From: Scott Cody
Sent: Wednesday, June 17, 2009 11:22 AM
To: Sakari Morvey
Subject: FW: Problems with Newly Released Revised WWC Early Childhood Report

-----Original Message-----
From: Mark Dynarski
Sent: Monday, June 15, 2009 3:27 PM
To: Christopher J. Lonigan, Ph.D.
Cc: Jill Constantine; Scott Cody
Subject: RE: Problems with Newly Released Revised WWC Early Childhood Report

I think you've raised some issues we need to consider and I am referring your message to the quality review team for consideration. It's not often we have a reader of a WWC report with quite your sophisticated knowledge of the standards.

Regards,
Mark

-----Original Message-----
From: Christopher J. Lonigan, Ph.D. [mailto:lonigan@psy.fsu.edu]
Sent: Saturday, June 13, 2009 12:15 PM
To: Mark Dynarski
Cc: Jill Constantine
Subject: Problems with Newly Released Revised WWC Early Childhood Report
Importance: High

Hi Mark (and Jill),

I was on the WWC site this morning and came across the updated Doors to Discovery report. The report notes that the original WWC report erroneously included results from Assel et al. (2006). The stated error was that the study has severe subcluster attrition, which was not attended to in the original WWC report.

I think that this conclusion is incorrect.

Assel et al. (2006) reported that the sites included in the study had 603 children.

Their analysis sample--based on the information reported in Appendix A was 580.

Attrition within assigned group was .04 (Let's Begin), .03 (Doors), and .05 (Control) for the analysis sample.

This does not trigger either an overall or differential attrition concern.

Moreover, even if it did trigger such a concern, one can compute group equivalence at pretest from the information provided (i.e., means and SDs for each group at pretest).

In place of Assel et al., the WWC included the PCER report. Of course, the PCER report is a subsample of the Assel et al. study. Therefore, if one study report has some mysterious subcluster attrition problem, then the other does as well. PCER's analysis sample was a subset of the children in Assel et al (so, if the "attrition" issue is based on the difference between the N = 603 and the analysis sample, the PCER Study has an even bigger problem).

The "Extent of Evidence" thing also needs to go away. PCER (as like Assel) found no significant or substantively important effects. The revised report includes a summary of another study that randomized **4** classrooms to control versus Doors. The Extent of Evidence is rated as medium to large. Yet, the significant contribution to this rating is a study showing no significant or substantively important effects. The large, no impact of curriculum study is giving the very small, no significant impact study standing in this case. This is INSANE! The domain average effect size is a simple average instead of a weighted average, which is bizarre. If one applied a weighted average, based on the larger PCER sample (N = 183) and the smaller Christie et al. sample (N = 37), the domain average effect size would clearly be closer to the PCER's .16 (oral language) than to the Christie et al.'s .30.

MOREOVER, in the Christie et al. Study two of the classrooms served ELL children and two served English-only children. Type of child served was used as a blocking factor prior to randomization. This does not meet the standards of randomization. The only acceptable outcome would be if one of the ELL classrooms and one of the non-ELL classrooms ended up in each condition. If it were the case that this outcome did not obtain, curriculum condition would be confounded with type of student population and the study would fail. Because this confounding is not an acceptable randomization outcome, the study cannot really be considered randomized (i.e., there was not equal probability of which group a classroom ended up in). The randomization is based on two units (randomization of two of anything cannot equate). There seems to be no way that one could imagine that classrooms would end up being equated on expectation. Hence, one could argue that this is really a QED. There is also reported to be severe attrition (40%), and SUBSTANTIAL differences at pretest. The study fails standards.

So, there are three problems here:

1. The Assel et al. study (which is the largest of the studies) is rejected on what appears to be an erroneous rationale.

2. The Christie et al. study is included with a evidence standards rating that seems impossible to justify.

3. The INSANE application of extent of evidence.

The first of these two certainly seem like they are violations of WWC procedures. This seems like a serious problem. I would like to hear how such decisions were justified. As a strong supported of WWC and someone who has worked inside WWC, it is import to me that the output fits the high quality rules-based system in place.

This has to be done right.

Chris

Christopher J. Lonigan, Ph.D.
Professor
Associate Director, Florida Center for Reading Research
Department of Psychology
Florida State University
1107 W. Call Street
Tallahassee FL 32306-4301

October 23, 2009

Professor Christopher J. Lonigan
Associate Director, Florida Center for Reading Research
Department of Psychology
Florida State University
1107 W. Call Street
Tallahassee FL 32306-4301

Dear Dr. Lonigan:

In response to your June 13, 2009, email concerning the What Works Clearinghouse (WWC) Doors to Discovery Intervention Report, we conducted a quality review. The WWC Quality Review Team responds to concerns raised by study authors, curriculum developers or other relevant parties about WWC reviews published on our website. These quality reviews are undertaken when concerned parties present evidence that a WWC review may be inaccurate. When a quality review is conducted, a researcher who was not involved in the initial review undertakes an independent assessment of the study in question. The researcher also investigates the procedures used and decisions made during the original review of the study. If a quality review concludes that the original review was flawed, a revision will be published. These quality reviews are one of tools used to ensure that the standards established by the Institute of Educational Sciences (IES) are upheld on every review conducted by the What Works Clearinghouse.

