National Institute for Direct Instruction

Throwing the Baby Out With the Bathwater?

The What Works Clearinghouse Criteria for Group Equivalence*

A NIFDI White Paper

December 13, 2013

Jean Stockard

Professor Emerita, University of Oregon

Director of Research and Evaluation, National Institute for Direct Instruction

jeans@uoregon.edu

jstockard@nifdi.org

*The following people provided helpful comments on this paper: Douglas Carnine, Kurt Engelmann, Zig Engelmann, Robert M. O'Brien, Jerry Silbert, Timothy Wood, and J. Walter Wood. Any errors or opinions expressed are the sole responsibility of the author.

Throwing the Baby Out With the Bathwater?
The What Works Clearinghouse Criteria for Group Equivalence

Abstract

The What Works Clearinghouse (WWC) provides summary reports of the effectiveness of educational programs and practices, but only a small proportion of studies considered for review meet their selection criteria. A common reason for rejecting studies from consideration regards the WWC's standard for equivalence of the intervention and comparison groups. This paper examines the criteria related to these decisions. Calculations based on the Central Limit Theorem illustrate how the probability of meeting the WWC criteria declines sharply when studies include multiple measures and/or comparisons (which is common in more sophisticated studies) and/or when sample sizes are smaller (which is common in highly controlled experimental designs). Descriptions of two well-designed studies rejected for inclusion in recent WWC reports illustrate the nature of the problem. Implications and recommendations for policy change are discussed.

Throwing the Baby Out With the Bathwater?
The What Works Clearinghouse Criteria for Group Equivalence

The What Works Clearinghouse (WWC) was established in 2002 as an initiative of the Institute for Education Sciences (IES) in the U.S. Department of Education. Its website describes its purpose:

> The goal of the WWC is to be a resource for informed education decision making. To reach this goal, the WWC *identifies studies that provide credible and reliable evidence* of the effectiveness of a given practice, program, or policy (referred to as "interventions"), and disseminates summary information and reports on the WWC website (WWC, 2013a, emphasis added).

While other groups review educational programs and issue summary reports, the WWC, because it is an official arm of IES, is arguably the most powerful and influential of such groups and thus is the focus of this paper.

In developing its reports the WWC uses a number of criteria to determine if studies are sufficiently rigorous to be reviewed, providing the "credible and reliable evidence" noted above. Yet, according to a 2010 report by the General Accounting Office, fewer than ten percent of the studies examined pass this screening (GAO, 2010, p. 13). For some programs, the percentage is far less (e.g. WWC, 2013b, c). One of the common reasons for rejecting studies is a conclusion that the intervention and comparison groups are not equivalent.

This paper examines the criteria that the WWC uses to reach this conclusion, using sampling theory and the Central Limit Theorem to calculate the probability that studies could meet the WWC standards. The analysis shows how difficult it is for studies to meet the criteria and the ways in which having multiple measures or comparisons within a study and/or smaller samples increases the probability that a study will be excluded. Ironically, the randomized control trials favored by the WWC (see WWC, 2013d, pp. 9-10) and other contemporary review groups often have relatively small samples. Studies considered to be high quality by the academic community generally include multiple dependent measures and/or comparisons. Thus, meeting the criteria typically used by review groups and the academic community to define high quality studies actually decreases the probability they will meet the WWC's criteria for inclusion. Descriptions of two well-designed studies that were rejected for inclusion in recent WWC reports illustrate the nature of the problem. The paper concludes with a short discussion of implications and recommendations.

Probabilities of Meeting the WWC Criteria

The WWC guidelines for determining if two groups that are compared in a study are equivalent require that information be provided on the groups' comparability prior to the intervention

> on observable characteristics….If the reported difference of *any* (emphasis added) baseline characteristic is greater than .25 standard deviations (based

> on the variation of that characteristic in the pooled sample), the intervention and comparison groups are judged to be not equivalent….For differences in baseline characteristics that are between .05 and .25 the analysis must include a *statistical adjustment* (emphasis in original) for the baseline characteristics to meet the baseline equivalence requirement. Differences of less than or equal to 0.05 require no statistical adjustment (WWC, 2013d, p. 13).

