Report of the What Works Clearinghouse Expert Panel

To: National Board for Education Sciences

From: Hendricks Brown, Ph.D.
David Card, Ph.D. (chair)
Kay Dickersin, Ph.D.
Joel Greenhouse, Ph.D.
Jeffrey Kling, Ph.D.
Julia Littell, Ph.D.

Re: Expert Report on the What Works Clearinghouse

Date: October 21, 2008

I. Introduction and Summary

We have been charged with the task of conducting a "…focused study addressing the fundamental question of whether the Clearinghouse's evidence review process and reports are scientifically valid-that is, provide accurate information about the strength of evidence of meaningful effects on important educational outcomes." (Our complete charge is reproduced as Appendix A, below).

Based on our investigation and analysis of the What Works Clearinghouse (hereafter, WWC), we have concluded that:

(1) WWC procedures and processes for identifying and extracting information from intervention studies are generally well documented and follow reasonable standards and practices for systematic reviews;

(2) WWC Intervention and Topic Area Reports provide succinct and meaningful summaries of the evidence on the effectiveness of specific education interventions.

Support for these conclusions is detailed in the remainder of the report. We have also formed a number of specific recommendations for the continued enhancement and improvement of WWC procedures, which are summarized in section IV. Primary among these recommendations is that the Department of Education commission a comprehensive review of the full range of WWC activities and procedures, with a time frame to allow a complete consideration of a number of issues we have not been able to fully evaluate in this report.

II. Description of the Panel's Activities and Materials Considered

The panel was convened in late July 2008, and held two telephone conferences meetings on August 25, 2008 and September 5, 2008, with representatives from the Institute of Education Sciences (IES) present at both meetings. The panel also met for a full day on September 11, 2008 with IES staff and members of the National Board for Education Sciences. During part of the meeting the project director of WWC, Dr. Mark Dynarski, and the deputy director, Dr. Jill Constantine, were present and answered questions from the panel.

The panel was provided with a number of confidential documents describing the procedures of the WWC, including an August 18, 2008 draft of the "What Works Clearinghouse: Standards and Processes" (S&P) manual, and full documentation for the reviews of three interventions: the "Check and Connect" program, which was reviewed as a dropout prevention intervention, the "Tools of the Mind" program, which was reviewed as an early childhood education intervention, and the "Accelerated Reader" program, which was reviewed as a beginning reading intervention. Included in the materials were the study protocols for three topic areas, copies of the studies that passed the initial eligibility screenings, and the WWC Intervention Reports for these three interventions. These documents allowed the panel to assess the implementation of WWC procedures and standards for a relatively large group of studies (a total of 119 studies, 19 of which passed the initial eligibility screen). The panel also benefited from an extensive series of written communications from IES and WWC staff, answering specific questions posed before, during, and after our September 11 meeting.

III. The WWC Review and Reporting Process

a. Goals of the WWC

In view of the time constraints faced by the panel, the panel determined that it would conduct its evaluation assuming that the goal of the WWC review and reporting process is to assess and summarize the strength of the evidence regarding the effectiveness of replicable interventions (programs, products, practices, and policies).[1] The panel notes that a more comprehensive review could (and presumably would) evaluate the mission of the WWC, and in particular the focus on judging efficacy of specific interventions.

b. Overview of the Review and Reporting Process

Building on existing research and accepted principles in the field of research synthesis, the panel identified six key steps in the WWC's review and reporting process that follow the delineation of a topic area for systematic review[2]:

1. formulation of inclusion criteria for studies in a topic area

[1] This is consistent with the description of the WWC at http://ies.ncee/wwc/aboutus/.
[2] The panel did not consider or try to evaluate the choice of topic areas considered by WWC.

2. development and implementation of search criteria for potentially included studies
3. implementation of initial eligibility screens
4. review and classification of studies passing initial eligibility screen
5. extraction and synthesis of estimated effects from studies deemed to contain useable evidence
6. summary and reporting of evidence on the effectiveness of specific interventions

For each of the six steps, the panel reviewed the standards adopted by the WWC in the "Standards and Processes" Manual, and assessed the application/implementation of these standards to the three interventions for which we had complete materials.

c. Step 1: Formulation of Inclusion Criteria

The protocol template describes the rules that will be used to classify a study as "meeting" or "not meeting" evidence screens during Stage 1 of the WWC review process. For each topic area, the WWC protocol template covers the following issues:

