Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Volume 13, Number 4 ISSN 1099-839X

An Analysis of the Fidelity Implementation Policies of the What Works Clearinghouse

Jean Stockard
University of Oregon and National Institute for Direct Instruction

Citation

Stockard, J. (2010). An Analysis of the Fidelity Implementation Policies of the What Works Clearinghouse. *Current Issues in Education*, 13(4). Retrieved from http://cie.asu.edu/

Abstract

A large body of literature documents the central importance of fidelity of program implementation in creating an internally valid research design and considering such fidelity in judgments of research quality. The What Works Clearinghouse (WWC) provides web-based summary ratings of educational innovations and is the only rating group that is officially sponsored by the U.S. Department of Education. Yet, correspondence with the organization indicates that it disregards information regarding implementation fidelity in its summary ratings, relying on “replicated findings” and suggesting that any fidelity issues that “may have arisen are averaged.” This paper demonstrates the fallacy in this logic. Simulations show that the policy minimizes the positive impact of highly effective programs and the negative impact of highly ineffective programs. Implications are discussed.

Keywords: implementation fidelity, What Works Clearinghouse, evidence screening, systematic reviews, program evaluation

About the Author(s)

Author: Jean Stockard

Affiliation: University of Oregon and National Institute for Direct Instruction

Address: P.O. Box 11248, Eugene, Oregon 97440

Email: jstockard@nifdi.org

Biographical information: A sociologist by training, Jean Stockard is Professor Emerita at the University of Oregon and Director of Research for the National Institute for Direct Instruction. Her recent publications span areas of sociology of education, family, youth, demography, and methodology.

Current Issues in Education

Mary Lou Fulton Teachers College • Arizona State University
PO Box 37100, Phoenix, AZ 85069, USA

Analysis of Fidelity Implementation Policies

Prompted by the long-standing concern with promoting students' academic achievement, reviews and meta-analyses of studies of educational reforms are relatively common (e.g. Adams and Engelmann 1996, AFT 1998, Beck and McCaslin 1978, Borman et al 2003, Herman, et al 1999). In addition, several groups have web-based summaries of program effectiveness oriented toward educators and the public at large (e.g. Social Programs that Work, http://www.evidencebasedprograms.org/; Promising Practices Network, http://www.promisingpractices.net/; and the Best Evidence Encyclopedia, http://www.bestevidence.org/). Yet, only one of these, the What Works Clearinghouse (WWC), is endorsed by the U.S. Department of Education. The WWC was established in 2002 by the U.S. Department of Education's Institute of Education Sciences "to provide educators, policymakers, researchers and the public with a central and trusted source of scientific evidence of what works to improve student outcomes" (WWC Statement of Work, p. 1). Given this official sponsorship by the federal government, any systematic errors with its conclusions should be the focus of serious concern.

Although the Clearinghouse began reviewing research on various topic areas in education in 2003, summary ratings did not begin to appear until 2007. Despite its relatively short period of existence, the WWC's policies and procedures have received substantial scholarly and professional criticism. These criticisms have included concerns about 1) the scope of studies that are reviewed, both the inclusion of poor quality work as well as the exclusion of large numbers of relevant studies; 2) the quality of reviews, including reporting inaccurate and misleading summaries, using measures of outcomes that fail to match the studied curricula, and employing inappropriate statistical techniques; 3) a narrow research agenda including a bias toward small studies in artificial settings and of short duration; 4) ignoring contextual factors and issues of external validity; 5) failing to incorporate the cumulated scientific knowledge regarding research procedures in its conduct of reviews; 6) employing analysis techniques that result in minimizing the actual size of effects; 7) not including rigorous and transparent quality control measures such as peer review, comparisons with previous summaries, and reliability checks of ratings; and 8) the suppression of criticism, thus violating basic norms regarding the open quality of scientific inquiry (e.g. Briggs 2008, Chatterji 2005, 2008, Confrey 2006, McArthur 2008, Schoenfeld 2006, Slavin 2008, Stockard 2008).

Much of the substance of these critiques focuses on the WWC's strong preference for strongly controlled randomized experiments. For instance, in an extensive discussion of the WWC's procedures, Robert Slavin (2008) cites meta-analytic studies showing that quasi-experimental designs with well matched experimental and control groups provide results very similar to randomized experiments. In addition, Slavin suggests that the focus on highly controlled, randomized designs often results in highlighting short-term studies with small samples. He notes that this focus can potentially introduce bias and have low external validity.

