From: Mark Dynarski
Sent: Friday, January 16, 2009 5:23 PM
To: Scott Cody; Andrei Streke
Cc: Michael Ponza; Jill Constantine
Subject: FW: Followup to IES Board mtg -- What Works Clearinghouse (WWC)
Attachments: New Chance demonstration.pdf; Reading Recovery Pinnell 1988 RCT.pdf; Reading Recovery Schwartz 2005.pdf

I am sure we will need to prepare responses to these issues.

Mark

From: Jon Baron [mailto:jbaron@excelgov.org]
Sent: Friday, January 16, 2009 4:53 PM
To: Mark Dynarski; Jill Constantine
Cc: Eric Hanushek; Sally Shaywitz (sally.shaywitz@yale.edu); (b)(6) GearyD@Missouri.edu; Phoebe Cottingham (Phoebe.Cottingham@ed.gov); Lynn Okagaki (Lynn.Okagaki@ed.gov); Kerachsky, Stuart; Garza, Norma; Betka, Sue; Carol D'Amico (cdamico@conexusindiana.com); Jeffrey Kling (JKLING@brookings.edu)
Subject: Followup to IES Board mtg -- What Works Clearinghouse (WWC)

Hi Mark, Jill, and others:

It was good to see you at the IES Board meeting, and I thought your update on the latest improvements in the WWC was extremely helpful. As I mentioned at the meeting, I agree that the Expert Panel did an excellent job in its focused review of the WWC's overall standards and methodologies, and believe the WWC has greatly improved over the last couple years. I also support the Panel's primary recommendation -- that the Department commission a comprehensive review of WWC. In particular, I'd suggest such a review examine how the WWC has applied its overall standards/methodologies to specific interventions and studies -- an aspect that the Panel did not have a chance to examine (except in a very limited way) given its short time frame.

I suggest this because, based on reviewing many of the underlying studies that WWC reviewed, I think the WWC, in a number of cases, has missed key flaws in study design and implementation – particularly in studies found to meet evidence standards without reservations. Below are four brief, representative examples (two I mentioned in the meeting, and two new ones). These are illustrative of a number of other examples that I and others have identified.

But as a first step, I'm submitting these examples through the feedback process you suggested. Thanks for your consideration of this input, and I look forward to your thoughts. We all support the WWC as the centerpiece of IES's research dissemination efforts, and want it to succeed. I hope this is helpful toward that end.

Jon

Jon Baron
Executive Director, Coalition for Evidence-Based Policy
Council for Excellence in Government
202-530-3279
www.excelgov.org/evidence

(1) Study: Large randomized controlled trial of the New Chance demonstration – a program for young welfare mothers that provides them and their children with comprehensive services (basic academic skills, job training, parenting workshops, child care, etc). The study had a sample of over 2000 women, and a follow-up 3.5 years after random assignment.

WWC review: The WWC overview, shown here, reports that the program produced a sizeable improvement in school completion. Most readers of this WWC overview would be led to believe that the program is backed by promising evidence of effectiveness.

The study's actual findings: The findings show a much different picture of the program's overall effects. It did produce a statistically-significant 11.8 percentage point increase in participants' receipt of a GED. However, it also produced:

- A statistically-significant *decrease* of 3.5 percentage points in their receipt of a high school diploma.
- No impact on their receipt of a trade certificate or license.
- No impact on their reading skills.
- No impact on their employment rate, earnings, or receipt of welfare (which, as you know, are key outcomes that remedial education seeks to improve).
- No impact on their children's school readiness.
- A statistically-significant *adverse* impact on their children's behavior

In light of these findings, the WWC overview could easily mislead readers, and encourage the adoption of a program that is costly and produces little or no meaningful improvement in the lives of young disadvantaged women or their children. I've attached the study report's executive summary, and marked in red pencil the key tables showing program impact (see pdf pages 13-14, 19-20, and 22).

(2) Study: *Reading Recovery: Early Intervention for At-Risk First Graders*, Pinnell et. al, 1988., a randomized controlled trial of Reading Recovery -- a one-on-one tutoring program for first graders at risk of reading failure. Of approximately 100 first graders identified as being the weakest readers, half were randomly assigned to Reading Recovery, and half were randomly assigned to a control group. This produced treatment and control groups of equal size (about 50 students each). However, the study authors informed the WWC that there were 53 students in the control group and just 38 in the treatment group. The reason for the smaller reported treatment group is the "intention-to-treat" problem described below.