- **Topic Area Focus** – Defining the outcomes that the interventions should affect, and the particular population subgroups of interest
- **Key Definitions** – Definitions of outcomes, intervention types, key subgroups
- **Inclusion Criteria** –
  a. Populations
  b. Types of interventions
  c. Types of research studies
  d. Topic relevance
  e. Timeframe relevance
  f. Sample relevance
  g. Study design relevance
  h. Outcome relevance
- **Specific Topic Parameters**
  a. Characteristics of interventions
  b. Elements of intervention replicability
  c. Outcomes relevant to the topic area
  d. Reliability of outcome measures
  e. Timeframe of review
  f. Defining characteristics of the target population
  g. Characteristics relevant to equating groups
  h. Effectiveness of the intervention across different groups
  i. Effectiveness of the intervention across different settings
  j. Measuring post-intervention effects
  k. Defining differential attrition
  l. Defining severe overall attrition
  m. Statistical properties important for computing effect sizes.

These issues can be grouped into five main criteria:

- Population(s) of interest

- Types of interventions
- Time period covered
- Types of outcomes
- Standards of evidence

With respect to the definition of the population of interest and the types of interventions to be considered in a topic area, the panel noted that WWC procedures rely on a combination of nominations from the field and discussions with Department of Education staff. Some topic areas (such as dropout prevention) necessarily involve a narrower population, whereas others (e.g., early reading) can refer to a broader population or to a specific subgroup. Since the choice of the target population is primarily a question of resource allocation and not scientific appropriateness, the panel did review this issue further. The panel infers from existing documents and information received from IES and WWC staff that WWC focuses on interventions that involve a well-defined set of activities that can be replicated in other settings (for example, "branded" interventions sold by publishing companies). The panel agrees with this focus, since systematic reviews of existing research are most likely to be informative when the intervention meets these criteria, and these types of interventions are likely to be of wide interest to the education community.

With respect to the time period covered, the protocols have generally limited the scope to studies conducted in the past 20 years (with some categories such as conference proceedings limited to the most recent seven years). The panel believes this limitation is appropriate.

With respect to outcomes of interest and standards of evidence, there are specific standards defined in each protocol for the following:

- Admissible research designs
- Reliability of outcome measures
- Characteristics relevant to equating groups
- Timing of measurement of post-intervention effects
- Defining differential attrition
- Defining severe overall attrition
- Statistical properties important for computing effect sizes

In general, the panel believes that the specification of minimum standards for reliability of outcomes, timing of measurement, attrition, and the information needed to construct effect sizes is appropriate for a systematic review.

The types of research designs considered within scope for WWC include randomized controlled trials (RCTs) and longitudinal quasi-experimental designs (QEDs) with pre-intervention equating. For the latter designs, the protocol specifies minimum standards for characteristics relevant to equating the treatment and comparison groups prior to the treatment intervention. The panel agrees with the use of such standards. In principle, regression discontinuity-based quasi-experimental designs are also considered in scope, but minimum standards for these designs have not yet been developed, and the panel did not consider the standards for these designs.

Three other issues potentially relevant for specifying standards of evidence are not explicitly covered by the protocols:

- Standards for non-compliance with assignment status
- Specification of intensity of treatment
- Specification of the control state

With respect to the first of these issues, the panel notes that non-compliance with assignment status (also known as "crossover") can lead to difficulties in interpreting intention-to-treat effects and in making comparisons across studies. Current WWC procedures leave the Principal Investigator(s) with discretion to cope with non-compliance but do not require a minimum standard or specify an adjustment process.

With respect to the second issue, the panel notes that comparisons across studies in which the intensity of treatment is varied (e.g., one year of exposure to treatment versus two, or an original version of a curriculum versus a revised or enhanced version) can lead to difficulties in making comparisons across studies. Similarly, the panel notes that the precise conditions for members of the control group (in a RCT) or comparison group (in a QED) can vary across studies even when the treatment is held constant, potentially leading to difficulties in interpreting differences in estimated effects across studies.