In other words, if any baseline (pretest) characteristic (such as a mean) of an experimental and control group differ by more than .25 of a standard deviation, the study is excluded from consideration. If the difference falls between .05 and .25 of a standard deviation, statistical adjustments must be used. There are several other requirements to establish baseline equivalence (WWC, 2013d, pp. 14-15) including the presence of either pre-intervention measures that are "analogous" to post-intervention measures or the use of control variables specified by the WWC and demonstrating equivalence separately for each outcome. These requirements are not the focus of the present discussion.

The stipulations regarding the magnitude of acceptable baseline differences appear to be extraordinarily stringent. The paragraphs below use basic sampling theory to calculate the probability of researchers obtaining samples that would meet these criteria. The probability of meeting the criteria is not large and declines substantially when more than one measure is used and/or when sample sizes are smaller. Calculations are given for samples of size 64 and 36, typical of many studies examined by the WWC.

Example with Sample Size of 64 per Group

Suppose that a researcher were interested in selecting two random samples from a normally distributed population, with a mean of μ and a standard deviation of σ. Sampling theory tells us that if repeated random samples were drawn from this population, they would comprise a normal distribution, the sampling distribution, with a mean of μ and a standard deviation of ($\sigma/\sqrt{n}$), where n is the sample size. The standard deviation of the sampling distribution is called the standard error. In other words, the mean of an infinite number of drawn samples equals the population mean, the standard error is a function of the standard deviation of the population and the sample size, and the distribution assumes the shape of a normal curve. We can use this logic (the Central Limit Theorem) to examine the probability that two randomly drawn samples, typical of those that the WWC prefers to have in studies that it examines, would have characteristics that met the criteria described above.

Consider a population with a mean of 50 and a standard deviation of 21, roughly equivalent to the Normal Curve Equivalent (NCE) distribution often used in education research. Suppose that a researcher drew two samples, each with 64 cases, from this population and designated one as the treatment group and one as the control group. For simplicity's sake we will assume that one of these samples (Sample A) perfectly represents the population, with a mean of 50 and a standard deviation of 21. (Note that this

assumption is conservative in nature, resulting in the maximum probability of cases matching the WWC criteria.) To meet the WWC criterion of a difference at baseline of less than .05 of a standard deviation, the mean of the second sample (Sample B) would need to be within (.05 * 21 =) 1.05 points of the mean of Sample A, falling between 48.95 and 51.05. To meet the criterion of .25 of a standard deviation, Sample B would need to have a mean within (.25*21=) 5.25 points of the mean of Sample A, falling between 44.75 and 55.25.

We can use basic sampling theory to estimate how likely such an outcome would be. We begin by calculating the probability that one sample would be greater than .05 s.d. away from the population mean, assuming that the samples each have an n of 64. In other words, what is the probability that Sample B would have an average between 48.95 and 51.05 [P(48.95<=M<=51.05)]? Given that the mean of the sampling distribution would be 50 and the standard error (the standard deviation of the sampling distribution) would be $21/\sqrt{64} = 21/8 = 2.65$, we can calculate the z scores associated with each of these values:

$Z_{48.95} = (48.95 - 50)/2.65 = -.40$ and,

$Z_{51.05} = (51.05 - 50)/2.65 = +.40$.

Using a normal curve table we find that the probability of falling between these two values, or

P[48.95<=M<=51.05] = .1554 + .1554 = .3108.

Thus, the probability of choosing a sample that differs from the population mean (and by assumption the mean of Sample A) by less than .05 of a standard deviation is .31. The probability that the difference is larger than that amount, and thus subject to more stringent criteria by the WWC, is .69 (=1.00 - .31).