Jeffrey Confrey (2006) compares the WWC's procedures to those used by the National Research Council's report *On Evaluating Curricular Effectiveness* (OECE). Like Slavin, Confrey criticizes the overreliance on randomized studies, detailing the way in which a sole focus on randomized designs ignores other issues in the quality of studies such as the nature of outcome measures, insights that are obtained through multiple methods of analysis, and issues regarding both internal and external validity. Confrey specifically criticized the way in which the WWC deals with the fidelity of implementation:

> In sum, if the intervention program is sufficiently described, and there is no evidence in the research report that particular kinds of disruptions occurred, the program is given a rating of "fully meets the criteria" for implementation fidelity. This process seems to underestimate dramatically the challenges associated with implementation.

> Based on the studies we reviewed for OECE, the variation within an implementation of a curriculum is substantial… (Confrey, 2006, p. 206).

This paper focuses on a specific aspect of the Clearinghouse's procedures for dealing with the fidelity of treatment interventions. This aspect has not received public scrutiny, but has the potential of resulting in misleading recommendations and conclusions. Below I expand on Confrey's statement, by discussing the well-established importance of treatment fidelity in research and describing the WWC's policies for addressing fidelity in their reviews. I then move to a description of the problems that result from the implementation of the WWC's policy: It results in summary ratings that minimize the positive impact of effective programs as well as the negative impacts of poor programs. Quite simply, it makes good programs look worse and poor programs look better than they actually are. The potential impacts are illustrated with examples using effect sizes and regression techniques. A concluding section discusses the potentially serious implications of the policy for educators, families, and students, as well as the public at large, and provides suggestions for change.

Treatment Fidelity and the WWC Policy

As summarized by O'Donnell (2008, pp. 33-34), "*Fidelity of implementation* (emphasis in original) is traditionally defined as the determination of how well an intervention is implemented in comparison with the original program design." The central importance of treatment fidelity to ensuring internal validity of an experiment has long been a standard element of strong research. Poor implementation of an intervention is a major threat to the internal validity of any type of research design, and numerous scholars call for paying close attention to fidelity of implementation in any type of review of research results (e.g. Crowley and Hauser 2007, Desimone 2002, Emshoff et al 1987, Gersten et al 2005, Haynes 1998, McMillan 2007, O'Donnell 2008, Ross 2007).

Somewhat surprisingly, the WWC website (as of October 7, 2009) pays relatively little attention to fidelity of implementation. A search of the website (using their search engine and the term "fidelity") revealed 26 reports on individual curricula that included information on fidelity within a given study. In addition, a document directed toward consumers and titled, *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide* (2003), listed fidelity as one of the "important factors to consider when implementing an evidence-based intervention in your schools or classrooms." The document notes that "whether an evidence-based intervention will have a positive effect in your schools or classrooms may depend critically on your adhering closely to the details of its implementation" (http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=14&tocId=7). I was not able to find any documentation on the website regarding how the fidelity of implementation affects the rating that a study may receive and nothing to indicate that Confrey's description, quoted above, has altered.

It was only through personal communications with the WWC that I was able to obtain additional details regarding the way in which they handle issues of fidelity. In a letter written to *Mathematica*, the organization that now holds the contract for WWC, in June of 2008, I mentioned the issue of treatment fidelity in the context of more general concerns with recognizing issues of both internal and external validity in their review procedures, echoing the criticisms cited above. The paragraph in which these issues were raised is reproduced below:

> Finally, the range of work considered has been limited by downgrading findings from studies that do not incorporate strict random assignment. While we all know that random assignment is the "gold standard" for experimental work, the WWC's over-reliance on this criterion ignores the realities of how school organizations work. With this strict attention to random assignment, other aspects of research designs that are

> an even greater threat to internal validity can be ignored. The most important of these is no doubt ensuring the fidelity of treatment implementation, making sure that a program is implemented as the developers designed it.

A response from a high-level WWC official was received in September of 2008:

> The letter notes that the WWC review process may downplay implementation fidelity. Definitions of implementation fidelity vary and many studies include little information to gauge fidelity, especially information about whether an intervention has been implemented within normal operating regimes of districts, schools, and teachers, not under specialized laboratory conditions. Moreover, there is no standard metric with which to rate and assess fidelity across studies that assures comparability. The WWC's approach emphasizes the importance of replicated findings, which ensures that any one study in which fidelity issues may have arisen are averaged with findings from other studies. Intervention reports include an "extent of evidence" classification that allows practitioners to place more weight if they choose on interventions for which the extent of evidence is large, meaning the results are drawn from multiple studies and a large number of classrooms and students.

Note that this response essentially confirms the description provided by Confrey – variations in implementation fidelity are simply ignored with an assumption that any bias will "balance out."