Finally, the panel notes that potential issues can arise when a study uses an outcome measure that directly tests for the content of the program itself. [3] In this case the outcome measure is said to be "over-aligned" or to be "treatment inherent." The topic area protocol for beginning reading specifies that RCTs with an "over-alignment problem" will be downgraded, while QEDs will fail the review standard. The protocol for early childhood programs does not appear to mention the issue of over-alignment. The panel recognizes the potential biases that can arise from over-alignment but did not have sufficient time to evaluate the impact of this issue on the WWC review process.

d. Development and Implementation of Search Criteria

Once a topic area is identified and a protocol is established, the WWC follows an iterative search process with two broad phases: (1) a broad search of the literature, based on key words specified in the protocol, to identify potential interventions; and (2) a more focused search for studies of the interventions identified in phase 1. The search parameters (keywords) for the broad search are identified in each topic area protocol. The process includes standard databases, a prescribed list of journals, conference programs, the websites of developers, publishers, and various research organizations, and direct queries to researchers and developers. WWC also receives direct submissions from the public (including researchers and product developers) that supplement other sources of unpublished studies.

The panel believes that the conceptual framework for searching used by the WWC is sound, and that the use of trained librarians in combination with topic area team members

[3] This issue is emphasized by Slavin and Madden (2008).

is in accord with accepted standards. However, documentation of the search procedures actually implemented in the specific topic areas, and information on the results of this process (such as "yield rates" from various sources) is limited, preventing the panel from drawing stronger conclusions. The panel understands that WWC is in the process of revising the protocol template to standardize the reporting of search procedures.

e. Implementation of Initial Eligibility Screens

After the search for potentially eligible studies, WWC staff conduct an initial screening based on the inclusion criteria set out in the topic area protocol. As noted above, these standards can be grouped into five main areas:

- Population(s) of interest
- Type of interventions
- Time period covered
- Types of outcomes
- Standards of evidence

To pass the initial screening stage, studies must pertain to the specified population of interest, during the time period covered, and must address an appropriate type of intervention. Eligible studies must use an eligible research design (in practice, either RCT or longitudinal QED with pre-intervention equating) with at least one "adequate outcome measure." The latter is defined as an instrument that has demonstrated evidence of reliability in a national probability sample. Studies must provide "adequate outcome reporting." This is interpreted as requiring that the study report means and standard deviations for the key outcome measures.

Studies that do not pass the initial screening are classified as "*Does Not Meet Evidence Screens"* and excluded from further review. Studies that pass the initial screens move to the next stage (Stage 2), in which reviewers determine whether the studies meet WWC "evidence criteria."

Initial screening is performed by a single reviewer who reads the title and abstract of a written report, records information on a Study Review Guide, and determines whether or not the study passes this stage. If there is insufficient information in the title and abstract, the initial screen is based on the full-text of a study report. The panel notes that the Study Review Guides appear to be linked to the topic area protocols and are not consistent across topic areas. The panel was unable to assess the completeness of the Study Review Guides for studies of the three interventions for which we had complete materials, or to verify how accurately these Guides are filled out.[4]

The panel believes that the general WWC approach to initial eligibility screening follows accepted practice in the field of systematic review, by determining whether each of the

[4] The materials provided for the Check and Connect review provided full-text reports for five of the six studies that were screened, and completed Study Review Guides for all six studies. In contrast, the materials for the Accelerated Reader review report eligibility decisions, but do not include completed Study Review Guides for all studies.

studies identified in the search process meets predetermined inclusion criteria. It appears that WWC screeners correctly applied standards for population, type of intervention, and date of study in the cases we were able to review. Reliability standards were not used in the review of dropout prevention studies because psychometric outcomes were not used. Common reliability standards (internal consistency 0.6; temporal stability 0.4; inter-rater reliability 0.5) were used for interventions in the beginning reading topic area and the early childhood topic area. It also appears that WWC screeners adhere to the generally accepted principle of retaining studies if there is any doubt about their eligibility (erring on the side of over-inclusion) until the last step of the screening process.

Nevertheless, the panel notes two specific concerns with the WWC initial eligibility screening process. First, the protocols for some topic areas (including the Dropout Prevention topic area) would lead reviewers to eliminate otherwise-eligible studies because those studies did not report group means and standard deviations. This may not be necessary if effect sizes can be calculated from other information. Second, the use of a single screener at the initial eligibility stage may lead to "over-rejection" of potentially eligible studies.[5] Recent studies have concluded that "it is desirable for more than one [reviewer] to repeat parts of the [screening] process" (Higgins & Green 2008) to check the reliability of the screening process

f. Review and Classification of Studies Passing Initial Eligibility Screen

Studies that have passed the WWC initial eligibility screening move to Stage 2 of the WWC review process. In this stage each study is reviewed independently by two reviewers using a Study Review Guide. The two reviews are combined by a senior reviewer, sometimes using additional information obtained from direct queries to the author of a study. The primary goal of the Stage 2 process is to classify eligible studies into three mutually exclusive categories: "Meets Evidence Standards," "Meets Evidence Standards with Reservations," or "Does not Meet Evidence Standards." The use of two independent reviewers is consistent with accepted scientific standards for the conduct of high quality systematic reviews.