Suppose that the researcher was looking at three separate variables (e.g. fluency, comprehension, and vocabulary, similar to the dimensions reviewed by the WWC for analyses of beginning reading), each of which came from a population with an NCE distribution. Also assume that Sample A mirrors the characteristics of the population on each of these variables and that the sample size for each group (Sample A and Sample B) is 64. For each of the three separate measures the probability of obtaining a sample that fell within .05 s.d. of the population mean would be .31. But, having such an outcome for all three of the variables would only be .31*.31*.31 = .029. In other words, if a researcher were to examine three outcomes, the probability that all three of these scores would be within .05 of a standard deviation of the mean would be only .03. The probability that the measures would not meet the criteria would be .97. The probability of having samples that met the criteria when 5 measures are examined is much lower. (See the first line of data in Table 1.)

Consider the .25 of a standard deviation outcome, or the probability that the two samples would differ by more than .25 of a standard deviation and thus be rejected under the WWC criteria from any consideration, even with statistical adjustments. Again, using the NCE distribution, as described above, .25 of an s.d. = 5.25. So, paralleling the logic above, one may determine the probability that a sample would have an average between 44.75

and 55.25, or P[44.75<=M<=55.25]. Using the mean of the sampling distribution (50) and the standard error (2.65) given above, we can calculate the z scores associated with each of these values:

$Z_{44.75}$ = (44.75 - 50)/2.65 = -1.98 and,

$Z_{55.25}$ = (55.25 – 50)/2.65 = +1.98

Using a normal curve table we find that the probability of falling between these two values is .94, approximately equal to the standard 95 percent confidence interval. And thus, the probability of having a sample value that was greater than this level, given the sample size of .64, equals .06.

Again, however, with multiple dependent measures, the probability of falling outside the acceptable range becomes greater. Using the logic outlined above, if the probability of obtaining one sample that fell with .25 s.d. of the population mean is .94, having such an outcome for all three of the variables would be .94*.94*.94 = .83. In other words, if a researcher were to examine three outcomes, the probability that all three of these scores would be within .25 of a standard deviation of the mean would be .83. There would be almost a 20 percent probability that at least one of the three measures would fail to meet the criteria and the study would then be excluded by the WWC. If the researcher examined 5 variables, the probability that the study would fall within the acceptable range (with all comparisons differing from Sample A by less than .25 s.d., but of course requiring additional statistical controls) would be .73. (See the second line of data in Table 1.)

Example with Sample Size of 36 per Group

The issue becomes more difficult when samples are smaller, for then the standard error becomes larger. Let us assume that the sample size for each group is 36. Then the standard error = 21/√36 = 21/6 = 3.5, substantially larger than with the sample size of 64. Consider first the criterion of having a sample mean within .05 s.d. (or +- 1.05) of the population mean. In this case,

$Z_{51.05}$ = 1.05/3.5 = .34 and $Z_{48.95}$ = -1.05/3.5 = -.34

Using a normal curve table one can find that there is a .26 (.13+.13) probability that the mean of Sample B, when the samples have an n of 36, will fall within .05 s.d. of the mean Sample A, which, by definition, is equivalent to the population mean. The probability that the samples will differ by more than .05 s.d. is .74 (=1.00 - .26). If three measures are involved, the probability, with a sample size of 36 for each group, that all three measures will meet this criterion is only .02 (.0175). With more measures in the analysis the probability is, of course, even lower.

The last line of Table 1 reports the same calculations for a sample size of 36 and utilizing the criterion of sample B differing by .25 s.d. from Sample A. With one measure the probability of meeting the criterion is .86, but the probability declines as more measures are included in the analysis. With 5 measures included the probability that a researcher using a sample size of 36 would meet the criterion of a difference of sample means of only .25 s.d. is less than .50.

In short, this demonstration shows how difficult it is for research projects to meet the WWC criteria regarding baseline equivalence of sample groups. The WWC strongly prefers randomized control trials, which are typically quite small, yet the probability of meeting the criteria declines as samples become smaller. In addition, the WWC reports include results from several measures. Yet, the more measures that a study reports, the greater is the chance that it will not meet the WWC criteria for baseline equivalence.

