and analyses to reduce bias caused by ad hoc changes in protocol.

Control for Assignment Bias: Criteria address equivalence in comparison conditions

Appropriateness of Comparison Condition: Criteria address the nature and appropriateness of selected comparison conditions

Control for Awareness of Study Condition: Criteria address bias created by participant, provider, data collector, and analyst awareness of participants’ study condition

Standardization of Data Collector/Participant Preparation: Criteria address issues of demand characteristics and data collector training

Implementation Fidelity: Criteria address whether conditions were implemented as designed

Outcome Measure Validity: Criteria address outcome measure validity

Outcome Measure Reliability: Criteria address outcome measure reliability

Adjustment for Missing Data: Criteria address bias created by participant attrition and missing data values

Appropriateness of Statistical Analysis: Criteria address threats to internal validity caused by inappropriate data analysis

Additional details on the extracted criteria can be found in Appendix A.

The variations in descriptions of methodological criteria found in EGSs caused some criteria to be more transparent than others. Criteria that operationalize key concepts transparently allow reviews to be systematic, consistent, and comparable. Conversely, rating systems with less clearly defined criteria rely more heavily on the experiences of individual reviewers, making ratings less consistent and comparable. While two systems may employ the same broad concept in their respective criteria, if it is not operationalized in the same way, it may not in fact be the same criterion, and the comparison is not as useful. To address this issue, raters evaluated the transparency of the extracted methodological criteria using the following scale.

1. Concept is not transparently operationalized, ratings rely primarily on reviewer opinion.
2. Concept is operationalized with some transparency, reviewers are guided (especially by examples) but some subjectivity is required.
3. Concept is operationalized transparently, reviewers make decisions with little or no subjectivity

Transparency ratings were reviewed by both raters and disagreements were discussed in an effort to reach consensus. A process for involving multiple outside raters to resolve persistent conflicts was outlined in the manual, but no such disagreements were found. Transparency ratings are included on the crosswalk in parentheses following each criterion.

Table 4a presents results for EGSs categorized as guidance documents. Table 4b presents results for EGSs categorized as checklists. Table 4c presents results for EGSs categorized as scales.

These crosswalks show that criteria related to control for assignment bias and adjustment for missing data are the most commonly included ones in all types of EGSs. Checklists and scales also tend to include criteria related to outcome measure validity and reliability. EGS scales tend to be more inclusive than guidance documents and checklists.

Table 4a. A Crosswalk of Criteria in Guidance Documents

Rating System	Pre-specification of Measures and Analyses	Assignment Bias	Comparison Condition	Awareness of Comparison Condition	Collector/ Participant Preparation	Implementation Fidelity	Measure Validity	Measure Reliability	Missing Data	Analysis Bias
ICH	<input checked="" type="checkbox"/> Pre-specification of: Primary outcomes (3) All analyses (3) All measures (3)	<input checked="" type="checkbox"/> Selection Bias (3)	<input checked="" type="checkbox"/> Selection of control group (3)	<input checked="" type="checkbox"/> Blinding (2)		<input checked="" type="checkbox"/> Evaluation of Usage (2)	<input checked="" type="checkbox"/> Validated Variable	<input checked="" type="checkbox"/> Reliable Variable	<input checked="" type="checkbox"/> Full analysis set (3)	<input checked="" type="checkbox"/> Adjustment for multiplicity (3)
APA		<input checked="" type="checkbox"/> Methodological Rigor (3)	<input checked="" type="checkbox"/> Treatment conditions for comparison (3)				<input checked="" type="checkbox"/> Methods (1)		<input checked="" type="checkbox"/> Attrition (2)	
Cochrane		<input checked="" type="checkbox"/> Concealed Random Assignment (3)		<input checked="" type="checkbox"/> Subject Awareness (2) Provider Awareness (2) Collector Awareness (2)		<input checked="" type="checkbox"/> Contamination and Cointervention (1)			<input checked="" type="checkbox"/> Attrition Bias (1)	
Child Trends		<input checked="" type="checkbox"/> Study Design (3)								
Campbell										
NIDA		<input checked="" type="checkbox"/> Similar comparison condition (1)	<input checked="" type="checkbox"/> Using comparison conditions (1)			<input checked="" type="checkbox"/> Program implementation monitoring (2)	<input checked="" type="checkbox"/> Tested Data-collection instruments (1)	<input checked="" type="checkbox"/> Tested Data-collection instruments (1)	<input checked="" type="checkbox"/> Post-intervention follow up (1)	<input checked="" type="checkbox"/> Appropriate statistical methods (1)

Table 4b. Crosswalk of Criteria in Checklist EGSs

Rating System	Pre-specification of Measures and Analyses	Assignment Bias	Comparison Condition	Awareness of Comparison Condition	Collector/ Participant Preparation	Implementation Fidelity	Measure Validity	Measure Reliability	Missing Data	Analysis Bias
WWC		<input checked="" type="checkbox"/> Study Design (3) Randomization : Were participants placed into groups randomly? (2) Baseline Equivalence (1)				<input checked="" type="checkbox"/> Fidelity (2) Disruption (1)	<input checked="" type="checkbox"/> Valid Outcome Measure (2) Outcome Timing (2) Alignment (1)	<input checked="" type="checkbox"/> Reliability (2)	<input checked="" type="checkbox"/> Differential Attrition (2) Overall Attrition (2)	<input checked="" type="checkbox"/> Statistical Independence (2) Statistical Assumptions (2) Formula (1)
SPR		<input checked="" type="checkbox"/> Assignment Bias (3) 5aii. Adjustment for pretest differences (3)	<input checked="" type="checkbox"/> 3a. Comparison Condition Required (3)			<input checked="" type="checkbox"/> E2d. Level of exposure (2)	<input checked="" type="checkbox"/> 2c. Outcome Measure Construct validity (2)	<input checked="" type="checkbox"/> 2c. Outcome Measure Reliability (2)	<input checked="" type="checkbox"/> 5aiv. Adjustment of differential attrition (2)	<input checked="" type="checkbox"/> 5ai. Level of analysis (2) 5aiii. Adjustment for measurement of multiple outcomes (3)
USPSTF		<input checked="" type="checkbox"/> Study Design (3) Adjustment for potential confounders (2)		<input checked="" type="checkbox"/> Masked Outcome Assessment (1)		<input checked="" type="checkbox"/> Maintenance of comparable groups (1)	<input checked="" type="checkbox"/> Valid Measurement (1)	<input checked="" type="checkbox"/> Reliable Measurement (1)	<input checked="" type="checkbox"/> Important differential loss to follow-up (1)	

Table 4b. Crosswalk of Criteria in Checklist EGSs (continued)

Rating System	Pre-specification of Measures and Analyses	Assignment Bias	Comparison Condition	Awareness of Comparison Condition	Collector/ Participant Preparation	Implementation Fidelity	Measure Validity	Measure Reliability	Missing Data	Analysis Bias
Blueprints		<input checked="" type="checkbox"/> Random assignment or matched control groups (3)					<input checked="" type="checkbox"/> Demonstrated Instrument Validity (1)	<input checked="" type="checkbox"/> Demonstrated Instrument Reliability (1)	<input checked="" type="checkbox"/> Attrition Bias (1)	
SDFS		<input checked="" type="checkbox"/> 1b. Evaluation design (1)					<input checked="" type="checkbox"/> 1c. Valid measures (2)	<input checked="" type="checkbox"/> 1c. Reliable measures (2)	<input checked="" type="checkbox"/> 1b. Analysis to control attrition (1)	<input checked="" type="checkbox"/> 1b. Analysis controls for threats to internal validity (1) 1d. Appropriate analyses (1)
PRS		<input checked="" type="checkbox"/> Methodological Rigor (3)				<input checked="" type="checkbox"/> 2. Staff Adequately Trained for Sensitivity (2) 3. Staff Adequately Trained to Deliver Core Elements (2) 4. Core Elements Clearly Defined and Maintained in Delivery (2)				

Table 4c. Crosswalk of Criteria in Scale EGSs

Rating System	Pre-specification of Measures and Analyses	Assignment Bias	Comparison Condition	Awareness of Comparison Condition	Collector/ Participant Preparation	Implementation Fidelity	Measure Validity	Measure Reliability	Missing Data	Analysis Bias
NREPP	<input checked="" type="checkbox"/> 2. A Priori Identification of Outcomes (3) 16. A Priori Identification of Methods (3)	<input checked="" type="checkbox"/> 12. Selection Bias (3)	<input checked="" type="checkbox"/> 7. Nature of Comparison Condition (3)	<input checked="" type="checkbox"/> 9. Participant Awareness of Condition 11. Data Collector Bias (2)	<input checked="" type="checkbox"/> 10. Standardized Data Collection (3) 8. Assurances to Participants (2)	<input checked="" type="checkbox"/> 5. Intervention Fidelity (3) 6. Comparison Fidelity (3)	<input checked="" type="checkbox"/> 4. Validity (2)	<input checked="" type="checkbox"/> 3. Reliability (2)	<input checked="" type="checkbox"/> 13. Attrition (2) 14. Missing Data (2)	<input checked="" type="checkbox"/> 1. Alpha Inflation (2) 15. Analysis Meets Data Assumptions (2) 17. Analysis Consistent with Study Theory (2)
Community Guide		<input checked="" type="checkbox"/> Study Design (2) 5b. Assessment and control of confounding variables (1)	<input checked="" type="checkbox"/> 2d. Inappropriate control or comparison condition (1)			<input checked="" type="checkbox"/> 3a. Exposure measurement (3) 3b. Exposure measure validity and reliability (2)	<input checked="" type="checkbox"/> 3b. Outcome measure validity (2)	<input checked="" type="checkbox"/> 3b. Outcome measure reliability (2)	<input checked="" type="checkbox"/> 5a. Attrition bias (2)	<input checked="" type="checkbox"/> 4. Statistical testing (3) Reporting of statistical analyses (3) Controlled for design effects (1) Controlled for repeated measures (2) Controlled for differential exposure (2) Appropriate multi-level

OTHER CONSIDERATIONS FOR DESCRIBING AND ASSESSING EGSS: PSYCHOMETRIC PROPERTIES AND FEASIBILITY OF EGSS

Table 5 lists criteria for comparing the psychometric properties and feasibility of EGSSs. Most of these are not criteria for avoiding Type I error. They speak to the utility and practicality of EGSSs. They resemble closely the criteria developed by the GRADE Working Group of the Cochrane Collaboration Methods Group (7, 10). We view these criteria as secondary in importance for choosing EGSSs and consequently beyond the scope of this paper to apply to the EGSSs reviewed. Nevertheless, some, if not all, of these nonmethodological criteria should be considered in evaluating EGSSs. Column 1 of Table 5 shows the nonmethodological criteria. Column 2 presents brief rationales for the criteria.