Problems with the WWC Policy

There are numerous logical and methodological problems with the policy outlined by the WWC. The sections below address these issues beginning with the claims regarding defining and measuring fidelity and then moving to empirical assessments of the claim that "replicated findings" provide a way to deal with any fidelity issues.

Defining and Measuring Fidelity

The first paragraph quoted above makes three general points regarding the definition and measurement of fidelity. First, it suggests that the WWC review process downplays the importance of fidelity because definitions of fidelity vary. Yet, as noted above, the definition of fidelity of implementation is well-grounded and highly developed within the field. In her extensive discussion of the area, O'Donnell (2008) cites at least 15 sources in the mainstream education literature that converge on the standard interpretation of the term. She also notes that even more studies discuss the term in areas outside of education. Contrary to the WWC's claim, the notion of fidelity of implementation as program integrity, the implementation of a program as it was intended, is long established and well accepted.

Second, the WWC letter suggests that "many studies include little information to gauge fidelity." The number of reports on the WWC website that include mentions of fidelity indicates that, in fact, many of them have sufficient information to gauge the extent of fidelity of implementation.[1] More important, however, is the point that a study's failure to include such information should be taken as a reason to question its conclusions (see Crowley and Hauser 2007, Emshoff et al 1987, Gersten et al, 2005, Haynes 1998, McMillan 2007, Ross 2007). Interestingly, two published criticisms of WWC decisions specifically address the WWC's decision to give high ratings to studies with serious fidelity issues (Slavin 2008, Confrey 2006, and see discussion in Briggs 2008, p. 18).

Third, the WWC letter notes that "there is no standard metric with which to rate and assess fidelity across studies that assures comparability." This statement can be logically challenged. Fidelity, by definition, is measured within a program, for the question is the extent to which an implementation adheres to the developer's guidelines. Logically, such comparisons would be within

[1] Because the WWC downplays the importance of fidelity of implementation, the number of studies that actually deal with this issue and were reviewed by the Clearinghouse may well be larger than the number found through the web search.

a program. Most well developed programs have guidelines for implementation and ways to assess the degree to which adopters meet those guidelines. Comparisons across programs could be easily made, such as using dummy variables to represent whether or not a program met the developer's standards (see Cook 2002, p. 186). A measure of fidelity must, by definition, be program specific, but an indicator of the extent to which fidelity occurred can easily be universal.

Can Fidelity Problems "Average Out?"

The second paragraph of the WWC's response quoted above claims that by consulting numerous reports "any one study in which fidelity issues may have arisen are averaged with findings from other studies." In other words, any problems with fidelity will simply "wash out," sometimes portraying a program as better than it really is and sometimes portraying it as worse. This assumption probably, however, only applies to programs that have little or no effect, that is, to programs in which students achieve, on average, only as well as (not better than or worse than) those in a control group. It is likely however, that even if there were no differences in central tendency (the means), poor fidelity would result in a larger variance. When programs are implemented in different ways, greater variability should occur.[2]

More generally, however, the assumption that the effect of fidelity problems will be random is seriously flawed. For both exemplary programs and for programs that are ineffective, poor implementation of a program would, very likely, produce results that are systematically biased. Thus, the WWC's policy has very serious consequences for consumers, minimizing the positive impact of good programs as well as the negative impact of poor programs (see also Engelmann 2008).

Basic to our analysis is the assumption of "regression toward the mean," the well-established statistical phenomenon where those with high scores, or low scores, on a measure will tend to have

[2] A recent article by Zvoch and associates (2007) included measures of fidelity in the analysis of multisite implementation of a childhood literacy program and found greater variability among the sites with low fidelity of implementation than among those with higher fidelity.

scores that are closer to the mean (lower for those who are high and higher for those who are low) at later testing periods. Simply through regression toward the mean we would expect those who score lower than a mean at pretest to score higher (and closer to the mean) at posttest. This concept can apply to programs as well as individuals. By chance, programs that typically do well will, on average, do less well (rather than even better) over time. Similarly, programs that do not perform well will, on average, do somewhat better (rather than worse) over time.[3]

These trends would be expected to be even stronger with poor implementation. A good program, when implemented poorly, would result in students doing less well than they would normally do. Similarly, a program that typically results in poor achievement would, when implemented with less fidelity, result in students doing better than they normally would with the program. With poor implementation, the good programs would be less good and bad programs would be less bad. In other words, *poor fidelity would result in biased results – and the nature of this bias is different for effective programs than for less effective programs*. The sections below illustrate the extent of this problem using two different methods: calculations of effect sizes and calculations of regression coefficients.