The Criteria in Practice

The discussion above helps explain why the WWC has, to date, found very few studies that meet their selection criteria. Two examples illustrate the way in which the criteria have been applied. Both studies were well designed and published in top ranked journals in the field. Both were rejected for consideration by the WWC because the comparison groups were not equivalent at baseline.

The first is a quasi-experimental study, supported by grants from IES, the body that supports the WWC, and the National Institute for Child Health and Human Development (NICHHD) (Crowe, Connor, & Petscher, 2009). The study was rejected for consideration in the WWC's analysis of Reading Mastery for Beginning Readers because "it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent" (WWC 2013b, p. 2). Crowe and associates examined growth in reading achievement during one school year of over 30,000 students in grades one to three who were randomly selected from almost 3000 classrooms. They compared changes in oral reading fluency from fall to spring of students in six different curricula using growth curves and hierarchical linear modeling. The authors reported descriptive statistics at baseline on oral reading fluency for each of the groups in the analysis (p. 192) and for the total group. Of the 15 possible comparisons of a curriculum with *Reading Mastery (RM)*, the subject of the WWC report, three exceeded the .25 criterion set by the WWC. On average, the *RM* sample differed from the other groups by .12 of the total s.d., while the absolute value of the deviations ranged from .03 to .40. As explained above, the WWC's criteria for inclusion of studies states that "If the reported difference of *any* (emphasis added) baseline characteristic is greater than .25 standard deviations…the intervention and comparison groups are judged to be not equivalent" (WWC, 2013d, p. 13). The fact that three of the fifteen comparisons exceeded this level apparently resulted in the study being rejected, even though the statistical analyses nicely controlled for any baseline differences. (Interestingly, the pretest differences did not exceed the .25 criterion for one of the three grades examined, but the WWC rejected all of the results from consideration.)

The second example involves a randomized control design reviewed in the 2013 WWC analysis of *Reading Recovery*, a tutoring program for primary students at risk of future reading difficulties. The authors (Iversen & Tunmer, 1993) used a standard and well regarded method of controlled random assignment of subjects, matching 34 triplets of first grade students on beginning literacy skills and then randomly assigned members of the triplet to one of three groups. As with the Crowe et al. study, the WWC rejected this article for

review because some of the differences between the intervention groups were greater than .25 of the pooled standard deviation (WWC 2013c, p. 34). The authors reported pretest data on 10 measures, resulting in 30 comparisons between the groups (Iversen & Tunmer, 1993, p. 119). Of the 30 comparisons between pretest means, 6 were larger than .25 of a standard deviation. The differences ranged from 0 to .55 of a standard deviation, and the average difference was .06 s.d. Note that the sample size in this study of 34 students per group is slightly less than that in the final two rows in Table 1 above. Using the methods described above, the probability that all 30 comparisons would be less than .25 s.d. is .005. The probability that all 30 comparisons would be less than .05 s.d., and thus not require further statistical adjustment, is very remote: 2.03^{-20}. In other words, simple calculations based on the Central Limit Theorem would indicate a very small chance, given the sample size and number of comparisons, that Iversen and Tunmer's study would pass the WWC's criteria for baseline equivalence.

Summary and Discussion

Both of these examples illustrate how difficult it can be for a well-designed study to meet the criteria established by the WWC regarding equivalence of study groups. Even though the sample used by Crowe, et al. was extraordinarily large and used stringent and highly regarded analytic methods, differences emerged on a small proportion of the comparisons that were larger than the WWC set criterion. The chance of such differences emerging was, of course, heightened by the multiple comparisons included. The randomized design of Iversen and Tunmer had a substantially smaller sample than used by Crowe and associates (although, with a total n of 102, larger than that in many randomized studies) and employed a relatively large number of measures. The analysis in the first section of this paper shows how, simply by chance, differences that surpass the WWC criteria would be likely to occur. Ironically, while such multiple comparisons and careful designs make studies valuable within the academic research community, they greatly heighten the probability that they will not be accepted by the WWC.