The findings of our quality review, discussed in detail below, will lead to revision of the Doors to Discovery Intervention Report. More specifically, we find that both the Assel et al (2006) and Preschool Curriculum Evaluation Research (PCER, 2008) studies had student-level subcluster attrition that exceeded the Early Childhood Education (ECE) topic area standards (40%). Regarding the other issues raised in your email, we find that the ECE protocol (available on the WWC website) and the WWC version 1 protocol were followed.

1) Assel et al. (2006)

The first issued raised in your email was the attrition rate in the Assel et al. (2006) study. The quality review findings are in agreement with the published description that "the subcluster attrition rate of children exceeded standards." The ECE protocol specifies 40% for acceptable subcluster attrition. The subcluster attrition rate was calculated as follows: The authors report 27 control and 25 Doors to Discovery classrooms, the PCER study reports an average of 18.6 students per classroom, and an author response reported parental consent of 65% for the treatment (both treatments combined) and 55% for the control. Based on these numbers, 389 students dropped out due to lack of parental consent (27*18.6*0.45 + 25*18.6*0.35). Of those with parental consent, the PCER study reports that 8 from each classroom were randomly selected for the study. The most complete post-test had 366 respondents (auditory comprehension). Based on these numbers, 50 of the chosen students were not included in the post-test (27*8 + 25*8 – 366). Taken together, including the students without parental consent plus the students without a post-test leads to an attrition rate of 45% ((389+50)/(27*18.6+25*18.6)).

With the high level of subcluster attrition, the study must establish baseline equivalence of the analytic sample. However, the study does not establish baseline equivalence. For example, in Appendix A, the total sample size for Expressive Vocabulary at pre-test was 409 but the analytic sample at post-test was 364. In addition, we have no way of telling the degree of overlap between the pre-test sample and the post-test sample. That is, we cannot assume that all students in the pre-test sample were included in the post-test sample. Appendix A suggests the samples were not consistent. In some cells, the table shows more students at post-test than at pre-test.

2) PCER study

The second issue raised in your email was the attrition rate in the PCER study which was based on the same experiment as the Assel et al study. Reassessing the PCER study, we find that it also had student-level subcluster attrition that exceeded 40%. The subcluster attrition rate was calculated as follows: The authors report 15 control and 14 Doors to Discovery classrooms an average of 18.6 students per classroom. Using the parental consent rates reported in the author responses (65% for both treatments combined and 55% for the control), we calculate that 217 students dropped out due to lack of parental consent (15*18.6*0.45 + 14*18.6*0.35). Of those with parental consent, an author response reports that 7 from each classroom were randomly selected for the study. The most complete post-test had 183 respondents (mathematics, Table C-6a). Based on these numbers, 20 of the chosen students were not included in the post-test (15*7 + 14*7 – 183). Taken together, including the students without parental consent plus the students without a post-test leads to an attrition rate of 44% ((217+20)/(15*18.6+14*18.6)).

For the analytic student sample, the study does not establish baseline equivalence. Appendix Table 6-A provides baseline test measures, but Ns are not the same for the pre and post tests. The total DTD sample size for math measures at pre-test was 100 and at the Spring Pre-K post test it was 94. The control sample for math was 94 at pre-test and 89 and post-test.

In the DTD Intervention Report the PCER study was categorized as "meets evidence standards." The Intervention Report will be revised to reflect the revised calculation of the attrition rate. The PCER study was also used for the Let's Begin with the Letter People Intervention Report which will also be revised.

3) Christie et al. (2003)

The third issue raised in your email regarded randomization in the Christie et al. (2003) study. In January 2009, the reviewers queried the authors for details about the randomization. There were four classrooms that were randomized: two that taught ELL students and two that taught English-only students. For each of the two types of classrooms, one class was chosen at random by coin toss to receive the treatment (i.e., stratification by classroom type). Thus, by design, each class had a 50% chance of receiving the treatment and there was one class of each type in the treatment group. If the analysis were performed separately by classroom type, the study would suffer from an N=1 confound. However, by combining results across the classroom types, there are two classrooms in the treatment group and two in the control group. The WWC does not downgrade studies because randomization was performed by strata or when the number of randomized units is small (except for the N=1 confound).

As you note, the study did have substantial subcluster attrition (41%, Appendix Table A1.2). Based on the high attrition, the authors were required to show baseline equivalence. According to the ECE protocol used at the time and currently available on the WWC website, baseline equivalence was defined as differences of less than 0.5 standard deviation units. Based on detailed results obtained from the authors, the pretest differences (0.40, 0.45, and -0.29; Appendix Table A1.2) were within the limits of the topic-specific standard. It should be noted that under WWC version 2 standards, the study would not meet evidence standards as differences in baseline measures of the outcome variables must be less than .25.