(The study also had a nonrandomized comparison group, but it wasn't relevant to the WWC review.)

WWC review: This study was found to meet WWC evidence standards without reservations.

Key flaw I believe the WWC missed: The study – rather than estimating the program's impact using *all* students randomly assigned to the treatment group – only used those treatment-group students "who at some time during their first-grade year had 60 or more lessons or were successfully discontinued (released) from the program" (see attached study report, marked paragraphs on pages 35-36 of the pdf). Treatment-group students who did not participate at this level were dropped from the treatment group. This is a straightforward violation of the intent-to-treat principle, which likely distilled the treatment group down to the more capable students, and this advantage in initial capabilities, rather than the program, could well have accounted for the superior outcomes for the treatment group versus the controls. (Based on a careful read of both the study report and the WWC report – which included correspondence with the study authors -- the number of students dropped from the treatment group was 15 out of the initial 53, leaving 38 students.) The WWC seems to have missed this key flaw in study implementation, found the

study to meet WWC standards without reservations, and appears to have based its report of the program's effects on the 38 students remaining in the treatment group. (For reference, the WWC's summary of this study's characteristics is posted here – see page 17.)

(3) Study: *Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention*, Robert Schwartz, 2005, a randomized controlled trial of Reading Recovery. The study randomly assigned 94 first-graders at risk of reading failure to Reading Recovery versus a control group, and measured reading outcomes halfway through the school year.

WWC review: This study was found to meet WWC evidence standards without reservations.

Key flaw I believe the WWC missed: In most cases, the reading outcomes in this study were measured by the Reading Recovery teachers themselves, through individually-administered tests of their first-grade students (see attached study report, marked paragraphs on pdf pages 5 and 10). Such tests inherently allow some subjective judgment (e.g., about whether a child correctly sounded-out a word or not). This raises a real possibility that the teachers' biases -- e.g., as proponents of Reading Recovery -- could have consciously or unconsciously influenced their outcome measurements, and thereby undermines the confidence one can have in the study results. (For reference, the WWC's summary of this study's characteristics is posted here – see page 19.)

(4) Study: Power4Kids – a large randomized controlled trial of 4 reading interventions for struggling readers in 3rd and 5th grade. The interventions were Wilson Reading, Kaplan Spell Read, Corrective Reading, and Failure Free Reading.

WWC review: This study is cited in WWC's reports on all 4 interventions as meeting WWC evidence standards without reservations.

Key flaw I believe the WWC missed: While this was an important study, the design confounded teacher quality/motivation with the interventions, undermining the study's ability to produce strong evidence regarding the impact of these interventions. Specifically, the study randomly assigned about 30 schools to one of these 4 reading interventions, and then randomly assigned students within each school to that school's assigned intervention or to a control group. Importantly, however, the study did not randomly assign *teachers* within each school to treatment versus control classes. Instead, the study purposely hired the most motivated and capable teacher volunteers within each school to teach the treatment classes, whereas the control classes were taught by the school's other teachers. (The teachers *were* randomly assigned with their school to one of the 4 intervention groups, but were not randomly assigned *within* each school to treatment versus control classes.) I've posted a copy of the Power4Kids study report here, and marked the relevant text on pages 32, 33, and 58 of the pdf. Thus, the treatment and control groups likely differed both in teacher quality and intervention.

(For reference, the WWC's summary of this study's characteristics is posted here – see pdf page 1.)

Attachment

Below are the findings of the Quality Review Team's examination of four studies identified by Jon Baron in his January 16 message to the What Works Clearinghouse. For three of the studies, the Quality Review Team concluded that the WWC had addressed the issues appropriately in light of the mission and standards of the WWC. For one of the studies (Pinnell et al., 1988), the team concluded that the potential problem identified by Mr. Baron did not occur. However, in investigating this issue, the team learned of a issue in the study that may affect its rating. The team is investigating this issue with the author and will revise the intervention report on the WWC website if necessary.

Study 1: New Chance Demonstration (Quint et al., 1997)

Regarding the New Chance study (Quint et al., 1997), Mr. Baron expressed concern that readers will view the overall program as effective whereas the conclusions of its authors indicate that some outcomes did not improve or were negatively affected by the program.

The Quality Review Team concluded that while the authors looked at a wide range of outcomes to arrive at their conclusions, the WWC reviewers followed its procedures correctly in not including all these outcomes in its intervention rating. The QRT noted that the WWC report explicitly noted that other outcomes were examined in the study.