Table 5. Other considerations for describing and assessing EGSSs: psychometric properties and feasibility of EGSSs

Nonmethodological Criteria	Rationales
Will the EGS work with only one type of intervention or problem or many different types?	The more consistent EGSSs are the more credible they will be. If EGSSs can be made more consistent we will approach the condition of having one generally applicable, maximally credible EGS. However, it may be that such EGSSs do not do justice to the differences in context discussed by Jüni et al. (1999) and others.
Will the EGS work with both Individual and multiple studies?	Ideally, effectiveness estimates should be based on evidence from multiple studies. This means EGSSs should have guidelines for aggregating quality ratings.
Is the EGS grounded in a credible definition or theory of methodological quality?	Policy oriented EGSSs should make clear the definition of quality that guided their development.
Does the EGS provide separate scoring for methodological and nonmethodological criteria (if included)?	EGSSs should focus on methodological criteria. The results of EGSSs can be used in making decisions and recommendations about utility and use. However, the two are separate tasks and ideally the latter task should involve stakeholders in addition to scientists such as consumers, providers, and purchasers.
Do EGS criteria reflect both intervention science theory and current evidence about methodological criteria that affect estimates of effect (West et al., 2002)?	Historically, EGS criteria have come from expert opinion as expressed in textbooks and journals. More recently, empirical studies have become available that relate quality scores to effect estimates. Recently, evidence quality issues have also come to be discussed in the media. Some criteria that are more widely endorsed are listed in section 3c, above.
Does the EGS require going beyond reports to obtain unreported aspects of methodological quality?	Policy oriented EGSSs should focus on studies, not reports of studies.
Are EGS criteria clear and can they be rated without subjective interpretation or averaging?	Policy oriented EGSSs should avoid multi-barreled or global questions. This is an aspect of transparency. Performance of shorter forms should be ensured by empirical work on scale development (1, 3-8)

Nonmethodological Criteria	Rationales
Does the EGS provide guidelines for scoring criteria and for avoiding bias?	EGSs should come with explicit guidelines for scoring criteria. This is another component of transparency. EGSs should also come with recommendations for minimizing rater bias. One recommendation in this regard is that raters assessing study quality do not know study authors and make quality ratings before knowing findings.
Does the EGS provide guidelines for summary score	EGSs should also come with explicit guidelines for combining/summarizing criteria scores. These should include guidelines for combining evidence quality scores with other information such as effect sizes, number of replications, etc. Options include: Simple averaging Weighted averaging Algorithm with other information (e.g., effect size)
Are the EGS criteria framed to provide actionable information	Criteria ratings should suggest ways to improve evidence quality
Do EGS criteria have interrater reliability?	EGSs should provide evidence that different raters use criteria in the same way.
Do EGS criteria yield a range of possible suggesting that the item can discriminate lower from higher quality evidence?	EGS criteria and summary scores should not exhibit floor and ceiling effects that limit the abilities of systems to discriminate.
Do EGSs exhibit convergent validity	EGSs scores should be correlated. Although an emerging principle is that EGSs should be specific to treatment and social context, intervention science requires they have some criteria in common and have some degree of positive relationship.
Do EGSs exhibit predictive or discriminant validity?	If possible, EGSs should be tested for other types of validity. Some ideas are discussed below.
On average, how long does the EGS take to implement?	The first priority of an EGS should be to appropriately grade the evidence for an intervention. However, shorter EGSs should be preferred when shorter instruments have been shown to yield as good information as longer ones.
Can the EGS be used with different audiences	Ideally an EGS should be transparent and comprehensible to consumers, providers, purchasers and policy makers. However, this may require some translation of the version actually used by raters.

Table 6, below describes how the EGSs reviewed selected and categorized interventions. This table shows that many EGSs had “threshold” rules for selecting interventions (such as no simple pre-post studies) that insured that a certain “floor” of evidence quality was maintained. This table also shows that EGSs ultimately used factors other than evidence quality in categorizing interventions. These other factors include, the significance and size of effects, the length of follow-up, and support from replications. The ways in which these factors are included in the EGSs reviewed are summarized in Table 7.

The way in which methods for categorizing interventions based on a combination of factors including, but not limited to evidence quality work and their implications for how interventions are categorized have not been analyzed by others or by us. This is clearly an area for future work.

Table 6. Evidence Grading System Methods for Selecting and Grading Interventions

Evidence Grading System	Intervention Selection Criteria	Method for Categorizing Interventions
Guidance Document EGSs		
ICH	Not Applicable	Not Applicable
APA	Not Applicable	Not Applicable
Cochrane	<ul style="list-style-type: none"> Selection criteria dependant upon topic of systematic review. 	Not Applicable
Child Trends	<ul style="list-style-type: none"> <i>Experimental</i> – study uses random assignment, or <i>Quasi-experimental</i>, study uses a matched comparison condition. 	<ul style="list-style-type: none"> 6 programs evaluated experimentally are assigned to one of three categories on the basis of study findings—“works,” “doesn’t work,” or “mixed reviews.” 6 programs evaluated in quasi-experimental studies are considered to be a source of “best bets” for program practices. That is, they seem to work, but the method of evaluating them cannot produce clear evidence of success. Also included in the “best bets” column of the “what works” tables are findings from multivariate, longitudinal studies.
Campbell	<ul style="list-style-type: none"> Selection criteria dependant upon topic of systematic review. 	Not Applicable
NIDA	<ul style="list-style-type: none"> Minimum selection criteria are unclear. <i>Each program was developed as part of a research protocol in which an intervention group and a comparison group were matched on important characteristics, such as age, grade in school, parents’ level of education, family income, community size, and risk and protective factors. The interventions were tested in a family, school, or community setting, all with positive results.</i> 	<ul style="list-style-type: none"> NIDA includes a number of research-based example programs categorized by their audience category (universal, selective, indicted, or tiered) and for whom they are designed (elementary, middle, or high school students).
Checklist EGSs		
WWC	<ul style="list-style-type: none"> Selection criteria dependant upon topic of systematic review. However, all studies must be Randomized Control Trials, Regression Discontinuity Designs, Quasi-Experimental Designs, or Single Subject Designs. 	<ul style="list-style-type: none"> "Meets Evidence Standards"--randomized controlled trials (RCTs) that do not have problems with randomization, attrition, or disruption, and regression discontinuity designs that do not have problems with attrition or disruption. "Meets Evidence Standards with Reservations"--strong quasi-experimental studies that have comparison groups and meet other WWC Evidence Standards, as well as randomized trials with randomization, attrition, or disruption problems and regression discontinuity designs with attrition or disruption problems. "Does Not Meet Evidence Screens"--studies that provide insufficient evidence of causal validity or are not relevant to the topic being reviewed.

Table 6 continued

Evidence Grading System	Intervention Selection Criteria	Method for Categorizing Interventions
SPR	Not Applicable; Guidelines are applicable to research on causality.	<ul style="list-style-type: none"> Efficacious – Meets all efficacy criteria; Effective – Meets all effectiveness criteria AND all efficacy criteria; Disseminable – Meets all dissemination criteria AND all effectiveness criteria.
USPSTF	<ul style="list-style-type: none"> Selection criteria dependant upon topic of systematic review. 	<ul style="list-style-type: none"> A. The USPSTF strongly recommends that clinicians provide [the service] to eligible patients. <i>The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.</i> B. The USPSTF recommends that clinicians provide [this service] to eligible patients. <i>The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.</i> C. The USPSTF makes no recommendation for or against routine provision of [the service]. <i>The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of benefits and harms is too close to justify a general recommendation.</i> D. The USPSTF recommends against routinely providing [the service] to asymptomatic patients. <i>The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.</i> I. The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. <i>Evidence that the [service] is effective is lacking, of poor quality, or conflicting and the balance of benefits and harms cannot be determined.</i>
Blueprints	<ul style="list-style-type: none"> <i>Experimental designs with random assignment or quasi-experimental designs with matched control groups.</i> 	<ul style="list-style-type: none"> <i>11 model programs or Blueprints have been proven to be effective in reducing adolescent violent crime, aggression, delinquency, and substance abuse and predelinquent childhood aggression and conduct disorders.</i> <i>19 programs have been identified as promising.</i>
SDFS	<ul style="list-style-type: none"> Solicited applications from any program sponsor who felt their program might be eligible. Reviewed 124 applications. 	<ul style="list-style-type: none"> 9 programs designated <i>Exemplary</i>. 33 programs designated <i>Promising</i>.
PRS	<ul style="list-style-type: none"> <i>1. Criteria for Scope – HIV Prevention focus and time period; 2. Criteria for Relevance – Relevant outcomes; 3. Criteria for Methodologic Rigor – Random or non-biased assignment.</i> 	<ul style="list-style-type: none"> Programs added to the <i>Compendium</i> if they meet additional methodological and scope criteria.

Table 6 continued

Evidence Grading System	Intervention Selection Criteria	Method for Categorizing Interventions
Scale EGS		
NREPP	<ul style="list-style-type: none"> • <i>Minimum Evidence-based Standards.</i> 	<ul style="list-style-type: none"> • Effective Program or Practice – High quality evidence combined with medium ($d > 0.5$) or large ($d > 0.8$) effect sizes and statistical significance. Replications by unaffiliated investigators have documented consistent effects of the intervention. • Conditionally Effective Program or Practice – High quality evidence combined with medium or large effect sizes and statistical significance. No independent replications by unaffiliated investigators have occurred. • Emerging Program or Practice – High quality evidence combined with smaller effect sizes ($d > 0.2$) and statistical significance, or high quality evidence combined with medium or larger effect sizes but no statistical significance. • Program or Practice of Interest – Lower quality evidence combined with small or medium effect sizes and statistical significance, or lower quality evidence combined with large effect sizes and no statistical significance. • Insufficient Current Support – Low quality evidence combined with small or medium effect sizes and no statistical significance.
Community Guide	<ul style="list-style-type: none"> • Selection criteria dependant upon topic of systematic review. 	<ul style="list-style-type: none"> • <i>Strength of Evidence of Effectiveness:</i> <i>Strong – Strongly recommended; Sufficient – Recommended; Insufficient empirical information supplemented by expert opinion – Recommended based on expert opinion; Available studies do not provide sufficient evidence to assess – Insufficient evidence to determine effectiveness; Sufficient or strong evidence of ineffectiveness or harm – Discouraged</i>

Table 7. Approaches to Including Effect, Length of Follow-up and Replications in Categorizing Interventions