The standards for review and classification of studies involve a number of features which are specified differently for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs). For studies that appear to be RCTs the factors are:

- Randomization
- Overall attrition
- Differential attrition
- Intervention contamination
- Teacher-intervention confound

For each factor, a primary standard is established in the Study Review Guide for a study to be classified as "meeting evidence standards." A secondary standard is also established such that if the study falls short of the primary standard, but meets or exceeds

[5] Existing research shows that single screeners can miss up to 24 percent of eligible studies (on average, screeners missed 8 percent of eligible studies; Edwards et al., 2002).

the secondary, then the study is "downgraded." A study that is downgraded on one factor is classified as "meeting evidence standards with reservations." A study that is downgraded on two or more factors is classified as "not meeting evidence standards."

The primary standard for randomization was adjusted as of January 1, 2007, to require that the study provide specific information on the assignment process to establish that the assignment process was random or functionally random. Prior to that date the standard required only that the author claim "random assignment." The panel agrees that the current standard is appropriate, although it notes that the precise definition of random assignment in a classroom setting, where students and teachers are both assigned to treatment or control arms, should be spelled out in detail. The panel also agrees that the inclusion of studies with functionally random assignment is justified, provided that "functionally random assignment" applies to teacher/classroom assignments to treatment and control arms, and that the study demonstrates equivalence using informative pre-assignment characteristics of the students and teachers assigned to the two groups.

The primary standards for overall and differential attrition were set at different thresholds for different topic areas. No standards were defined for beginning reading; the thresholds for early childhood interventions were 20 percent overall, 40 percent within cluster, and 7 percent differential; the thresholds for dropout prevention intervention studies were 30 percent overall and 5 percent differential. The panel notes that there is no standard threshold of attrition in the literature but also believes that there is little scientific basis for relaxing the standard of evidence in different topic areas.

WWC policy on intervention contamination is that for an RCT to meet evidence standards "…there should be no evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants" (Standards and Processes, Appendix B). The panel notes that as a practical matter full information on whether such disruptions occurred is unlikely to be available, but should be taken into account when available.

Standards for teacher-intervention confound are: (1) if there is only one teacher per condition and there is no evidence that teacher effects are negligible the study does not meet evidence standards; (2) if there is only one teacher per condition and the study supplies evidence that teacher effects are "minimal" the study meets evidence standards with reservations; and (3) if there is more than one teacher per condition, or one teacher per condition with strong evidence that teacher effects are negligible the study meets evidence standards. The panel agrees with the basic principle of down-weighting reported evidence from studies with potential teacher confounding, although it also notes that the "strikes against" downgrading process is inherently arbitrary. In the opinion of the panel, case (2) – with only one teacher per condition – is inherently a weak design and arguably fails to meet the standards of evidence.

The protocols and Study Review Guides also specify specific adjustments to be made to estimates of statistical significance that correct for mismatch between unit of assignment

and unit of analysis (e.g., classes are random assignment but the analysis is conducted on student data without clustering by class). These are discussed further in the next section.

For longitudinal QEDs the factors are:

- Equating and baseline equivalence
- Overall attrition
- Differential attrition
- Intervention contamination
- Teacher-intervention confound
- Mismatch between unit of assignment and unit of analysis

Unlike RCTs, the highest classification that can be achieved by a QED is "meeting evidence standards with reservations." Studies that fail to meet the (primary) standard for any of the factors are classified as "not meeting evidence standards." The standards for attrition, intervention confound and teacher confound are essentially the same as the standards for RCTs. The primary difference in standards is that for QEDs the study must establish equivalence of the treatment and comparison groups using a pre-test or proxy of a pre-test (and in some cases other characteristics, as specified in the topic area protocol).