4) Extent of Evidence Rating

The fourth issue raised in your email calls into question the calculation of extent of evidence rating. The findings in the Doors to Discovery Intervention Report are based on the WWC protocol for "potentially positive effects" which calls for evidence of a positive effect with no overriding contrary evidence:

- At least one study showing a statistically significant or substantively important positive effect.
- No studies showing a negative effect and the number of studies showing indeterminate effects is not greater than the number showing statistically significant or substantively important positive effects.

However, we agree that in this case the "potentially positive effect" resulted from a small study (Christie et al, 2003) which was implicitly supported by extent of evidence taken from a large study that showed indeterminate effects (PCER, 2008). The Statistical, Technical, and Analysis Team of the WWC is aware of the limitations of the current extent of evidence categorization and will be considering options to make it more informative.

I hope that this letter has addressed your concerns.

Sincerely,

(b)(6)

Deborah Reed
WWC Quality Review Team

From: What Works
Sent: Monday, October 26, 2009 12:35 PM
To: 'lonigan@psy.fsu.edu'
Subject: What Works Clearinghouse (WWC 2009005)
Attachments: Response 2009005.pdf

Dear Dr. Lonigan,

Attached is a response to the questions you raised in your June 15 message to the What Works Clearinghouse (WWC).

Thank you,

What Works Clearinghouse

-----Original Message-----
From: Christopher J. Lonigan, Ph.D. [mailto:lonigan@psy.fsu.edu]
Sent: Saturday, June 13, 2009 12:15 PM
To: Mark Dynarski
Cc: Jill Constantine
Subject: Problems with Newly Released Revised WWC Early Childhood Report
Importance: High

Hi Mark (and Jill),

I was on the WWC site this morning and came across the updated Doors to Discovery report. The report notes that the original WWC report erroneously included results from Assel et al. (2006). The stated error was that the study has severe subcluster attrition, which was not attended to in the original WWC report.

I think that this conclusion is incorrect.

Assel et al. (2006) reported that the sites included in the study had 603 children.

Their analysis sample--based on the information reported in Appendix A was 580.

Attrition within assigned group was .04 (Let's Begin), .03 (Doors), and .05 (Control) for the analysis sample.

This does not trigger either an overall or differential attrition concern.

Moreover, even if it did trigger such a concern, one can compute group equivalence at pretest from the information provided (i.e., means and SDs for each group at pretest).

In place of Assel et al., the WWC included the PCER report. Of course, the PCER report is a subsample of the Assel et al. study. Therefore, if one study report has some mysterious subcluster attrition problem, then the other does as well. PCER's analysis sample was a subset of the children in Assel et al (so, if the "attrition" issue is based on the difference between the N = 603 and the analysis sample, the PCER Study has an even bigger problem).

The "Extent of Evidence" thing also needs to go away. PCER (as like Assel) found no significant or substantively important effects. The revised report includes a summary of another study that randomized **4** classrooms to control versus Doors. The Extent of Evidence is rated as medium to large. Yet, the significant contribution to this rating is a study showing no significant or substantively important effects. The large, no impact of curriculum study is giving the very small, no significant impact study standing in this case. This is INSANE! The domain average effect size is a simple average instead of a weighted average, which is bizarre. If one applied a weighted average, based on the larger PCER sample (N = 183) and the smaller Christie et al. sample (N = 37), the domain average effect size would clearly be closer to the PCER's .16 (oral language) than to the Christie et al.'s .30.

MOREOVER, in the Christie et al. Study two of the classrooms served ELL children and two served English-only children. Type of child served was used as a blocking factor prior to randomization. This does not meet the standards of randomization. The only acceptable outcome would be if one of the ELL classrooms and one of the non-ELL classrooms ended up in each condition. If it were the case that this outcome did not obtain, curriculum condition would be confounded with type of student population and the study would fail. Because this confounding is not an acceptable randomization outcome, the study cannot really be considered randomized (i.e., there was not equal probability of which group a classroom ended up in). The randomization is based on two units (randomization of two of anything cannot equate). There seems to be no way that one could imagine that classrooms would end up being equated on expectation. Hence, one could argue that this is really a QED. There is also reported to be severe attrition (40%), and SUBSTANTIAL differences at pretest. The study fails standards.

So, there are three problems here:

1. The Assel et al. study (which is the largest of the studies) is rejected on what appears to be an erroneous rationale.

2. The Christie et al. study is included with a evidence standards rating that seems impossible to justify.

3. The INSANE application of extent of evidence.

The first of these two certainly seem like they are violations of WWC

procedures. This seems like a serious problem. I would like to hear how such decisions were justified. As a strong supported of WWC and someone who has worked inside WWC, it is import to me that the output fits the high quality rules-based system in place.

This has to be done right.

Chris

Christopher J. Lonigan, Ph.D.
Professor
Associate Director, Florida Center for Reading Research
Department of Psychology
Florida State University
1107 W. Call Street
Tallahassee FL 32306-4301