Evidence Grading System Guidance Document EGSs	Approach to Including Effect	Approach to Length of Follow-up	Approach to Replications
ICH	<ul style="list-style-type: none"> • Trial effects and how they were calculated should be reported. • Significance tests and whether they are one-tailed or two-tailed should be reported. 	<ul style="list-style-type: none"> • Trial length should be reported. 	<ul style="list-style-type: none"> • Trials using the same outcome may be synthesized in order to gain a better estimate of efficacy.
APA	<ul style="list-style-type: none"> • <i>5.9 Clinical significance. Ideally, outcome descriptions should specify clinical significance (i.e., actual clinical benefit) in addition to reporting any statistical significance.</i> 	<ul style="list-style-type: none"> • <i>5.5 Long-term consequences of treatment. [T]reatments that have enduring effects following termination are to be preferred over those that do not.</i> 	<ul style="list-style-type: none"> • No Criterion: <i>Replication across multiple studies and multiple settings is desirable.</i>
Cochrane	<ul style="list-style-type: none"> • Quantitative synthesis of relevant study outcomes included in systematic reviews. 	<ul style="list-style-type: none"> • Follow-up times for relevant outcomes are considered on a case-by-case basis as determined by expert opinion. 	<ul style="list-style-type: none"> • Systematic reviews include ALL published and unpublished studies with relevant outcomes.
Child Trends	<ul style="list-style-type: none"> • Effects must be statistically significant. 	No Criterion, poses research question: <i>Is there a minimum frequency and duration of participation before programs become effective?</i>	No Criterion
Campbell	<ul style="list-style-type: none"> • Quantitative synthesis of relevant study outcomes included in systematic reviews. 	<ul style="list-style-type: none"> • Follow-up times for relevant outcomes are considered on a case-by-case basis as determined by expert opinion. 	<ul style="list-style-type: none"> • Systematic reviews include ALL published and unpublished studies with relevant outcomes.
NIDA	No Criterion	<ul style="list-style-type: none"> • <i>Principle 13: Prevention programs should be long-term with repeated interventions (i.e., booster programs) to reinforce the original prevention goals. Research shows that the benefits from middle school prevention programs diminish without followup programs in high school.</i> 	No Criterion
Checklist EGSs			
WWC	<ul style="list-style-type: none"> • Effects in individual studies are assessed in three domains: 1) statistical significance; 2) practical significance; 3) magnitude. • Quantitative synthesis of relevant study outcomes included in systematic reviews. 	<ul style="list-style-type: none"> • Follow-up times for relevant outcomes are considered on a case-by-case basis as determined by expert opinion. 	<ul style="list-style-type: none"> • Systematic reviews include ALL published and unpublished studies with relevant outcomes.
SPR	<ul style="list-style-type: none"> • <i>5b. Statistically significant effects</i> Effects must be statistically significant. 	<ul style="list-style-type: none"> • <i>5d. Duration of effect</i> Effect must be sustained for a minimum of six months. 	<ul style="list-style-type: none"> • <i>5e. Replication</i> Two replications are required. This requirement can be overridden by expert opinion if there is not sufficient research.

Table 7 continued

Evidence Grading System	Approach to Including Effect	Approach to Length of Follow-up	Approach to Replications
USPSTF	<ul style="list-style-type: none"> The USPSTF classifies benefits, harms, and net benefits on a 4-point scale: “substantial,” “moderate,” “small,” and “zero/negative.” 	<ul style="list-style-type: none"> Follow-up times for relevant outcomes are considered on a case-by-case basis as determined by expert opinion. 	<ul style="list-style-type: none"> Systematic reviews include ALL published and unpublished studies with relevant outcomes.
Blueprints	<ul style="list-style-type: none"> Effects must be statistically significant. 	<ul style="list-style-type: none"> Effects must be shown to last at least one year; no long-term effect means a program can be <i>promising</i>. 	<ul style="list-style-type: none"> At least one replication is necessary.
SDFS	<ul style="list-style-type: none"> 1a. The program evaluation indicates a measurable difference in outcomes that is based on statistical significance testing or a credible indicator of magnitude of effect. 	No Criterion	No Criterion
PRS	<ul style="list-style-type: none"> 3. Positive results on PRS relevant outcome(s) 4. Positive results that represent a statistically significant difference between the intervention and control and comparison condition. 	No Criterion	No Criterion
Scale EGSs			
NREPP	<ul style="list-style-type: none"> Outcomes are rated on effect size (small, medium, or large) and statistical significance. 	<ul style="list-style-type: none"> Follow-up times are considered to be part of outcomes. 	<ul style="list-style-type: none"> 19. <i>Replications</i> Outcomes are scored based on their number of replications and whether or not they are conducted by an independent investigator. Outcomes of replications are synthesized by reviewers. Programs that have not been replicated by an independent investigator cannot be designated <i>effective</i>.
Community Guide	<ul style="list-style-type: none"> Qualitative or Quantitative synthesis of relevant study outcomes included in systematic reviews. 	<ul style="list-style-type: none"> Follow-up times for relevant outcomes are considered on a case-by-case basis as determined by expert opinion. 	<ul style="list-style-type: none"> Systematic reviews include ALL published and unpublished studies with relevant outcomes.

FUTURE DIRECTIONS FOR THE DEVELOPMENT AND INVESTIGATION OF EVIDENCE GRADING SYSTEMS

What types of interventions and problems should be evidence based?

Can and should EGSs be applied to psychosocial interventions with complex arrangements and soft boundaries (Wolfe, 2000) and to interventions such as community coalitions and other types of advocacy and preventative services? We believe that EGSs can be applied to certain interventions within any area. However, we also believe there is a possibility that certain types of interventions should not be assessed or endorsed on the basis of intervention science. The discussion of these areas is beyond the scope of this paper. We believe that identifying those interventions that lie outside the boundaries of intervention science should be done by appropriate stakeholders such as policy makers and consumers.

Should EGSs be policy oriented or basic science oriented?

This is the question of what types of EGSs should be developed in the future? There appears to be a movement to establishing both governmental and non-governmental organizations to assess the evidence for interventions and disseminate these assessments in registries and similar formats. If this continues, these efforts may be able to require that intervention developers and testers submit applications containing all the information EGSs requires in predefined formats. If this trend does continue, it may be better to focus on developing application based EGSs than to focus on report-based (especially publication-based) ones that require a high degree of inference and often fail to obtain needed information.

Can the same EGSs be used with RCTs and observational studies and how confident can we be in observational studies no matter how well done?

Although it is under debate as to whether observational studies can control for selection and produce truly comparable study groups, it is becoming clearer that this may be more difficult than originally imagined (West et al., 2002). However, substantial opinion exists that observational research is the only type of research we currently have and possibly may be the only type of research we can have for certain interventions and problems. One implication of these two perspectives taken together is that in some cases we will have to make policy and clinical decisions based on observational data, but every step must be taken to make observational studies as rigorous as possible, applying the same criteria used for RCTs to consider as many sources of potential selection bias or other biases as possible (West et al., 2002). As noted above, several analysts have postulated that observational studies differ from RCTs in a number of ways related to evidentiary quality, not just with respect to assignment. Although the biases stemming from lack of random assignment may not be totally removable, if these other aspects of quality are made explicit and observational studies are held to high standards of evidence in these other respects, differences between RCTs and observational studies may be reduced.

Another implication is that while observational evidence may be more useful than no evidence, we should be as imaginative and creative as possible in developing high quality and randomized study methods and not simply assume that such studies are impossible. To quote Kunz and Oxman (1998):

“Observational studies often provide valuable information...However, it is important to remember that it is only possible to control for confounders that are known and measured in observational studies, and we should be wary of hubris and its consequences in assuming that we know all there is to know about any disease” (5-1188).

Or, as Kenneth Arrow (quoted in Bernstein, 1996, p.7) has warned:

“[O]ur knowledge of the way things work, in society or nature, comes trailing clouds of vagueness. Vast ills have followed a belief in certainty.”

How should decisions be made in the future about criteria selection and refinement and scoring approaches for categorizing interventions?

More studies of EGS criteria are needed and hopefully will occur. The creation of registries that record details of study designs and results should facilitate this. We need to know much more about issues such as the interrelationships among criteria, the associations between quality scores and effect size estimates, and the ways in which various methods of summing and using evidence quality scores alone and in combination with factors related to effect, length of follow-up, and replications work. We believe our lack of information at this time necessitates a more inclusive and experimental approach to criteria selection, retention and use. The risks of ignoring methodological elements that might contribute to quality seem greater than the risks of including ones that ultimately prove unnecessary.

How should EGSs be evaluated?

All but one of the scales and checklists reviewed by Moher et al. (1995) “evolved” with little or no attention to standard scale development techniques. All but one of the scales and checklists used criteria from standard clinical trial textbooks without reviewing relevant empirical evidence for the criteria. Scales also differed in the topics they covered and did not always present explicit rationales for the ones they included. Because of these and other shortcomings, Moher et al. (1995) recommend caution in assessing quality using any scale that has been inadequately developed (p. 68).

Developing scales to assess quality should be considered similar to developing any other instrument” (Moher et al., 1995). The items listed in Table 5, above begin to address the issues of scale development that should be addressed.

The issue of how to test EGS validity is an interesting one. It is relatively easy to imagine how to test EGS convergent validity and such tests have been reported (Jüni et al., 1999). It is more challenging to imagine how to test discriminant validity. This is because it is difficult to conceptualize how to decide what evidence is of high quality without using some type of “gold standard” EGS. If such a gold standard existed, it would obviate the need for further scale development.

The predictive validity of EGSs might be tested by replicating low quality studies in a high quality manner and making predictions about how effect estimates would differ. Based on the analyses cited above, we would definitely expect higher quality replications to show less heterogeneity and possibly smaller estimates of effect than the lower quality original studies. We might also do the reverse experiment of implementing lower quality replications of higher quality studies, although this might be ethically dubious.

How can EGSs be applied to the monitoring of interventions in routine clinical and long term use?

Most intervention evaluations involve relatively small samples and are of relatively short duration with follow-up extending less than two years. Nevertheless, an important aspect of how well interventions perform is their long term effects in widespread use. This is particularly true for adverse events which tend to be rare. Recent events in the pharmaceutical area have highlighted the importance of collecting higher quality data about interventions in routine clinical and long term use. Nevertheless, monitoring methods used currently are primarily observational, and much less attention has been paid to how to conceptualize and grade the quality of evidence from observations of interventions in such follow-up monitoring. EGSs in the future should pay more attention to these issues and do a better job of identifying interventions that are safe and effective in widespread and long term use.

How can international collaboration be developed?

Considerable work on EGSs is going on outside the United States of America. The reasons for international collaboration are many (West et al., 2002). Differences in EGSs confuse stakeholders and undermine the credibility of all EGSs. No country profits from reinventing EGS technology developed elsewhere. Joint efforts can speed the development of new EGS technology. The International Commission on Harmonization was initiated in recognition of these realities (Leff, 2005).

REFERENCES

- Alderson P, Green S, Higgins JPT, editors. Cochrane Reviewers' Handbook 4.2.2[updated March 2004]. <http://www.cochrane.org/resources/handbook/hbook.htm> (accessed 10th March 2005).
- American Psychological Association. Criteria for Evaluation Treatment Guidelines. August 2000. www.apa.org/practice/guidelines.html (accessed 10th March 2005).
- American Psychological Association. Treatment Guideline Checklist. www.apa.org/practice/guidelines/treatcrit.html (accessed 10th March 2005).
- Baruch, RF. Randomized Experiments for Planning and Evaluation: A Practical Guide. Sage, 1997.
- Bernstein, Reter L. Against the Gods: The Remarkable Story of Risk. New York: John Wiley & Sons, Inc. 1996
- Briss, P. A., Zaza, S., Pappaioanou, M., Fielding, J., Wright-De Agüero, L., Truman, B. I. et al. (2000). Developing an evidence-based guide to community preventive services—Methods. *American Journal of Preventive Medicine*, 8(1S), 35-43.
- The Campbell Collaboration. (2001). Campbell systematic reviews: Guidelines for the preparation of review protocols Version 1.0. www.campbellcollaboration.org
- Campbell, DT. and Stanely, JC. Experimental and Quasi-experimental Designs for Research. Boston: Houghton Mifflin Co. 1966
- The Centers for Disease Control, HIV/AIDS Prevention Research Synthesis Project. (1999). *Compendium of HIV prevention interventions with evidence of effectiveness*. Atlanta: Centers for Disease Control and Prevention.
- The Centers for Disease Control and Prevention (n.d.). "Guide to Community Preventive Services Overview." www.thecommunityguide.org/overview/default.htm.
- The Cochrane Collaboration (2004) "What is the Cochrane Collaboration." www.cochrane.org/docs/descrip.htm.
- Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovich, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7(27).
- Emerson, JD, Burdick, E, Hoaglin, DC, Mosteller, F, Chalmers, TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clinical Trials*. 1990 Oct. 11(5): 339-352.
- International Conference on Harmonization; Guidance on Statistical Principles for Clinical Trials; Availability, 63 Fed. Reg. 49,583 (Sep. 16, 1998).
- International Conference on Harmonisation. (n.d) <http://www.ich.org/>.
- Juni, Peter, Witschi, Anne, Bloch, Ralph, Egger, Matthias. The hazards of scoring the quality of clinical trials for meta-analyses. *JAMA*, 282(11) Sept. 1999. 1054-1060.