The panel agrees with the general principle that well implemented RCTs represent the strongest form of evidence that is available on the effectiveness of education interventions. We also agree that useful information likely can be gleaned from RCTs with relatively minor flaws in design or implementation, whereas RCTs with substantial flaws are less likely to provide such information. The panel also agrees that well-implemented quasi-experimental studies that compare post-intervention outcomes for a treated group and a comparison group that are closely equated on a pre-test (or other close proxy of the main outcome) provide potentially useful evidence on the effectiveness of education interventions, albeit of lower strength than from the best RCTs. Thus, while we believe the WWC's "two strikes" standard for RCTs and "one strike" standard for QEDs is arbitrary, we conclude that the WWC grading system allows meaningful information to be extracted and presented to users of the WWC.

g. Extraction and Synthesis of Estimated Effects

The methods and procedures for the extraction and synthesis of results from eligible studies are specified in a document entitled "Technical Details of WWC-Conducted Computations" (http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=9). A link to this document is usually found in the respective intervention reports. The computations described in this document are consistent with the standard prescriptions for effect size estimation for continuous and dichotomous outcomes, respectively. In the case studies reviewed by the panel, the application of these procedures for converting study outcomes to effect sizes was appropriate. However, the panel was unable to determine how these procedures were used or modified under nonstandard conditions (e.g., when distributional assumptions were not satisfied).

WWC specifies an adjustment to be made to estimates of statistical significance of estimated effects that correct for mismatch between unit of assignment and unit of

analysis (i.e., when classes are randomly assigned to treatment but the original analysis was conducted on individual level student data without clustering by class). This approach is based on a procedure suggested by Hedges (2007). This is only an approximate adjustment, and does not take into account study-specific factors that could lead to a larger or smaller adjustment than the one prescribed by Hedges' approach. For this reason the panel believes that a study-specific correction, implemented by the study authors using the actual study data, would be preferable.

WWC protocols specify the use of Benjamini-Hochberg corrections when multiple comparisons are reported in a study. These procedures appear to have been implemented in summarizing results for several beginning reading and early childhood interventions. The panel notes that other correction procedures are potentially preferable in certain situations.

Procedures for combining effect sizes across studies (i.e., methods for synthesizing results) are not specified, perhaps because very few interventions have more than one or two studies to combine. The panel notes that in the future WWC will have to develop standards for combining evidence from multiple studies, including the issues of how to compare effect sizes across studies that vary the intensity of an intervention, and across studies with different control conditions.

h. Reporting of Evidence on the Effectiveness of Interventions

WWC reports its overall findings in a highly transparent and timely manner, after outside experts have conducted a review. These results are reported on the WWC website and are readily available for policymakers and practitioners. The presentation is intended to be clear to all audiences, calibrated against standards that have been set by WWC, and scientifically sound.

The panel believes that for the most part, WWC achieves these objectives in its reviews. All systematic reviews have to make important decisions regarding what quality studies they pay attention to, what degrees of missing data and the like that are required, which method is to be used to measure effect sizes, and what cut-offs are required to achieve the highest level of impact. The panel believes that WWC has done a reasonable job of making many of these critical decisions, and a very good job of applying these standards once they have been set.

i. Overall Evaluation of the WWC Process

Overall, the panel believes that the WWC review and processes are based on scientifically appropriate methodologies for the task of judging the strength of the evidence regarding the effectiveness of the interventions identified in the topic areas, although the panel did not have time or resources to fully investigate the application of these methodologies in every review. Moreover, the panel believes that the Intervention

and Topic Area Reports provide succinct and meaningful summaries of the evidence on effectiveness of specific interventions.

IV. Recommendations

1. *Full Review*. The panel recommends that IES commission a full review of the What Works Clearinghouse, including a review of the Clearinghouse's mission, and of the WWC Practice Guides, which we have not attempted to evaluate. The panel also recommends that IES consider instituting a regular review process to ensure that WWC is using the most appropriate standards in its work.

2. *Protocol Templates*. The panel recommends that the WWC review and update the protocol templates, focusing on the following issues:

(i) standards for crossover and assignment noncompliance, and for adjusting intention to treat effects across studies.

(ii) standards for documenting the program received in the control arm of RCTs (or by members of the comparison group in QEDs), and potentially incorporating this information in making comparisons across studies and/or interventions.

(iii) revised standards for multiple comparisons. We recommend that WWC review the treatment of multiple comparisons in light of the recent research report by Peter Schochet entitled "Guidelines for Multiple Testing in Impact Evaluations."

(iv) attrition standards. We recommend that WWC reconsider the current process of setting different attrition standards in different topic areas.

(v) potential conflicts of interest. We recommend that WWC establish a new protocol to keep track of potential conflicts of interest, such as cases where a study is funded or conducted by a program developer, and consider making that information available in its reports.