The Impact of Poor Fidelity on Effect Sizes. Tables 1 and 2 summarize the possible implications of inadequate fidelity of implementation for two different hypothetical curricular programs: one that is more effective than a control program (Table 1) and one that is less effective than a control (Table 2). Effect sizes (the difference between the mean scores of experimental and control groups divided by the common standard deviation) are used as the metric of comparison.[4]

[3] In lay person's terms, this phenomenon corresponds to common-day sayings such as, "when things are bad, the only way to go is up;" and "those on top are poised for a fall."

[4] Effect sizes are measures commonly used to estimate the strength of the relationship between two variables. Unlike tests of statistical significance, they are not affected by sample size and thus provide estimates that can be compared from one study to another. They are often used in meta-analyses and other quantitative summaries of research literature. This paper uses Cohen's d, a widely used measure (Cohen 1992).

For sake of illustration, it is assumed that the scores involved are normal curve equivalent (NCE) scores with, for the control group, a consistent mean of 50 and a standard deviation of 21.

Consider first the data in Table 1 regarding an exemplary program. The first column of Panels A and B of Table 1 gives four possible mean values of the experimental group that we will assume are the "true" values and the second column gives the values of Cohen's d that result when comparing these experimental means with those of the control group. It can be seen that the d values vary from .24, when the experimental mean is 55, to .95, when the experimental mean is 70. The higher values in this column are, in fact, similar to those often obtained in studies of highly effective programs, especially when they are implemented with fidelity (see Adams and Engelmann 1996). The lower values correspond to the level generally considered the point of educationally significant (d = .25, see Fashola and Slavin 1997).

The other columns in part A of Table 1 examine the likely result if the effective program is implemented with less than optimal fidelity, but the impact is simply the "random" effects that the WWC assumes will occur. In these calculations, as implied by the WWC assumption, the means of the experimental group do not change. However, the standard deviation is assumed to change, for with lower levels of fidelity, even if the means stay constant, greater variability would be expected. This is at the basis of the WWC notion of "averaging out." Three different

Table 1

Effect Sizes of Comparisons with an Effective Program with Different Levels of Fidelity

A. Random Influences of Fidelity Problems (The WWC Assumption)

		Alternative d			% Change from "True" d		
M Exp. Gp.	"True" d	SD = 25	SD = 30	SD = 35	SD = 25	SD = 30	SD = 35
55	0.24	0.22	0.20	0.18	-10%	-18%	-26%
60	0.47	0.43	0.39	0.36	-8%	-17%	-24%
65	0.71	0.65	0.59	0.54	-8%	-17%	-25%
70	0.95	0.87	0.78	0.71	-9%	-18%	-25%

B. Systematic Influences of Fidelity Problems with Smaller Experimental Means, but Constant Standard Deviation

		Effect Sizes		% Ch. from "True" d	
M Exp. Gp.	"True" d	M 5 pts less	M 10 pts less	M 5 pts less	M 10 pts less
55	0.24	0.00	-0.24	-100%	-200%
60	0.47	0.24	0.00	-49%	-100%
65	0.71	0.47	0.24	-34%	-66%
70	0.95	0.71	0.47	-25%	-51%

C. Systematic Influences of Fidelity Problems with Smaller Experimental Means and Varying Standard Deviations

	Effect Size				% Change from "True" d			
	M 5 pts. Less		M 10 pts less		M 5 pts. Less		M 10 pts less	
M Exp. Gp.	SD=25	SD=30	SD=25	SD=30	SD=25	SD=30	SD=25	SD=30
55	0.00	0.00	-0.22	-0.20	-100%	-100%	-190%	-182%
60	0.22	0.20	0.00	0.00	-54%	-58%	-100%	-100%
65	0.43	0.39	0.22	0.20	-39%	-45%	-69%	-72%
70	0.65	0.59	0.43	0.39	-31%	-38%	-54%	-59%

Note: If the impact of fidelity implementation is random (Panel A), it is assumed that this affects only the standard deviation of the experimental group and not the mean. If the impact of fidelity implementation is systematic (Panels B and C), both the mean and the standard deviation can be affected. Cohen's d is calculated by subtracting the mean of the experimental group from the mean of the control group and dividing by the common standard deviation. The control group is always assumed to have a mean of 50 and a standard deviation of 21.06, the mean and standard deviation of the experimental group varies as shown in the table. For the "true" condition, with perfect fidelity, both the experimental group and the control group are assumed to have the s.d. of 21.06, reflecting the definition of normal curve equivalent scores. The common standard deviation is calculated as the average of control and experimental group s.d. for each case, assuming that the two groups are of equal size.

values of the standard deviation are given, and it can be seen that in all situations the value of the effect size becomes lower than the "true" value. The final columns in Panel A report the percentage change from the "true" d value, with results ranging from eight to 26 percent. (Percentages were calculated by comparing the "true" values in the second column with the other values of d going across each row and using the "true" value as the base of the percent.)