- Kunz, Regina, Andrew Oxman. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal*. Vol. 317, October, 1998. 1185-1190
- Leff, HS. Evidence in Intervention Science in RE Drake, MR Merrens, DW Lynde. Evidence-Based Mental Health Practice: A Textbook. WW Norton. In Press.
- Lipsey, Mark W. Wilson, David, B. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist* 1993 or after.
- Mihalic, S., Irwin, K., Elliott, D., Fagan, A., & Hansen, D. (2001). Blueprints for violence prevention. *Office of Juvenile Justice and Delinquency Prevention: Juvenile Justice Bulletin*, July 2001.
- Moher, David, Ba' Pham, Alison Jones, Deborah J. Cook, Alejandro R. Jadad, Michael Moher, Peter Tugwell, Terry P, Klassen. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses. *The Lancet*, Vol 352 August 22, 1998 pp609-613
- National Institute on Drug Abuse. (2003). *Preventing drug use among children and adolescents: A Research-based guide for parents, educators and community leaders*. (NIH Publication No. 04-4212(A), 2nd ed.) Washington, D.C.: U.S. Government Printing Office.
- Orne, M.T. (1959, September). The demand characteristics of an experimental design and their implications. Paper presented at *The Problem of Experimenter Bias*, symposium conducted at the 67th Annual Convention of the American Psychological Association, Cincinnati, Ohio.
- Redd, Z., Cochran, S., Hair, E., & Moore, K. (2002). *Academic achievement programs and youth development: A Synthesis*. Washington, D.C.: Child Trends.
- Rosenthal, Robert. Experimenter and clinician effects in scientific inquiry and clinical practice. *Prevention & Treatment*, Vol 5 October 18, 2002, pp1-12.
- Schulz, Kenneth, F, Chalmers, Iain, Richard, J., Altman, Douglas, G. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of American Medical Association*. Volume 273 (5). Feb. 1995; 408-412.
- Society for Prevention Research. (2004). *Standards of evidence: Criteria for efficacy, effectiveness and dissemination*. Falls Church, VA.
- Stokes, Donald E. Pasteur's Quadrant: Basic Science and Technological Innovation. Washington, DC: Brookings Institution Press, 1997
- The Substance Abuse and Mental Health Services Administration. (n.d.) "NREPP Overview." <http://modelprograms.samhsa.gov/template.cfm?page=nreppover>.
- The UK Cochrane Centre. The Cochrane Collaboration Methods Groups Newsletter. Volume 8, June, 2004
- U.S. Department of Education Safe, Disciplined, and Drug-Free Schools Expert Panel. (2001). *Exemplary & promising safe, disciplined, and drug-free schools programs*. Washington, D.C.: U.S. Department of Education.

U.S. Preventative Services Task Force. (n.d.) "Clinical Services."

www.ahrq.gov/clinic/cps3dix.htm.

West, Suzanne, King, Valerie, Carey, ;Timothy, S., Lohr, Kathleen, N., McKoy, Nikki, Sutton, Sonya F., Lux, Linda.. Systems to Rate the Strength of Scientific Evidence. Research Triangle Park, North Carolina: Research Triangle Institute-University of North Carolina Evidence-based Practice Center April, 2002

What Works Clearinghouse (n.d.). "Overview."

www.whatworks.ed.gov/whatwedo/overview.html.

Wolf, N. (2000). Using randomized controlled trials to evaluate socially complex services: Problems , challenges and recommendations. The Journal of Mental Health Policy and Economics, 3, 97-109.

APPENDIX A: Full Profile of Evidence Grading Systems Reviewed

Appendix A provides a complete profile of each evidence grading system that was reviewed. In addition to listing the full methodological criteria of each system, all other criteria and relevant elements, including a description of the systems purpose and review process, are present in this appendix. Finally, the names and descriptions of other documents related to the system and a website where they can be downloaded or ordered is provided.

Evidence Grading System: *Academic Achievement Programs and Youth Development: A Synthesis*

Developer: Child Trends

Description of Purpose and Scope:

This synthesis of research on academic achievement programs describes how such programs may help children and adolescents develop a broad array of strengths and abilities in the areas of education and cognitive attainment, health and safety, social and emotional well-being, and, as they move into adulthood, self-sufficiency. Specifically, the synthesis addresses the following questions: What do academic-oriented programs look like? What impacts do they have? What resources do they provide to promote healthy development? What positive outcomes are achievable through academic-oriented programs? What methods of implementation characterize effective programs? (Redd et al., 2002, p. v)

Review Process: Child Trends reviewers selected 12 programs with academic achievement as a main component. All selected programs had been previously evaluated using experimental (randomized control trials) and quasi-experimental (matched control trials) designs. Reviewers then examined the full methodology of each program's evaluation, treating programs that had been evaluated using experimental methods separately from those that had been evaluated using quasi-experimental methods. Programs that had been evaluated in experiments were separated into three categories on the basis of study findings "works," "doesn't work," "mixed reviews." Programs that had been evaluated using quasi-experimental methods were all considered as "best bets" and explored for "best-practice" type recommendations.

System Criteria and Elements: Child Trends does not provide a well-differentiated set of criteria that guided their classification of programs into evidence categories. However, they do provide a list of common program elements that derive from the exploration of programs that were evaluated using randomized control trials (Part II), as well as a few common outcomes elements found in quasi-experimental studies (Part III). The criteria below were extracted from the Child Trends report *Academic Achievement Programs and Youth Development: A Synthesis* by Redd et. al. (2002).

Part II. Outcomes Positively Affected by Academic Achievement Programs

Educational Achievement and Cognitive Attainment – Evidence shows that academic achievement programs can improve educational outcomes for young people who participate in them, although there is great variability across programs and outcomes.

Health and Safety – Studies indicate that academic-oriented programs which also aimed to improve health and safety were only sometimes effective at meeting this goal.

Social and Emotional Well-Being – Academic achievement programs showed some evidence of effectiveness at improving social and emotional well-being, although programs that specifically target this developmental outcome are more effective.

Self-Sufficiency – Academic achievement programs showed mixed effectiveness at improving self-sufficiency in young adulthood.

Part III. Program and Participant Characteristics Associated with Positive Outcomes

Quasi-experimental evidence from several studies suggests that young people who participate in programs longer and more intensively do better than those who participate for a shorter time or less intensively. Findings of one such study suggest that the development of close tutoring and mentoring relationships improves academic outcomes. Quasi-experimental studies also suggest that programs with a strong academic focus are more consistently effective at improving academic achievement outcomes.

Other Resources: Child Trends offers a full copy of this report, *Academic Achievement Programs and Youth Development: A Synthesis*, as well as other related publications at www.childtrends.org.

Evidence Grading System: *The Blueprints for Violence Prevention Initiative*

Developer: The Center for the Study and Prevention of Violence & Office of Juvenile Justice and Delinquency Prevention

Description of Purpose and Scope: *The Blueprints for Violence Prevention Initiative is a comprehensive effort to provide communities with a set of programs whose effectiveness has been scientifically demonstrated. With the Office of Juvenile Justice and Delinquency Prevention's (OJJDP's) support, the Initiative also provides the information necessary for communities to begin replicating programs locally* (Mihalic et al, 2001, p. 1).

Review Process:

System Criteria and Elements: Reviewers at the Center for the Study of Prevention of Violence at the University of Colorado, with the support of the Office of Juvenile Justice and Delinquency Prevention, selected 11 programs that met their methodological criteria.

- **Evidence of Deterrent Effect When Using a Strong Research Design:** The Blueprints advisory board accepts evidence of deterrent effects for three key indicators—violence (including childhood aggression and conduct disorder), drug use, and/or delinquency— as evidence of program effectiveness.
- **Sustained Effects:** Designation as a Blueprints program requires a sustained effect at least 1 year beyond treatment, with no subsequent evidence that this effect is lost.
- **Multiple Site Replication:** Replication establishes the strength of a program and its prevention effects and demonstrates that it can be successfully implemented in other sites.
- **Sample sizes must be large enough to provide statistical power to detect effects.** It is more difficult to detect statistically significant differences between groups when small sample sizes are used.
- **Attrition may be indicative of problems in program implementation or may be a failure to locate subjects during a follow-up period.** Attrition is dangerous, particularly because it can compromise the integrity of the original randomization or matching process. It reduces confidence that the original and final samples are comparable and that the final experimental and control comparisons reflect only treatment effects.
- **Tests to measure outcomes must be administered fairly, accurately, and consistently to all study participants.** For example, the use of inconsistent measures over time may produce less reliable test scores. The instruments used to measure outcomes should be demonstrated to be reliable and valid.
- **Analysis of mediating factors.** The Blueprints advisory board looks for evidence that change in the targeted risk or protective factor(s) mediates the change in violent behavior. This evidence clearly strengthens the claim that program participation is responsible for changes in violent behavior, and it contributes to the theoretical understanding of the causal processes involved. In its reviews of different programs, the advisory board has discovered that many programs reporting significant deterrent “main effects” have not collected the data necessary to complete an analysis of mediating factors.

Other Resources: Additional information on the *Blueprints* project can be found at www.colorado.edu/cspv/blueprints/.

Evidence Grading System: *Cochrane Reviewers' Handbook*

Developer: The Cochrane Collaboration

Description of Purpose and Scope: *The Cochrane Collaboration is an international non-profit and independent organisation, dedicated to making up-to-date, accurate information about the effects of healthcare readily available worldwide. It produces and disseminates systematic reviews of healthcare interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions.* (Cochrane, "What is," 2004)

Review Process: Cochrane reviewers complete a systematic review and meta-analysis of studies in a particular healthcare topic area. The review is carried out using a standardized protocol based on recommendations and guidelines from the *Cochrane Reviewers' Handbook*.

System Criteria and Elements: *The Cochrane Reviewers' Handbook* has a through set of recommendations for eliminating bias in systematic reviews and meta-analysis. However, listed below are only the elements of the *Handbook* that relate to grading the methodological quality of individual studies (Cochrane, 2004).

- Selection Bias (systematic differences in comparison groups)
 - Concealed Random Assignment
- Performance Bias (systematic differences in care provided apart from the intervention being evaluated)
- Attrition Bias (systematic differences in withdrawals from the trial)
- Detection Bias (systematic differences in outcomes assessment)
 - Subject Awareness of Assignment
 - Provider Awareness of Assignment
 - Collector Awareness of Assignment
- Contamination and Cointervention

Other Resources: The full text of the *Cochrane Reviewers' Handbook* along with the corollary *Glossary* can be obtained at www.cochrane.org/.