(vi) randomization. We recommend that the WWC precisely define the standards for "randomization" in a multi-level setting.

3. *Documentation of Search Process*. The panel recommends that the WWC expand the protocol templates to specify more explicit documentation of the actual search process used in each topic area, and maintain a record of the results of the search process that can be used to guide decision making on future modifications of the search process.

4. *Reliability of Eligibility Screening*. The panel recommends that the WWC conduct regular studies of the reliability of the eligibility screening process, using two independent screeners, and use the results from these studies to refine the eligibility screening rules and screening practices.

5. *Documentation of Screening Process*. The panel recommends that WWC reports include a QUOROM-type flow chart documenting the flow of studies through each review and number of studies excluded at each point, and a Table of Excluded Studies, listing specific reasons for exclusion for each study.

6. *Misalignment Adjustment*. The panel recommends that in cases where a study analysis is "misaligned," WWC staff request that study authors re-analyze their data correctly, taking into account the unit of randomization and clustering. We recommend that the results from the process be compared to the simple ex post adjustment procedure currently specified, to develop evidence on the validity of the latter.

7. *Combining Evidence Across Multiple Studies*. We recommend that WWC re-evaluate procedures for combining evidence across studies, with specific attention to the issue of how the rules for combining evidence can be optimally tuned, given the objectives of the WWC review process and the sample sizes in typical studies for a topic area.

8. *Reporting.*
(i) The panel recommends that published reports on the website include the topic area protocols, as well as more information on the screening process results that led to the set of eligible studies actually summarized in the Topic Area reports.

(ii) The panel recommends that WWC make readily available its "Standards and Procedures" manual, including appendices, as well as all other relevant documents that establish and document its policies and procedures.

9. *Practice Guides*. The panel recommends that the Practice Guides – which contain material that does not meet the high standards of evidence for other WWC products – be clearly separated from the Topic and Intervention Reports.

10. *Outreach and Collaboration with Other Organizations.* The panel recommends that the WWC build and maintain a relationship with national and international organizations focusing on systematic reviews, specifically with the goals of having Review Team leaders engaged in the broader scientific community, and in bringing the latest standards and practices to the WWC. The panel also recommends that the WWC convene working groups with a mixture of researchers (including specialists in education research and systematic reviews) to address the development of new standards for the review and synthesis of studies.

Appendix A: Panel Charge

DEPARTMENTS OF LABOR, HEALTH AND HUMAN SERVICES, AND EDUCATION, AND RELATED AGENCIES APPROPRIATION BILL, 2009
REPORT OF THE COMMITTEE ON APPROPRIATIONS
U.S. SENATE,
ON S. 3230
JULY 8, 2008.

The Committee requests that the National Board for Education Sciences, as the body responsible for oversight of the Institute of Education Sciences, convene a blue-ribbon panel of leading experts in rigorous, particularly randomized, evaluations to assess the What Works Clearinghouse. While the Committee believes a comprehensive assessment should be undertaken given the significant investment made in the Clearinghouse, an immediate priority should be a focused study addressing the fundamental question of whether the Clearinghouse's evidence review process and reports are scientifically valid-that is, provide accurate information about the strength of evidence of meaningful effects on important educational outcomes. The Committee requests that the Board convene the panel within 60 days of enactment of this act, and that the panel complete its work and submit a report, including any recommendations for improvements in the Clearinghouse, to the Board, the Director, and Congress no later than 4 months thereafter. The Committee intends for panel members to be free of conflicts of interest.

Appendix B: References

Edwards, Phil, Mike Clarke, Carolyn DiGuiseppi, Sarah Pratap, Ian Roberts, and Reinhard Wentz. (2002). "Identification of Randomized Controlled Trials in Systematic Reviews: Accuracy and Reliability of Screening Records." *Statistics in Medicine*, 21(11): 1635-1640.

Hedges, Larry (June 2007). "Correcting a Significance Test for Clustering." *Journal of Educational and Behavioral Statistics*, 32: 151-179.

Higgins, Julian P.T., and Sally Green, editors. (2008). *Cochrane Handbook for Systematic Reviews of Interventions.* New York: Wiley.

Slavin, Robert E, and Nancy A. Madden. (March 2008). "Understanding Bias Due to Measures Inherent to Treatments in Systematic Reviews in Education." Paper presented at the Annual Meeting of the Society for Research on Effective Education, Crystal City, VA.