Panel B and Panel C of Table 1 presents results that are more likely to occur when an effective program is implemented with less than adequate fidelity. In such a situation, it would be logical to expect that performance would be lower – that the mean value would decline. As with Panel A, the first two columns of Panel B give the "true" experimental mean and effect size for comparison purposes. The next two columns present the effect sizes that would occur if the average value of the experimental group declined by either 5 points (column 3) or by 10 points (column 4), but assuming no change in the standard deviation.[5] Again, all of the effect sizes are smaller, but the percentage change from the "true value" is much larger than in Panel A, ranging from 25 to 200 percent. Finally, the calculations in Panel C assume that lower fidelity with an effective program has two effects: the average score is lower than it would otherwise be and the standard deviation is larger, as in Panel A. Again, of course, the effect sizes are substantially lower and the declines are marked.

Now consider the possible implications of poor fidelity of implementation of a program that is, in reality, ineffective and, in fact, produces poorer results than the control group. These results are shown in Table 2. As with Table 1, the results with the "true" values are shown in the first two columns of Panels A and B. It can be seen that the "true" effect sizes vary from -.24 to -.95. The data in Panel A illustrate what would happen if the fidelity problems produced only random changes: the means would stay the same, but the standard deviations would become larger. It can be seen that, in all cases, the absolute values of the effect sizes become smaller – that is, less negative.

Table 2

Effect Sizes of Comparisons with an Ineffective Program with Different Levels of Fidelity

A. Random Influences of Fidelity Problems (The WWC Assumption)

		Alternative d			% Change from "True" d		
M Exp. Gp.	"True" d	SD = 25	SD = 30	SD = 35	SD = 25	SD = 30	SD = 35
45	-0.24	-0.22	-0.20	-0.18	-10%	-18%	-26%
40	-0.47	-0.43	-0.39	-0.36	-8%	-17%	-24%

[5] These differences in the means are probably reasonable estimates. A recent analysis (Stockard, forthcoming) compared reading achievement NCE scores between a control group and two implementations of a highly effective program – one with high fidelity and one with lower fidelity. The differences in NCE scores between the two experimental groups ranged between 5 and 17 NCE points depending on the measure and time during the implementation.

35	-0.71	-0.65	-0.59	-0.54	-8%	-17%	-25%
30	-0.95	-0.87	-0.78	-0.71	-9%	-18%	-25%

B. Systematic Influences of Fidelity Problems with Larger Experimental Means, but Constant Standard Deviations

		Effect Sizes		% Change from "True" d	
		Increased Mean		Increased Mean	
M Exp. Gp.	"True" d	5 pts.	10 pts	5 pts.	10 pts
45	-0.24	0.00	0.24	-100%	-200%
40	-0.47	-0.24	0.00	-49%	-100%
35	-0.71	-0.47	-0.24	-34%	-66%
30	-0.95	-0.71	-0.47	-25%	-51%

C. Systematic Influences of Fidelity Problems with Larger Experimental Means and Varying Standard Deviations

	Effect Size				% Change from "True" d			
	M 5 pts higher		M 10 pts higher		M 5 pts higher		M 10 pts higher	
M Exp. Gp.	SD=18	SD=15	SD=18	SD=15	SD=18	SD=15	SD=18	SD=15
45	0.00	0.00	0.26	0.28	-100%	-100%	-207%	-216%
40	-0.26	-0.28	0.00	0.00	-46%	-41%	-100%	-100%
35	-0.51	-0.55	-0.26	-0.28	-28%	-22%	-64%	-61%
30	-0.77	-0.83	-0.51	-0.55	-19%	-12%	-46%	-42%

Note: If the impact of fidelity implementation is random (Panel A), it is assumed that this affects only the common standard deviation and not the mean. If the impact of fidelity implementation is systematic (Panels B and C), both the mean and the standard deviation can be affected. Cohen's d is calculated by subtracting the mean of the experimental group from the mean of the control group and dividing by the common standard deviation. The "true" calculations assume that the standard deviation for both groups is 21.06, the values of the experimental mean and standard deviation vary as shown in each of the rows of the table. The common standard deviation is calculated as the average of the experimental and control group s.d., assuming that the groups are of equal size.

The results in Panels B and C illustrate the results if the impact of poor fidelity is systematic. With programs that are ineffective, it would be expected that poor implementation would lead to higher average scores (through regression to the mean) and smaller standard deviations. The smaller standard deviations result from having values that are “less ineffective.” As can be seen in Table 2, the result is that the effect sizes are less negative. That is, the extent to which tahe programs are ineffective, as shown by the “true” values of d, is disguised.