Evidence Grading System: Criteria for Evaluating Treatment Guidelines

Developer: American Psychological Association (APA)

Description of Purpose and Scope:

This document presents a set of criteria to be used in evaluating treatment guidelines that have been promulgated by health care organizations, government agencies, professional associations, or other entities. Although originally developed for mental health interventions, the criteria presented are equally applicable in other health service areas... This document is organized on the basis of two related dimensions for the evaluation of guidelines. The first dimension is treatment efficacy, the systematic and scientific evaluation of whether a treatment works. The second dimension is clinical utility, the applicability, feasibility, and usefulness of the intervention in the local or specific setting where it is to be offered. This dimension also includes determination of the generalizability of an intervention whose efficacy has been established (APA, 2002, p. 1052-53).

Review Process: The American Psychological Association does not expressly use these criteria for formalized reviews and recommendations of treatment guidelines. The criteria below were extracted from the American Psychological Association's *Criteria for Evaluating Treatment Guidelines* (2002).

System Criteria and Elements:

Treatment Efficacy

Criterion 1.0 Guidelines should be based on broad and careful consideration of the relevant empirical literature.

Criterion 2.0 Recommendations on specific interventions should take into consideration the level of methodological rigor and clinical sophistication of the research supporting the intervention.

Criterion 2.1 Guidelines consider clinical opinion, observation, and consensus among recognized experts representing the range of views in the field.

Criterion 2.2 Systematized clinical observation is weighted more heavily than unsystematized observation in evaluating treatment efficacy.

Criterion 2.3 The evaluation of treatment efficacy places greatest emphasis on evidence derived from sophisticated empirical methodologies, including quasi experiments and randomized controlled experiments or their logical equivalents.

Criterion 3.0 Recommendations on specific interventions should take into consideration the treatment conditions to which the intervention has been compared.

Criterion 3.1 Guidelines consider whether the treatment gets better results than doing nothing.

Criterion 3.2 Guidelines consider whether the intervention offers the patient any benefit beyond simply being in treatment.

Criterion 3.3 Guidelines consider whether an intervention's results are better than the results of other interventions.

Criterion 4.0 Guidelines should consider available evidence regarding patient–treatment matching.

Criterion 5.0 Guidelines should specify the outcomes the intervention is intended to produce, and evidence should be provided for each outcome.

1. Participant selection; 2. Treatment goals; 3. Quality of life, life functioning; 4. Attrition; 5. Long-term consequences of treatment; 6. Indirect consequences of treatment; 7. Patient satisfaction with treatment; 8. Iatrogenic negative effects or side effects of treatment; 9. Clinical significance; 10. Methods.; 11. Treatment goals.

Clinical Utility

Criterion 6.0 Guidelines should reflect the breadth of patient variables that may influence the clinical utility of the intervention.

Criterion 6.1 Guidelines take into account the complexity and idiosyncrasy of patients' clinical presentations, including severity, comorbidity, and external stressors.

Criterion 6.2 Guidelines take into consideration culturally relevant research and expertise.

Criterion 6.3 Guidelines take into consideration research addressing the issue of the patient's gender (a social characteristic) and sex (a biological characteristic).

Criterion 6.4 Guidelines take into account research and relevant clinical consensus concerning the age and developmental level of the patient.

Criterion 6.5 It is recommended that guidelines take into account research and clinical consensus on other relevant patient characteristics.

Criterion 7.0 It is recommended that guidelines take into account data on how differences between individual health care professionals may affect the efficacy of the treatment.

Criterion 7.1 It is recommended that guidelines take into account the effect of the health care professional's training, skill, and experience on treatment outcome.

Criterion 7.2 It is recommended that guidelines take into account the effects on treatment outcome of interactions between the patient's and the health care professional's characteristics, including but not limited to language, ethnicity, background, sex, and gender.

Criterion 8.0 It is recommended that guidelines take into account information pertaining to the setting in which the treatment is offered.

Criterion 9.0 Guidelines should take into account data on treatment robustness.

Criterion 10.0 Guidelines should take into account the intervention's level of acceptability to the patients who are to receive the service.

Criterion 10.1 Guidelines provide for informed patient choice among comparable interventions.

Criterion 10.2 Guidelines consider patients' willingness and ability to participate in recommended interventions.

Criterion 11.0 Guidelines should explicitly note and evaluate possible adverse effects of interventions as well as their benefits.

Criterion 12.0 Guidelines should address the preparation of the health care professionals to deliver the intervention.

Criterion 13.0 When guidelines include consideration of costs, it should be reported separately from consideration of effectiveness.

Criterion 14.0 When guidelines consider costs, they should consider the direct, indirect, short-term, and long-term costs to the patient, to the professional, and to the health care system, as well as the costs associated with withholding treatment.

Other Resources: As a companion to the *Criteria for Evaluating Treatment Guidelines*, the APA has a *Treatment Guideline Checklist* available at www.apa.org/practice/guidelines/.

Evidence Grading System: *The Guide to Community Preventive Services*

Developer: The U.S. Preventative Services Task Force

Description of Purpose and Scope: *The U.S. Preventive Services Task Force (USPSTF) was convened by the Public Health Service to rigorously evaluate clinical research in order to assess the merits of preventive measures, including screening tests, counseling, immunizations, and preventive medications. (USPSTF, "Clinical," n.d.)*

Review Process: The USPSTF conducts reviews in the following steps: (1) *scope and selection of topics*, (2) *review of the evidence*, (3) *assessing the magnitude of net benefit*, (4) *extrapolation and generalization*, (5) *translating evidence into recommendations*, (6) *drafting the report*, and (7) *external review*. (Harris et al, 2001, p. M-16)

System Criteria and Elements: While the USPSTF does not provide a list of numbered "criteria" per se, the list below provides the criteria and sub-criteria elements that are used to evaluate study quality by reviewers. These criteria were extracted from the Russell et al paper (2001).

For systematic reviews

- Standard appraisal of included studies
- Validity of conclusions
- Recency and relevance
- Comprehensiveness of sources/search strategy used

For case-control studies

- Accurate ascertainment of cases
- Nonbiased selection of cases/controls with exclusion criteria applied equally to both
- Response rate
- Diagnostic testing procedures applied equally to each group
- Appropriate attention to potential confounding variables

For randomized controlled trials

- Initial assembly of comparable groups: For RCTs: adequate randomization, including concealment and whether potential confounders were distributed equally among groups; For cohort studies: consideration of potential confounders with either restriction or measurement for adjustment in the analysis; consideration of inception cohorts
- Maintenance of comparable groups (includes attrition, crossovers, adherence, contamination)
- Important differential loss to follow-up or overall high loss to follow-up
- Measurements: equal, reliable, and valid (includes masking of outcome assessment)
- Clear definition of interventions
- All important outcomes considered
- Analysis: adjustment for potential confounders for cohort studies, or intention-to-treat analysis for RCTs

For diagnostic accuracy studies

- Screening test relevant, available for primary care, adequately described
- Study uses a credible reference standard, performed regardless of test results
- Reference standard interpreted independently of screening test
- Handles indeterminate results in a reasonable manner
- Sample size
- Administration of reliable screening

Other Resources: Current USPSTF reviews can be found at www.ahrq.gov/clinic/cps3dix.htm.

Evidence Grading System: *The Guide to Community Preventive Services*

Developer: The Centers for Disease Control and Prevention

Description of Purpose and Scope: The Guide to Community Preventive Services (Community Guide) serves as a filter for scientific literature on specific health problems that can be large, inconsistent, uneven in quality, and even inaccessible. The Community Guide summarizes what is known about the effectiveness, economic efficiency, and feasibility of interventions to promote community health and prevent disease. The Task Force on Community Preventive Services makes recommendations for the use of various interventions based on the evidence gathered in the rigorous and systematic scientific reviews of published studies conducted by the review teams of the Community Guide. (CDC, "Overview," n.d.)

Review Process: Systematic reviews are conducted by the Task Force on Community Preventive Services in the following manner: *The steps for obtaining and evaluating evidence into recommendations involve: (1) forming multidisciplinary chapter development teams, (2) developing a conceptual approach to organizing, grouping, selecting and evaluating the interventions in each chapter; (3) selecting interventions to be evaluated; (4) searching for and retrieving evidence; (5) assessing the quality of and summarizing the body of evidence of effectiveness; (6) translating the body of evidence of effectiveness into recommendations; (7) considering information on evidence other than effectiveness; and (8) identifying and summarizing research gaps.* (Briss et al., 2000, p. 35)

System Criteria and Elements: While the *Community Guide* does not provide a list of numbered "criteria" per se, the list below provides the criteria and sub-criteria elements that are used to evaluate study quality by Task Force reviewers. These criteria have been extracted from several *Community Guide Documents*, including their *Data Abstraction Form* and the paper by Briss et al. (2000).

- Study Design
- Populations Description
- Intervention Description
- Specified Sampling Frame or Universe of Selection
- Specified Screening Criteria
- Was the population that served as the unit of analysis the entire eligible population or a probability sample at the point of observation?
- Other Selection Bias Issues
- Was exposure measured?
 - Exposure measure validity
 - Exposure measure reliability
- Outcome measure validity
- Outcome Measure reliability
- Conducted statistical testing (when appropriate)?
- Reported which statistical tests were used?
- Controlled for design effects in the statistical model?
- Controlled for repeated measures in the analysis, for study designs in which the same population was followed with repeated measurements over time?
- Accounting for different levels of exposure in segments of the study population in the analysis?
- If the authors analyzed group-level and individual-level covariates in the same statistical model, was the model designed to handle multi-level data?
- Attrition Bias
- Were confounding variables assessed prior to intervention exposure?
- Control for confounding variables
- Potential Biases
- Other limitations on interpretation

Other Resources: Current reviews and recommendations as well as data abstraction forms used in the systematic reviews can be found at www.thecommunityguide.org/default.htm.

Evidence-Grading System: *Guidelines for the Preparation of Review Protocols*

Developer: The Campbell Collaboration

Description of Purpose and Scope: *A Campbell Systematic Review is meant to review and synthesize evidence on social and behavioural interventions and public policy, including education, criminal justice, and social welfare, among other areas. The primary concern is with evidence on overall intervention or policy effectiveness and how effectiveness is influenced by variations in process and implementation, intervention components and recipients, as well as other factors* (Campbell, 2000, p. 1).

Review Process: Campbell reviewers complete a systematic review and meta-analysis of studies related to a particular topic area. The review is carried out using a standardized protocol based on the *Guidelines for the Preparation of Review Protocols*.

System Criteria and Elements: While Campbell has a well-articulated manual for the execution of systematic reviews, there is no explicit information on how the quality of individual studies is evaluated. Since Campbell operates an RCT database and allows reviewers to reference other systems that are similar to Campbell (e.g. Cochrane) when developing protocols, it is likely that study quality is graded by Campbell similarly to how it is graded in comparable systems.

Other Resources: Campbell Reviews and trial registries can be found at www.campbellcollaboration.org.