The WWC systematic review procedures require principal investigators to establish outcome domains for review areas. Because the WWC is designed to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence for what works in education, these outcome domains generally focus on education outcomes. Additionally, reviews typically focus on education outcomes that are directly relevant to the topic area. For example, for the Dropout Prevention topic, the review examined outcomes related to students' attendance in school, progress in school, and completion of a high school diploma or GED.

For the New Chance study, the report of the WWC review noted that it did not include other study outcomes: "The study also examined the program's effects on college credits, trade certification, reading scores, living arrangements, fertility, mental and physical health, employment, earnings, AFDC receipt, and child outcomes. These outcomes, however, do not fall within the three domains examined by the WWC's review of dropout prevention interventions (staying in school, progressing in school, and completing school)" (p.6).

Mr. Baron also notes that the WWC report combined a negative effect on diploma completion and a positive effect on GED completion. While the WWC report indicates that New Chance effects are potentially positive in this domain, the report also clearly and repeatedly differentiates effects for these two outcomes. For example, the report states: "the positive effect on completion came entirely from New Chance's positive and statistically significant effect on the likelihood of receiving a GED certificate. New Chance was found to have a small, but statistically significant, negative effect on the likelihood of earning a high school diploma" (p.3).

When there are multiple outcomes within a domain, the WWC's process is to compute a simple average of the individual outcomes. It is possible that large positive effects on one outcome within a domain offset small negative effects on another outcome in the same domain. The WWC recognizes that educators, researchers, and policymakers may attach different importance to these individual outcomes, but the WWC makes no judgments as to their relative importance as doing so necessarily would involve a degree of subjectivity.

Study 2: Reading Recovery (Pinnell et al., 1988)

Mr. Baron expressed concern about the WWC's review of Pinnell et al. (1988). The study's statement that the treatment group included "those children who at some time during their first-grade year had 60 or more lessons or were successfully discontinued (released) from the program" could be interpreted to mean some students assigned to the treatment group were dropped because they received fewer than 60 lessons, which would violate the intent to treat principle, as Mr. Baron points out.

The WWC review of Pinnell et al. (1988) is based on five separate (but related) manuscripts as well as direct correspondence with the study author. The Quality Review Team has correspondence with the primary author that indicates the authors did not drop students from the treatment group if they received less than 60 lessons. Some students had fewer than 60 lessons because they were successfully discontinued from Reading Recovery before 60 lessons; these students were retained in the study and analyzed as part of the treatment group.

In investigating this issue, the Quality Review Team identified a possible issue in the WWC review, which, if confirmed, explains why the treatment and control groups are different sizes though the authors state that half were randomly assigned to treatment and half to control. There is an indication in author correspondence that 37 of the 51 students originally assigned to the control group were assigned through a random process. The remaining 14 students may have been assigned through a non-random process, which the team is attempting to confirm with the author. If the additional correspondence is not informative, the WWC will assume the treatment and comparison groups were formed by a non-random process. The study then will be rated as a quasi-experimental design.

Study 3: Reading Recovery (Schwartz, 2005)

Mr. Baron expressed concern that the findings from the Schwartz (2005) study of Reading Recovery could be biased because the testers were not blind to treatment status. As Mr. Baron notes, the author raises this possibility (page 266).

While the Quality Review Team agrees that there is some possibility that the findings of the Schwartz study could be biased by the testers opinions about Reading Recovery, the team believes it is appropriate for the WWC to include these findings as valid. WWC standards do not require that testers be blind to treatment status. The WWC does have standards that outcomes need to demonstrate an adequate degree of validity and reliability and WWC principal investigators can reject an outcome measure as invalid if they believe tester bias could influence the results.

In the case of Schwartz (2005), the principal investigator of the Beginning Reading topic area chose to include in the review the battery of tests in the Schwartz study because these measures have been demonstrated to have adequate validity and reliability. Moreover, these tests were not developed by the study authors and are commonly used measures of reading achievement. It also is relevant to note that the study's "transition" tests (which are used as post-tests for estimating effects) were not administered by the students' Reading Recovery teacher but by a different Reading Recovery teacher (consistent with Reading Recovery guidelines). These teachers were not blind to the treatment status but did not have first-hand knowledge of the students' performance during Reading Recovery instruction.