The Impact of Poor Fidelity on Regression Coefficients. The same substantive results occur if one addresses this issue using regression models. In other words, one reaches the same conclusion when approaching this issue with the language of regression analysis rather than the language associated with difference of means and effect sizes.

Assume that there are two groups, an experimental and a control group and that the two groups are of equal size. By definition, the variance of this grouping variable, which we will term x, is

$$\text{var}(x) = p(1-p) = .5(1-.5) = .25 \qquad (1).$$

Equation (1) simply reflects the computational formula for the variance in a binomial distribution as equal to the product of the two proportions (p * q), where q = 1 – p and where, with groups of equal size, p = .5.

Now suppose that a program is implemented with complete fidelity. In other words, there is no error at all in the implementation. For sake of example, we will assume that the "true" covariance between the treatment (x) and the outcome (y) (cov (xy)) is 4.[6] The regression coefficient (b_{yx}) is defined as the covariance of the treatment and outcome (cov (xy)) divided by the variance of the predictor, or treatment, variable. With the values defined here,

$$b^*_{yx} = \text{cov}(xy) / \text{var}(x) = 4 / .25 = 16 \qquad (2),$$

where b* refers to the "true" value of the regression coefficient. In other words, when the program is implemented with complete fidelity the regression coefficient (b_{yx}), which is defined as the difference between the experimental and control group (the two values of x), is 16 points.

Now suppose that the program is implemented with less than complete fidelity. In regression terminology, poorer fidelity results in lower reliability of the experimental treatment. This simply means that the differences that are observed between the experimental and control group do not reflect the "true" differences that exist. Observed differences reflect the true differences (var(x))

[6] The choice of "4" is arbitrary. The important point for the example is that it is the "true" value, or the covariance that occurs with perfect fidelity of implementation.

as well as error (termed var(e), for variance of the error term). In other words, under poor implementation, it is harder to tell the "real" or "true" differences between the two groups, and the observed variance of this grouping variable reflects both true differences and error.

As a result, when a program is implemented with less than complete fidelity, the estimates of the regression coefficient are altered:

$$b_{yx} = \text{cov}\ (xy)\ /\ [\text{var}(x) + \text{var}(e)] \qquad (3).$$

(Note that the asterisk has been omitted from the b coefficient to indicate that it is not the "true" value.) The covariance is unchanged, but the estimate of the variance of the independent variable is increased because it now includes the error resulting from less than perfect implementation (var(x) + var (e)). The result is simple. With lower reliability (higher error) the denominator of equation 3 becomes larger and the absolute value of b decreases. In other words, the error introduced through poor implementation obscures the "true" difference between the experimental and control group and increases the estimated variance of the independent variable (by var (e)). This change in the estimate of variance alters the regression coefficient.

In a regression framework, varying levels of fidelity of implementation are usually illustrated through the use of reliability coefficients, a standard method of estimating measurement error. By definition, reliability coefficients vary from 1.0, indicating no error in measurement and thus perfect implementation, to 0.0, indicating simple random implementation procedures. The computational formula for the reliability coefficient, r_{ii}, is

$$r_{ii} = \text{var}\ (x)\ /\ [\text{var}(x) + \text{var}\ (e)] \qquad (4).$$

Thus, in a situation of error-free implementation, when the error variance equals zero,

$$r_{ii} = \text{var}\ (x)\ /\ [\text{var}(x) +\ 0] = \text{var}(x)/\text{var}(x) = 1.0 \qquad (5).$$

However, as implementation fidelity declines and the error variance increases, the denominator of equation 4 rises and the reliability coefficient declines.

From rearranging equation (4), we can estimate the sum of the variance of x and the error variance as the variance of x divided by the reliability coefficient:

$$\mathrm{var}(x) + \mathrm{var}(e) = \mathrm{var}(x) \ / \ r_{ii} \qquad (6).$$

Substituting in equation (3), we can then estimate the regression coefficient, b_{yx}, as

$$b_{yx} = \mathrm{cov}\ (xy) \ / \ [\mathrm{var}(x)/r_{ii}] = [r_{ii} * \mathrm{cov}(xy)]/\ [\mathrm{var}\ (x)] \qquad (7).$$

Given that cov (xy) /var (x) is the “true” value of the regression coefficient (b^*_{yx}), the biased estimate of b_{yx} that results from decreasing reliability can be written simply as

$$b_{yx} = (r_{ii}\ )(\ b^*_{yx}) \qquad (8),$$

where b^*_{yx} is the “true” value of b_{yx} and r_{ii} is the reliability coefficient. In other words, the estimated value of the regression coefficient, or the difference between the experimental and control group, is linearly related to the reliability coefficient. If there is perfect reliability (perfect implementation or no error), the estimated coefficient equals the true coefficient ($r_{ii} = 1.0$ and $b^*_{yx} = b_{yx}$). However, as reliability declines the absolute value of the estimated coefficient becomes lower than the true coefficient.