Evidence-Grading System: *HIV/AIDS Prevention Research Synthesis Project*

Developer: The Centers for Disease Control and Prevention

Description of Purpose and Scope: *The CDC's PRS project aims to conduct systematic reviews that address the population, intervention, study design, setting, and outcome factors associated with intervention effectiveness, to identify methodologically rigorous studies that have statistically significant positive results, and to identify gaps in the existing research and directions for future study. (CDC PRS, 1999, p. 4-1)*

Review Process: All relevant HIV prevention studies from 1988 onward will be reviewed and cataloged by the PRS project. Studies that were conducted in the United States and that meet both relevance and effectiveness criteria will be added to the *Compendium of HIV Prevention Interventions with Evidence of Effectiveness*.

System Criteria and Elements: These criteria were extracted from the PRS publication *Compendium of HIV Interventions with Evidence of Effectiveness* (1999).

PRS Criteria

1. Criteria for Scope: Select studies that focus on HIV prevention interventions and that are not being studied extensively elsewhere.
 - A.) HIV prevention focus
 - B.) Reported from 1988 onward
 - C.) Published or unpublished
 - D.) Conducted inside or outside the U.S.
 - E.) Not drug treatment only
 - F.) Not biomedical only, (e.g., vaccine trials, AZT to prevent perinatal transmission)
 - G.) Not occupational exposure
 - H.) Not blood supply exposure

2. Criteria for Relevance: Select studies that aim to reduce sex- or drug-related risk behaviors or incidence rates of HIV or other STDs. PRS has determined that relevant outcomes are those that directly impact the transmission of HIV or are indicators of HIV transmission. A relevant outcome must include one or more of the following outcomes:

- A.) Sex-related behaviors: use of male condoms, use of female condoms, use of condom negotiation, not having sex, if condom not used, having unprotected sex, number of sex partners, mutually monogamous relationship, partner selection, return to abstinence, initiation of first sexual intercourse, exchanging sex for money/drugs
- B.) Drug-related behaviors: multi-person use of drug paraphernalia, cleaning/bleaching drug paraphernalia, use of new sterile needles/syringes, injecting drugs, initiation of drug injection, non-injecting drug use, sex with substance use, return of used syringes.
- C.) HIV testing behavior: being tested, learning test results, repeat testing.
- D.) Health Outcomes: incidence rate of HIV, AIDS, STDs, HBV, or HCV, prevalence rate of HIV, AIDS, STDs, HBV, or HCV.

3. Criteria for Methodological Rigor: All relevant studies are evaluated for methodological rigor, regardless of the findings (i.e., studies with negative or null findings are included). These criteria are based on study design and vary by intervention category. Behavioral and social intervention studies are classified as methodologically rigorous if they used random assignment to intervention and control groups (experimental designs) and reported at least post-intervention data. Behavioral and social studies are also considered rigorous if they used non-biased assignment (e.g., systematic assignment) to intervention and comparison groups (quasi-experimental designs) with equivalence of groups or used statistical adjustment for any nonequivalence, and reported pre- and post-intervention data. Policy interventions used these designs or less rigorous designs, such as designs with pre-post data but without a comparison group.

A.) For *behavioral and social* intervention studies: Random assignment to intervention and comparison groups *WITH* post-intervention data. Non-random assignment to intervention and comparison groups using a non-biased method *WITH* pre-post data *AND* equivalence of groups *OR* adjustment for apparent non-equivalence of groups.

B.) For *policy* studies: Random assignment to intervention and comparison groups *WITH* post-intervention data. Non-random assignment to intervention and comparison groups using a non-biased method, *WITH* post-intervention data *AND* apparent equivalence of groups *OR* adjustment for apparent non-equivalence of groups. *OR* Pre-post data with no comparison group.

Other Resources: The complete *Compendium of HIV Prevention Interventions with Evidence of Effectiveness* as well as the an intervention checklist that allows organizations to evaluate their own HIV prevention programs can be found at <http://www.cdc.gov/hiv/pubs/hivcompendium/>.

Evidence-Grading System: *International Conference on Harmonisation*

Developer: EU, EFPIA, MHLW, JPMA, FDA, and PhRMA

Description of Purpose and Scope: *The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) is a unique project that brings together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration.*

The purpose is to make recommendations on ways to achieve greater harmonisation in the interpretation and application of technical guidelines and requirements for product registration in order to reduce or obviate the need to duplicate the testing carried out during the research and development of new medicines.

The objective of such harmonisation is a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines whilst maintaining safeguards on quality, safety and efficacy, and regulatory obligations to protect public health. (ICH, n.d.)

Review Process: The International Conference on Harmonisation does not conduct reviews, but is advised by and in turn, advises the regulatory bodies of its member countries.

System Criteria and Elements: These criteria have been extracted from the ICH document that was published in the Federal Register of the United States (Federal Register, 1998).

- The primary objectives of any study should be clear and explicitly stated.
- The methods of monitoring adverse events by changes in clinical signs and symptoms and laboratory studies should be described.
- The protocol should specify procedures for the follow-up of patients who stop treatment prematurely.
- Selection of control group – Trials should have an adequate control group. Comparisons may be made with placebo, no treatment, active controls, or of different doses of the drug under investigation.
- Randomization - In conducting a controlled trial, randomized allocation is the preferred means of assuring comparability of test groups and minimizing the possibility of selection bias. No one involved in the conduct of the trial is aware of the specific treatment allocated to any particular subject, not even as a code.
- Blinding is an important means of reducing or minimizing the risk of biased study outcomes. A trial where the treatment assignment is not known by the study participant because of the use of placebo or other methods of masking the intervention is referred to as a single blind study. When the investigator and sponsor staff who are involved in the treatment or clinical evaluation of the subjects and analysis of data are also unaware of the treatment assignments, the study is double blind.
- Methods used to evaluate patient usage of the test drug should be specified in the protocol and the actual usage documented.
- Timely adverse event reporting during a study is essential and should be documented.
- The study protocol should have a specified analysis plan that is appropriate for the objectives and design of the study, taking into account the method of subject allocation, the measurement methods of response variables, specific hypotheses to be tested, and analytical approaches to common problems including early study withdrawal and protocol violations.
- Safety data should be collected for all clinical trials, appropriately tabulated and with adverse events classified according to their seriousness and their likely causal relationship
- The primary variable ("target" variable, primary endpoint) should be the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial. There should generally be only one primary variable.
- The use of a reliable and validated variable with which experience has been gained either in earlier studies or in published literature is recommended.
- To avoid multiplicity concerns arising from post hoc definitions, it is critical to specify in the protocol the precise definition of the primary variable as it will be used in the statistical analysis. When the clinical effect defined by the primary objective is to be measured in more than one way, the protocol should identify one of the measurements as the primary variable on the basis of clinical relevance, importance, objectivity, and/or other relevant characteristics, whenever such selection is feasible. This approach addresses the multiplicity problem without requiring adjustment to the Type I error.

- Whatever data capture instrument is used, the form and content of the information collected should be in full accordance with the protocol and should be established in advance of the conduct of the clinical trial.
- For the purpose of overseeing the quality of the trial, the checks involved in trial monitoring may include whether the protocol is being followed, the acceptability of data being accrued, the success of planned accrual targets, the appropriateness of the design assumptions, success in keeping patients in the trials, and so on.
- Inclusion and exclusion criteria should remain constant, as specified in the protocol, throughout the period of subject recruitment.
- When designing a clinical trial, the principal features of the eventual statistical analysis of the data should be described in the statistical section of the protocol.
- The intention-to-treat principle implies that the primary analysis should include all randomized subjects.
- The methods of dealing with missing values are sensible, particularly if those methods are predefined in the protocol.
- Multiplicity may arise, for example, from multiple primary variables, multiple comparisons of treatments, repeated evaluation over time, and/or interim analyses (see section 4.5). Any aspects of multiplicity that remain after steps of this kind have been taken should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.

Other Resources: Additional information on ICH members and activities can be found at www.ich.org/.

Evidence Grading System: *National Registry of Evidence-Based Programs and Practices*

Developer: Substance Abuse and Mental Health Services Administration

Description of Purpose and Scope: The National Registry of Evidence-based Programs and Practices (NREPP) is a voluntary rating and classification system for mental health and substance abuse prevention and treatment interventions. The system is designed to categorize and disseminate information about programs and practices that meet established evidence rating criteria. SAMHSA is committed to making NREPP a leading national resource for contemporary and reliable information on the scientific basis and practicality of interventions to prevent and/or treat mental and addictive disorders. (SAMHSA, "Overview" n.d.)

Review Process: Candidate programs and practices will work with a Review Coordinator to ensure that their application is complete. The Review Coordinator will determine if and when the application is ready for formal review by three independent scientist-reviewers. All programs and practices with a strong scientific basis (as determined by the scientist-reviewers) will be included in NREPP, but will also undergo a second review by a panel of stakeholders who will assess the utility and practicality of the intervention (SAMHSA, "Overview," n.d.)

System Criteria and Elements: The criteria below were extracted from SAMHSA's NREPP website (SAMHSA, n.d.).

The 19 criteria applied to each program-identified outcome measure are:

1. Alpha Inflation: If there are multiple measures of the same outcome, p values for these measures should be adjusted for alpha inflation or "capitalizing on chance." If an outcome has only one measure, no adjustment is necessary, and the p level is as accurate as any appropriately adjusted measures
2. A Priori Identification of Outcomes: Outcome measures should be identified prior to the implementation of a study to avoid capitalizing on chance. Ideally, this should be determined from a document written before study implementation. The use of a measure in a pretest can be considered evidence that the measure was selected a priori. Some studies will select portions of an instrument as an outcome measure, and in these cases there must be evidence that the specific measure was identified a priori, not the instrument as a whole.
3. Reliability: 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." Studies done by investigators other than the applicant that show acceptable statistically measured reliability are considered the most convincing evidence of reliability.
4. Validity: 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." Studies done by investigators other than the applicant that show statistically measured validity are considered the most convincing evidence of validity.
5. Intervention Fidelity: The "experimental" intervention implemented in a study should have fidelity to the intervention proposed by the applicant. Instruments that have tested acceptable psychometric properties (for example, inter-rater reliability, validity as shown by positive association with outcomes) provide the highest level of evidence.
6. Comparison Fidelity: A study's comparison condition should be implemented with fidelity to the comparison condition proposed by the applicant. Instruments for measuring fidelity that have tested acceptable psychometric properties (for example, inter-rater reliability, validity as shown by predicted association with outcomes) provide the highest level of evidence. Fidelity measures for placebo and no service conditions should examine any harmful ingredients that would allow the experimental condition to outperform these controls by being "the lesser of two evils."
7. Nature of Comparison Condition: Interventions should be tested against nonspecific control designs, commonly known as psychosocial placebo controls², or against no service controls, unless interventions with proven effectiveness with respect to the outcome under study already exist. In order to validly test the effectiveness of an intervention, nonspecific factors must be held constant between intervention and comparison groups.
8. Assurances to Participants: Study participants should always be assured that their responses will be kept confidential and not affect their care or services. When these procedures are in place, participants are more likely to disclose valid data. Data not collected directly from participants do not raise this issue.
9. Participant Awareness of Condition: Participants can be biased by how an intervention is introduced to them and by an awareness of their study condition. Although some participant biases or expectancies may be unavoidable as part of the consenting process, information used to recruit and inform study participants should be carefully crafted to equalize expectations. Masking treatment conditions as much as possible during implementation of the study provides the strongest control for participant expectancies. When intervention conditions cannot be completely masked, expectations from study conditions should be controlled by statistical methods.