Study 4: Power 4 Kids (Torgesen et al., 2006)

Mr. Baron expressed concern that the study of Power 4 Kids contained in the Torgesen et al., report includes a confound because the study purposely hired teachers to teach the treatment classes and measured effects could reflect the quality of teachers that were hired rather than the effects of the interventions.

The Quality Review Team concluded that the WWC review followed its procedures correctly.[1] The Power 4 Kids study tested whether small-group, pull-out instruction with trained reading teachers could improve student reading performance. The authors state on pages 3-4:

[1]Individuals on the Quality Review Team and some of the authors of the Power 4 Kids study are employees of Mathematica Policy Research.

….this study is an evaluation of interventions that both focus on particular content and are delivered in a particular manner. Our decision to manipulate both of these dimensions simultaneously is consistent with one of the important goals of the study: to examine the extent to which reading skills of struggling readers in grades three and five could be significantly accelerated if high quality instruction was delivered with sufficient intensity and skill. It also means, of course, that if there is a significant impact of an intervention compared to the control group, the impact could be related to either the increased intensity of instruction or to the particular focus of the intervention.

The WWC protocol calls for including a study in the WWC review if the study's context and procedures reflect what reasonably would be encountered in a real-world school setting. The Quality Review Team concluded that the context and procedures in the Power 4 Kids study reflect conditions likely to be encountered in a real-world setting. It is a common practice for schools to have trained reading specialists who use branded curricula in a small-group setting to assist struggling readers. The Power 4 Kids study examines effectiveness of four different branded curricula in this context. All WWC intervention reports that cite this study contain details about the training received by intervention teachers, and all indicate that instruction was provided in a small-group, pull-out setting.

From:	Mark Dynarski
Sent:	Thursday, February 19, 2009 1:06 PM
To:	Jon Baron
Cc:	Jill Constantine; Scott Cody; Eric Hanushek; Sally Shaywitz (sally.shaywitz@yale.edu); FPHandyman@aol.com; GearyD@Missouri.edu; Phoebe Cottingham (Phoebe.Cottingham@ed.gov); Lynn Okagaki (Lynn.Okagaki@ed.gov); Kerachsky, Stuart; Garza, Norma; Betka, Sue; Carol D'Amico (cdamico@conexusindiana.com); Jeffrey Kling (JKLING@brookings.edu)
Subject:	response regarding WWC reviews and NBES meeting discussion
Attachments:	2009001 Dynarski response.docx

I am attaching the findings from the WWC's quality review team regarding the four studies you had identified in your e-mail of January 16, 2009 as possibly having been reviewed incorrectly by the WWC. As the attachment notes, the team concluded that the WWC applied its standards correctly in all four cases. In conducting the investigation, the team identified a potential issue in one study of Reading Recovery about which it is gathering more information from the author.

During the Board meeting, you stated that you believed the WWC has quality control problems that affect perhaps a third of its reviews. I want to assure you that the WWC has high quality-control standards that render a thirty-percent error rate a near impossibility. There are five steps in the quality control process.

1) All studies are reviewed by two certified reviewers (these are reviewers who have completed the training course, scored high enough on the post-training test, and successfully completed two reviews according to a senior WWC researcher). If the reviewers differ in their assessment of a study, a senior team member examines the differences and consolidates the reviewers into a master review.

2) Reviews are synthesized in interventions reports and that report is reviewed by the principal investigator.

3) That report is reviewed independently by a senior WWC manager who is not part of the review team.

4) The report is reviewed by the IES officer for the WWC.

5) The report is reviewed independently by anonymous peer reviewers.

No process can be completely free of errors and the WWC has a Quality Review Team that investigates issues brought to its attention by users, developers, and authors. This team operates independently of the team that conducted the initial review.

We recognize that WWC users may have their own standards or preferences and desire that the WWC should carry out its reviews according to these standards. For example, the point you raise about the New Chance study having effects that did not favor the intervention but were not a focus of the WWC review is not asserting that the WWC did not apply its standards correctly, but rather that if it had used different standards (i.e. examined outcomes outside the WWC set of focal outcomes established *a*

priori for the dropout prevention review topic) it possibly would have reached different conclusions. Any standards-based review effort is based on tradeoffs about the utility and transparency of the standards, and we are always monitoring WWC standards to ensure that its reports are based on sound protocols and reviews of the extant research while being useful to educators and policymakers.

Regards,
Mark