Table 3 summarizes this pattern of results for programs with positive effects and programs with negative effects. Three different covariance values are included, with absolute values ranging from 2.0 to 4.0 (across columns two through seven). The first column on the left lists different levels of reliability, from a coefficient of 1.0, indicating perfect fidelity, to .30, indicating very low reliability and low fidelity. Values within the table are the regression coefficients, b_{yx}, which were computed using equation 8, for each level of reliability. As would be expected, the absolute value of the estimate of b declines linearly with decreasing reliability levels. Most importantly, as with the results with effect sizes, both the positive impact of effective programs and the negative impact of ineffective programs are obscured. With effective programs, the positive impact is smaller with less reliability (lower fidelity), as indicated by

Table 3

Regression Coefficients for Varying Levels of Association and Varying Levels of Fidelity

	Covariance (x,y) Effective Programs			Covariance (x,y) Ineffective Programs		
Reliability	4	3	2	-4	-3	-2
1.0	16.0	12.0	8.0	-16.0	-12.0	-8.0
0.9	14.4	10.8	7.2	-14.4	-10.8	-7.2
0.8	12.8	9.6	6.4	-12.8	-9.6	-6.4
0.7	11.2	8.4	5.6	-11.2	-8.4	-5.6
0.6	9.6	7.2	4.8	-9.6	-7.2	-4.8
0.5	8.0	6.0	4.0	-8.0	-6.0	-4.0
0.4	6.4	4.8	3.2	-6.4	-4.8	-3.2
0.3	4.8	3.6	2.4	-4.8	-3.6	-2.4

Note: The values in the cells represent the regression coefficient (b) for each level of reliability and for different values of the covariance, calculated using equation 8 in the text. Lower reliability coefficients correspond to lower levels of fidelity of treatment (larger error variances).

smaller coefficients. With ineffective programs, the negative impact is also smaller with poor implementation fidelity. Thus, with lower fidelity the "true" difference of both effective and ineffective programs is obscured.

Discussion

The results presented above seem to refute the assumption that guides the WWC's approach to considering the fidelity of implementation in its deliberations and summary ratings. Even if the results of poor fidelity of implementation were random, the efficacy of both good and poor programs would be misrepresented. If, as is more likely, the results of poor implementation are not random, the impact would be even greater. Good programs appear less effective and poor programs appear better than they actually are.

The logic and calculations presented in this paper are simple, but the implications are potentially serious. Even though the chance of a false negative (concluding that a good program is harmful) or false positive (concluding that a harmful program is good) is not high,[7] the procedures that the WWC has adopted would appear to almost definitely promote inaccurate conclusions. Their ratings would automatically minimize the extent to which individual programs can either help or hurt students. In addition, the procedures lead to false impressions regarding the extent to which educational programs have the potential to change achievement patterns by minimizing accurate information about the actual range of change when programs are well implemented.

The WWC policy may result in especially misleading results regarding effective programs, which are, of course, of most interest to policy makers, educators, and families. Effective programs may be more difficult to implement precisely because they address numerous factors that affect student performance. Such models are typically "extensive-requirements" reform programs, rather than "minimal-requirements" programs. The greater complexity and intricacy of the models, while resulting in substantially stronger results, also present greater challenges to full implementation and high fidelity (Engelmann and Engelmann 2004). It would not be unreasonable to expect that issues with fidelity could affect a higher proportion of such extensive requirements models than those that were simpler.

The impact of the WWC policy should also be considered in light of the very few studies that the WWC has determined meet their standards of evidence. While their reviews of individual curricula generally indicate that numerous studies, often dozens, have been examined, only a handful, usually less than three for any particular program, are deemed to have met the set criteria for inclusion. As noted earlier, the standards and their application have received an extraordinary amount of criticism. Yet, the simple fact that so few studies are actually reviewed to develop the summary ratings makes the importance of considering fidelity of implementation even more crucial.

[7] With the regression simulation, the calculated values of b never cross zero, but come increasingly close to this point as the error variance approaches infinity. This result reflects the constant covariance used in the analysis. With the effect size simulations, in which the difference of means varies, some of the simulations result in changing signs from positive to negative or negative to positive.