10. Standardized Data Collection: All outcome data should be collected in a standardized manner. Data collectors trained in and monitored for adherence to standardized protocols provide the highest quality evidence of standardized data collection.

11. Data Collector Bias: Data collector bias is most strongly controlled when data collectors are not aware of the conditions to which study participants have been assigned. When data collectors know specific intervention conditions, expectations from awareness of study conditions should be controlled by training and statistical methods.

12. Selection Bias: Random assignment of participants provides the strongest evidence of control for selection bias. When participants are not randomly assigned, covariates and confounding variables should be controlled as indicated by theory and research.

13. Attrition: Study results can be biased by participant attrition. Statistical methods as supported by theory and research can be employed to control for attrition that would bias results, but studies with no attrition needing adjustment provide the strongest evidence that results are not biased.

14. Missing Data: Study results can be biased by missing data. Statistical methods as supported by theory and research can be employed to control for missing data that would bias results, but studies with no missing data needing adjustment provide the strongest evidence.

15. Analysis Meets Data Assumptions: The methods used to analyze the data for each outcome measure should be consistent with the study design and type of data collected.

16. A Priori Identification of Methods: Analytic methods should be identified prior to inspection or analysis of the data to avoid capitalizing on chance. If this practice is not adequately conveyed through reports or published articles, applicants should submit documents such as proposals and IRB submissions. Analyses done in prior exploratory studies of an intervention can be taken as evidence that methods were selected a priori.

17. Analysis Consistent with Study Theory: The methods used to analyze the data for each outcome measure should be consistent with the theory and hypotheses underlying the intervention or program.

18. Anomalous Findings: Anomalous findings that contradict or would not be expected, given intervention theory, suggest the possibility of confounding causal variables. Anomalous findings are suggestive of negative evidence for the validity of an intervention. Studies with no anomalous findings, given intervention theory, provide the strongest evidence.

19. Replications: Replications of findings from additional studies of the same intervention done by independent investigators provide the strongest confidence of the effectiveness of an intervention. Evidence about replications can be provided in two ways. Applicants can refer to other studies in narrative and summary ways in introductions and literature reviews in reports and articles. They can also submit data from multiple studies. This criterion can be applied to both types of evidence. Data from multiple studies also are taken into account when multiple measures within and across studies are synthesized and in the NREPP decision tree.

Other Resources: The full, Likert-type scale items used to rate each NREPP criterion as well as more information about the review process can be found at <http://modelprograms.samhsa.gov/template.cfm?page=nreppover>.

Evidence-Grading System: *Preventing Drug Use Among Children and Adolescents*

Developer: National Institute on Drug Abuse

Description of Purpose and Scope: *One of the goals of the National Institute on Drug Abuse (NIDA) is to help the public understand the causes of drug abuse and to prevent its onset...NIDA hopes that this revised guide is helpful to drug abuse prevention efforts among children and adolescents in homes, schools, and communities nationwide (NIDA, 2003, p. v-1)*

Review Process: The NIDA describes its selection of example programs to include in the guide *Preventing Drug Use Among Children and Adolescents* as follows: *Each program was developed as part of a research protocol in which an intervention group and a comparison group were matched on important characteristics, such as age, grade in school, parents' level of education, family income, community size, and risk and protective factors. The interventions were tested in a family, school, or community setting, all with positive results. Prevention research continues to identify effective programs and strategies, thus this list is not meant to be exhaustive (2003, p. 26).*

System Criteria and Elements: These criteria were extracted from the NIDA guide *Preventing Drug Use Among Children and Adolescents* (2003).

Recommendations for Community Program Evaluators

- using control or comparison groups who did not receive the intervention, but whose characteristics are similar to those who did receive it
- monitoring the quality of program implementation
- using tested data-collection instruments
- ensuring that post-intervention follow-up includes a large percentage of the target population
- using appropriate statistical methods to analyze the data.

Prevention Principles

Principle 1 Prevention programs should enhance protective factors and reverse or reduce risk factors.

Principle 2 Prevention programs should address all forms of drug abuse, alone or in combination, including the underage use of legal drugs (e.g., tobacco or alcohol); the use of illegal drugs (e.g., marijuana or heroin); and the inappropriate use of legally obtained substances (e.g., inhalants), prescription medications, or over-the-counter drugs.

Principle 3 Prevention programs should address the type of drug abuse problem in the local community, target modifiable risk factors, and strengthen identified protective factors.

Principle 4 Prevention programs should be tailored to address risks specific to population or audience characteristics, such as age, gender, and ethnicity, to improve program effectiveness.

Principle 5 Family-based prevention programs should enhance family bonding and relationships and include parenting skills; practice in developing, discussing, and enforcing family policies on substance abuse; and training in drug education and information.

Principle 6 Prevention programs can be designed to intervene as early as preschool to address risk factors for drug abuse, such as aggressive behavior, poor social skills, and academic difficulties.

Principle 7 Prevention programs for *elementary school children* should target improving academic and social-emotional learning to address risk factors for drug abuse, such as early aggression, academic failure, and school dropout.

Principle 8 Prevention programs for middle or junior high and high school students should increase academic and social competence.

Principle 9 Prevention programs aimed at general populations at key transition points, such as the transition to middle school, can produce beneficial effects even among high-risk families and children. Such interventions do not single out risk populations and, therefore, reduce labeling and promote bonding to school and community.

Principle 10 Community prevention programs that combine two or more effective programs, such as family-based and school-based programs, can be more effective than a single program alone.

Principle 11 Community prevention programs reaching populations in multiple settings—for example, schools, clubs, faith-based organizations, and the media—are most effective when they present consistent, community-wide messages in each setting.

Principle 12 When communities adapt programs to match their needs, community norms, or differing cultural requirements, they should retain core elements of the original research-based intervention.

Principle 13 Prevention programs should be long-term with repeated interventions (i.e., booster programs) to reinforce the original prevention goals. Research shows that the benefits from middle school prevention programs diminish without follow-up programs in high school.

Principle 14 Prevention programs should include teacher training in good classroom management practices, such as rewarding appropriate student behavior. Such techniques help to foster student's positive behavior, achievement, academic motivation, and school bonding.

Principle 15 Prevention programs are most effective when they employ interactive techniques, such as peer discussion groups and parent role-playing, that allow for active involvement in learning about drug abuse and reinforcing skills.

Principle 16 Research-based prevention programs can be cost-effective. Similar to earlier research, recent research shows that for each dollar invested in prevention, a savings of up to \$10 in treatment for alcohol or other substance abuse can be seen.

Other Resources: The complete guide, *Preventing Drug Use Among Children and Adolescents* and further information on NIDA research and methods are available at www.drugabuse.gov.

Evidence-Grading System: The Safe and Drug-Free Schools Program

Developer: U.S. Department of Education

Description of Purpose and Scope: *In 1994, Congress directed the Office of Educational Research and Improvement (OERI), U.S. Department of Education, to establish “panels of appropriate qualified experts and practitioners” to evaluate educational programs and recommend to the Secretary of Education those programs that should be designated as exemplary or promising. Under the Education, Research, Development, Dissemination, and Improvement Act of 1994, each panel, in making this recommendation, was directed to consider 1) whether based on empirical data a program was effective and should be designated as exemplary or 2) whether there was sufficient evidence to demonstrate that the program showed promise for improving student achievement and should be designated as promising. The purpose of these panels was and still is to provide teachers, administrators, policymakers, and parents with solid information on the quality and effectiveness of programs and materials so that they can make better-informed decisions in their efforts to improve the quality of student learning (Department of Education, 2001, p. 1)*

Review Process: The SDFS program accepted applications from programs that considered themselves appropriate for evaluation. A committee of reviewers then evaluated the submitted programs in two stages. The first stage evaluated program efficacy. Those programs that had high efficacy scores were then classified by reviewers as either *exemplary* or *promising*. *Exemplary* programs were then reviewed by an *Impacts Panel* that further described program elements.

System Criteria and Elements: These criteria have been extracted from the *Exemplary and Promising Safe, Disciplined, and Drug Free Schools Program* publication (Department of Education, 2001).

A. EVIDENCE OF EFFICACY

- *Criterion 1 The program reports relevant evidence of efficacy/effectiveness based on a methodologically sound evaluation.*
 - Condition a. The program evaluation indicates a measurable difference in outcomes that is based on statistical significance testing or a credible indicator of magnitude of effect. Relevant outcomes are factors related to making schools safe, disciplined, and drug-free: a reduction in substance use, violence, and other conduct problems and positive changes in scientifically established risk and protective factors for these problems.
 - Condition b. The program evaluation used a design and analysis that adequately controls for threats to internal validity, including attrition.
 - Condition c. The program evaluation used reliable and valid outcome measures.
 - Condition d. The program evaluation used analyses appropriate to the data.

B. QUALITY OF PROGRAM

- *Criterion 2 The program's goals with respect to changing behavior and/or risk and protective factors are clear and appropriate for the intended population and setting.*
 - Condition a. The program's goals are explicit and clearly stated.
 - Condition b. The program's goals are appropriate to the intended population and setting.
- *Criterion 3 The rationale underlying the program is clearly stated, and the program's content and processes are aligned with its goals.*
 - Condition a. The rationale (e.g., logic model, theory) underlying the program is clearly stated and includes appropriate documentation (e.g., literature reviews and previous research).
 - Condition b. The program's content and processes are aligned with its goals.
- *Criterion 4 The program's content takes into consideration the characteristics of the intended population and setting (e.g., developmental stage, motivational status, language, disabilities, culture) and the needs implied by these characteristics.*
- *Criterion 5 The program implementation process effectively engages the intended population.*
 - Condition a. The program provides a relevant rationale to participants for its implementation.
 - Condition b. The program actively engages the intended population.
 - Condition c. The program attends to participants' prior knowledge, attitudes, and commonly held conceptions.
 - Condition d. The program implementation methods promote participants' collaboration, discourse, and reflection.

Where applicable:

- Condition e. The methods foster the use and application of skills.

- Condition f. The program promotes multiple approaches to learning.

C. EDUCATIONAL SIGNIFICANCE

- *Criterion 6 The application describes how the program is integrated into schools' educational missions.*

D. USEFULNESS TO OTHERS

- *Criterion 7 The program provides necessary information and guidance for replication in other appropriate settings.*
 - Condition a. The program clearly outlines the essential conditions required to replicate it with fidelity in other settings (e.g., strategies, resources, implementation plans, and materials).
 - Condition b. The program includes guidelines and materials for training and supporting those who are to replicate it.

Other Resources: More information about the *Safe and Drug-Free Schools Program* can be found at www.ed.gov.

Evidence Grading System: *Standards of Evidence*

Developer: Society for Prevention Research

Description of Purpose and Scope: *We hope...to provide a set of shared standards to be used by diverse organizations seeking to identify tested and effective prevention programs worthy of replication, adoption or dissemination. We believe that the promulgation and widespread use of these criteria will lead to consistent and high standards for determining whether programs have been scientifically shown to be efficacious, effective or ready for dissemination (SPR, 2003, p. i)*

Review Process: The Society for Prevention Research does not use the *Standards of Evidence* guidelines to conduct reviews.