Given the very small number of studies on which the summaries are based, it is extremely important that the possibility of systematic bias be removed. It could be relatively easy to add such elements to a review procedure. As noted earlier, simple rating scales could indicate the extent to which study reports note fidelity of implementation. Those with low fidelity to a given model could be discounted or omitted from analyses.

More generally, this paper adds to the already large body of critiques of the procedures of the WWC. The literature is replete with well-developed guidelines for reviewing literature and developing systems of consolidating findings in ways that pay careful attention to issues of both internal and external validity – including issues of fidelity. The results of the analysis above suggest that the WWC would benefit from consulting this literature closely. If the Clearinghouse were simply a project of a lone faculty member or small nonprofit with an obscure website the issues would be far less serious. However, given the extensive funding received by the Clearinghouse (in excess of $50 million at last report) as well as the endorsement of the federal Department of Education, one could suggest that students, families, the educational community, and the general public deserve more.

References

Adams, G. L. & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle: Educational Achievement Systems.

American Federation of Teachers. (1998). *Building on the best, learning from what works: Seven promising reading and language arts programs*. Washington, D.C.: AFT.

American Psychological Association. (2001). Publication Manual of the American Psychological Association (5th ed.). Washington, DC: Author

Beck, I.L. & McCaslin, E.S. (1978). *An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs*. LRDC Report No. 1978/6 Pittsburgh.

Borman, G. D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73,* 125-230.

Briggs, D. C. (2008). Synthesizing causal inferences. *Educational Researcher 37*, 15-22.

Chatterji, M. (2005). Evidence on "What Works": An argument for extended-term mixed-method (ETMM) evaluation designs *Educational Researcher 34*, 14-24.

Chatterji, M. (2008). Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher 37,* 23-26.

Cohen, J (1992). "A power primer". *Psychological Bulletin* 112: 155–159

Confrey, J. (2006). Comparing and contrasting the national Research Council Report *On Evaluating Curricular Effectiveness* with the What Works Clearinghouse Approach. *Educational Evaluation and Policy Analysis 28* (3) 195-213.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis 24*, 175-199.

Crowley, J. J. & Hauser, A.G. (2007). Evaluating whole school improvement models: Creating meaningful and reasonable standards of review. *Journal of Education for Students Placed at Risk, 12,* 37-58.

Desimore, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72,* 433-479.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W.S., & Erickson, S. (1987). Innovation in Education and Criminal Justice: Measuring Fidelity of Implementation and Program Effectiveness. *Educational Evaluation and Policy Analysis, 9,* 300-311.

Engelmann, S. (2008). *Machinations of the What Works Clearinghouse*. Unpublished Paper, http://www.zigsite.com/PDFs/MachinationsWWC%28V4%29.pdf (last downloaded March 31, 2010).

Fashola, O. S. & Slavin, R.E. (1997). Promising programs for elementary and middle schools: Evidence of effectiveness and replicability. *Journal of Education for Students Placed at Risk, 2,* 251-307.

Gersten, R., Fuchs, L.S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M.S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71,* 149-164.

Haynes, N. M. (1998). Lessons learned. *Journal of Education for Students Placed at Risk, 3,* 87-99.

Herman, R., Aladjam, D., McMahon, P., Masem, E., Mulligan, I., Smith, O., O'Malley, A., Quinones, S., Reeve, A., & Woodruff, D. (1999). *An educator's guide to schoolwide reform*. Washington, D.C.: American Institutes for Research.

McArthur, G. (2008). Does What Works Clearinghouse work? A brief review of Fast ForWord. *Australasian Journal of Special Education, 32,* 101-107.

McMillan, J. H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical assessment, Research and Evaluation, 12* (15).

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33-84.

Ross, S. M. (2007). Achievements, challenges, and potential advancements in reviewing educational evidence for consumers. *Journal of Education for Students Placed at Risk, 12,* 91-100.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35 (2),* 13-21.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37,* 5-14.

Stockard, J. (2008). *The What Works Clearinghouse Beginning Reading Reports and Rating of Reading Mastery: An Evaluation and Comment.* Technical Report # 2008-4. Eugene, Oregon: National Institute for Direct Instruction.

Stockard, J. (forthcoming). Direct instruction and first grade reading achievement: The role of technical support and time of implementation. *Journal of Direct Instruction*.

Zvoch, K., Letourneau, L.E., & Parker, R.P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation, 28*, 132-150.

Current Issues in Education

http://cie.asu.edu

Volume 13, Number 4 ISSN 1099-839X

ASU Mary Lou Fulton Teachers College
ARIZONA STATE UNIVERSITY