System Criteria and Elements: Criteria that are recommended but not required by SPR to prove efficacy, effectiveness, or disseminability claim are not listed below. These criteria were extracted from the SPR *Standards of Evidence* publication (2003).

Efficacy Criteria

1. Specificity of Efficacy Statement – Program X is efficacious for producing Y outcomes for Z population.
2. Intervention Description and Outcomes
 - a. Program or policy description Program or policy description – The intervention must be described at a level that would allow others to implement/replicate it (this may require a more detailed description than what is presented in most research journals).
 - b. Outcomes – What is measured?
 - i. The stated public health or behavioral outcome(s) of the intervention must be measured. For example, a measure of attitudes about violence cannot substitute for a “measure” of actual violent behavior.
 - ii. For outcomes that may decay over time, there must be at least one long-term follow-up at an appropriate interval beyond the end of the intervention (e.g., at least 6 months after the intervention, but the most appropriate interval may be different for different kinds of interventions).
 - c. Outcomes - Measurement properties
 - i. Construct validity – Valid measures of the targeted behavior must be used, following standard definitions within the appropriate related literature.
 - ii. Reliability Internal consistency (alpha), test-retest reliability, and/or reliability across raters must be reported.
 - iii. Where “demand characteristics” are plausible, there must be at least one form of data (measure) that is collected by people different from the people applying or delivering the intervention.
3. Clarity of Causal Inference – The design must allow for unambiguous causal statements.
 - a. Comparison Condition – The design must have at least one comparison condition that does not receive the tested intervention.
 - b. Assignment – The assignment to conditions needs to be done in such a way as to maximize confidence that the intervention, rather than some other alternative explanation, causes the reported outcomes.
 - i. For most kinds of interventions, random assignment (of sufficient sample size without significant pretest differences) is essential.
 - ii. For some kinds of large-scale interventions (e.g., policy interventions, wholestate interventions) where randomization is not practical or possible, repeated time-series designs without randomization can be convincing given large effects and long baselines
 - iii. Well-conducted regression-discontinuity designs also can be convincing because as in randomized studies, the selection model is completely known.
 - iv. Matched control designs are credible only with demonstrated pretest equivalence using adequately powered tests on multiple baselines or pretests of multiple outcomes and important covariates, and as long as assignment was not by self-selection, but instead is by some other factor (e.g., geography, every 2nd case, or all sites applying for a service every alternate month).
4. Generalizability of Findings
 - a. Sample is defined – The report must specify what/who the sample is and how it was obtained.
5. Precision of Outcome
 - a. Statistical analysis must allow us to unambiguously establish the causal relations between the intervention and the outcomes (main effects).
 - i. In testing main effects, the analysis must be at the same level as the randomization and include all cases assigned to treatment and control conditions (except for attrition, see below).
 - ii. Test for pretest differences.
 - iii. When multiple outcomes are analyzed, there must be adjustment for multiple comparisons, (i.e., correction of the experiment-wise (Type I) error).

- iv. Analyses to minimize the possibility that observed effects are significantly biased by differential measurement attrition, which occurs when the characteristics of participants that are not available for the post-test are not equally distributed across study groups, are essential.
- b. Statistically significant effects
 - i. Results must be reported for every measured outcome, regardless of whether they are positive, non-significant or negative.
 - ii. Efficacy can be claimed only for constructs with a consistent pattern of statistically significant positive effects.
 - iii. For an efficacy claim, there must be no negative (iatrogenic) effects on important outcomes.
- c. Practical value – It is necessary to demonstrate practical significance in terms of public health impact.
- d. Duration of effect – In general, for outcomes that may decay over time, there must be a report of significant effects for at least one long-term follow-up at an appropriate interval beyond the end of the intervention (e.g., at least 6 months).
- e. Replication
 - i. Consistent findings are required from at least two different high-quality studies/replicates that meet all of the above criteria and each of which has adequate statistical power. Recognizing the importance of the replication standards, we note that flexibility may be required in the application of this standard for some kinds of interventions until enough time passes to allow the research enterprise to meet this high standard.
 - ii. When more than two efficacy and effectiveness studies are available, the preponderance of evidence must be consistent with that from the two studies of highest quality.

Criteria for Effectiveness

1. To claim effectiveness, studies must meet all of the conditions of efficacy trials plus the following.
2. Program Description and Outcomes
 - a. Program definition - Manuals and, as appropriate, training and technical support must be readily available.
 - b. Intervention delivery - The intervention should be delivered under the same types of conditions as one would expect in the real world (e.g., by teachers rather than research staff).
 - c. Theory
 - i. A clear theory of causal mechanisms should be stated.
 - ii. A clear statement of “for whom?” and “under what conditions?” the intervention is expected to be effective should be made.
 - d. Measures - Level of exposure should be measured, where appropriate, in both treatment and control conditions.
 - i. Integrity and level of implementation/delivery of intervention.
 - ii. Acceptance/compliance/adherence/involvement of target audience and subgroups of interest in the intervention activities.
3. Clarity of Causal Inference - The same standards as stated for efficacy apply, though the challenges are greater. Randomization is still the best approach, but the other alternatives suggested (regression discontinuity, time-series, high quality matched controlled designs) may be used.
4. Generalizability of Findings
 - a. Representative sample
 - b. Generalizability of Findings
5. Precision of Outcome
 - a. Practical value - To be considered effective, the effects of an intervention must be practically important. Evaluation reports should report some evidence of practical importance.
 - b. Replication - Consistent findings are required from at least two different high-quality trials that meet all of the above criteria and each of which has adequate statistical power.

Criteria for Broad Dissemination

1. To claim readiness for broad dissemination, a program must meet all of the criteria for effectiveness plus the following.
2. The program must have the ability to go to scale, including providing all program materials and necessary services (e.g., manual, training and technical support).
3. Clear cost information must be readily available.
4. Monitoring and evaluation tools must be available to providers.

Other Resources: The complete *Standards of Evidence* publication and supplementary information can be found at www.preventionresearch.org/.

Evidence Grading System: *The What Works Clearinghouse*

Developer: U.S. Department of Education

Description of Purpose and Scope: *The WWC promotes informed education decision making through a set of easily accessible databases and user-friendly reports that provide education consumers with ongoing, high-quality reviews of the effectiveness of replicable educational interventions (programs, products, practices, and policies) that intend to improve student outcomes. To do this, the WWC uses standards for reviewing and synthesizing research. The WWC is currently conducting systematic reviews of existing research, and producing study, intervention, and topic reports (WWC, "Overview", n.d.)*

Review Process: *The What Works Clearinghouse (WWC) reviews studies in three stages. First, the WWC screens studies to determine whether they meet criteria for inclusion within the review activities for a particular topic-area...Second, the WWC determines whether the study provides strong evidence of causal validity ("Meets Evidence Standards"), weaker evidence of causal validity ("Meets Evidence Standards with Reservations"), or insufficient evidence of causal validity ("Does Not Meet Evidence Screens")... Third, all studies that meet the criteria for inclusion and provide some evidence of causal validity are reviewed further to describe other important characteristics...Finally, the WWC summarizes the evidence of all interventions for a topic in the topic report. Neither the What Works Clearinghouse (WWC) nor the U.S. Department of Education endorses any interventions.(WWC, "WWC Study," p. 2)*

System Criteria and Elements: The criteria below were extracted from the What Works Clearinghouse *WWC Study Review Standards* (n.d.).

Relevance Screening Criteria

- Relevance of Intervention: Is the intervention relevant to the WWC review?
- Relevance of Sample: Is the study's sample relevant to the WWC review?
- Recency of Study: Was the study conducted during a time frame appropriate to the WWC's review?
- Relevant Outcome Measure: Does the study contain at least one outcome measure relevant to the WWC's review?
- Valid Outcome Measure: Does the content of the outcome measure have face validity or adequate reliability? The study author must provide the title of the test and 1) test items that are relevant to the topic, 2) a description of the test items showing that the items are relevant to the topic, or 3) evidence of test reliability.
- Adequate Reporting: Can an effect size be calculated for at least one relevant, valid outcome measure?

Causal Validity Standards

- Study Design: Does the study design appear to be a randomized controlled experiment (RCT), a quasi-experiment with matching (QED), or a regression discontinuity design (RD)?
- What is the study design? (RCT, QED, RD)

RCTs

- Randomization: Were participants placed into groups randomly?
- Baseline Equivalence: Were the groups comparable at baseline, or was incomparability addressed by the study authors and reflected in the effect size estimate?
- Differential Attrition: Is there a differential attrition problem that is not accounted for in the analysis?
- Overall Attrition: Is there a severe overall attrition problem that is not accounted for in the analysis?
- Disruption: Is there evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants?

RDs

- Comparability: Were the groups comparable at baseline, or was incomparability addressed by the study authors and reflected in the effect size estimate? In the context of regression discontinuity studies, "comparability" means that a single regression line for the variable used to create the groups describes the sample.
- Differential Attrition: Is there a differential attrition problem that is not accounted for in the analysis?
- Overall Attrition: Is there a severe overall attrition problem that is not accounted for in the analysis?
- Disruption: Is there evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants?

QEDs

- Baseline Equivalence: Were the groups equivalent at baseline, or was incomparability addressed by the study authors and reflected in the effect size estimate?
- Differential Attrition: Is there a differential attrition problem that is not accounted for in the analysis?
- Overall Attrition: Is there a severe overall attrition problem that is not accounted for in the analysis?
- Disruption: Is there evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants?

Other Study Characteristics

Intervention Fidelity

- Documentation: Is the intervention described at a level of detail that would allow its replication by other implementers?
- Fidelity: Is there evidence that the intervention was implemented in a manner similar to the way it was defined?

Outcome Measures

- Reliability: Is there evidence that the scores on the outcome measure were acceptably reliable?
- Alignment: Is there evidence that the outcome measure was overaligned to the intervention?

People, Settings, and Timing

- Outcome Timing: Does the study measure the outcome at a time appropriate for capturing the intervention's effect?
- Subgroup Variation: Does the study include important variations in subgroups?
- Setting Variation: Does the study include important variations in study settings?
- Outcome Variation: Does the study include important variations in study outcomes?

Testing within Subgroups

- Analysis by Subgroup: Can effects be estimated for important subgroups of participants?
- Analysis by Setting: Can effects be estimated for important variations in settings?
- Analysis by Outcome Measures: Can effects be estimated for important variations in outcomes?
- Analysis by Type of Implementation: Can effects be estimated for important variations in the intervention?

Analysis

- Statistical Independence: Are the students statistically independent (i.e., the outcomes for some participants in a group are unrelated to the outcomes of others in that group) or, if there is dependence, can it be addressed in the analysis?
- Statistical Assumptions: Are statistical assumptions necessary for analysis met?
- Precision of Estimate: Is the sample large enough for sufficiently precise estimates of effects?

Statistical Reporting

- Complete Reporting: Are findings reported for most of the important measured outcomes?
- Formula: Can effects be estimated using the standard formula (or an algebraic equivalent)?

Other Resources: The most current study and topic reports produced by *What Works Clearinghouse* are available at www.whatworks.ed.gov/.