The rational response of researchers who want to have their studies accepted by the WWC could be to limit their studies to very few measures and analyses. For instance, if Iversen and Tunmer had reported on only some of the measures in their analysis, their work would probably have met WWC acceptance criteria. If Crowe and associates had reported results from only one grade level, rather than three, their results would also have potentially been accepted. Yet, to have done so would have made their findings far less valuable for both educational researchers and the general public. It appears clear that the extraordinarily stringent requirements for group equivalence, simply by statistical realities, can result in the rejection of the vast majority of relevant studies and especially those that are more sophisticated. In other words, the application of the current criteria for group equivalence will often result in "throwing the baby out with the bath water."

A number of changes to the WWC criteria for group equivalence could address this issue. First, it would seem appropriate to accept all studies that use randomized assignment

regardless of the extent of pretest differences. It is difficult to envision any rational, statistical justification for trying to improve upon randomization. Second, social scientists have developed sophisticated statistical techniques for analyzing data from quasi-experimental and field settings and adjusting for pretest differences, such as the growth models used by Crowe and associates. Given the power and wide acceptance of these approaches within the academic research community, it would seem reasonable to accept studies that use such statistical controls for analysis. In addition, because quasi-experimental, field based designs can have higher external validity than the more tightly controlled, yet often relatively artificial, characteristics of randomized control trials, the inclusion of studies with such statistical controls would potentially be of even greater importance to the development of sound educational policy (McMillan, 2007, author 2013). Third, given that the chance of finding cases that violate the criteria of group equivalence increases markedly with multiple measures and comparisons, it would be appropriate to look at the average difference across all measures and comparisons rather than omitting an entire study when only one difference surpasses a given level. Fourth, the WWC should consider any differences in pretest equivalence in conjunction with the magnitude of treatment effects. If treatment effects are large the criteria for similarity of baseline measures should be modified accordingly.

The WWC criteria for group equivalence are one element of the "exclusive" approach it has taken to selecting studies for review, using a variety of criteria to control for methodological variations and limit the set of studies examined. Having to also pass these other criteria increases the probability that a given study will not meet the WWC standards for inclusion in a review. In short, the WWC approach appears to involve a painstaking search for a seemingly very elusive "perfect study." The result, as noted above, is that the reports focus on a limited slice of possible studies and, of course, the probability of error in conclusions is greatly heightened when a sample of studies is so small and selective.

In contrast to the WWC approach, the classic methodological literature explains that there can be no "perfect" experiment and that is important to look at a variety of results in a range of settings. As Cook and Campbell, authors of one of the most influential books on research design, put it, "we stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results" (1979, p. 31, emphasis in original). Meta-analyses and the tradition of "generalized causal inference" (Shadish, Cook, and Campbell, 2002) are representative of this long accepted approach. The WWC could adopt this more inclusive, and more currently accepted, methodology by examining all of the available evidence. The methodological variations in the studies could be noted and empirical analyses could examine the extent to which these variations affect the results. Based on the data presented in this paper it seems reasonable to suggest that, if the WWC used an approach that examined the totality of the research base, it would be much more likely to meet its stated goal of providing "credible and reliable evidence."

Table 1

Probability that Samples Would Fail to Meet WWC Criteria Regarding Equivalence of Baseline Measures by Sample Size and Number of Dependent Variables

		Probability of Meeting Criteria		
Level	Sample Size	1 Measure	3 measures	5 measures
.05 s.d.	64	0.31	0.03	0.003
.25 s.d.	64	0.94	0.83	0.73
.05 s.d.	36	0.26	0.01	0.001
.25 s.d.	36	0.86	0.64	0.47

Note: As explained in the text, the probabilities are based on the assumption that one sample mirrors the population exactly. If this assumption did not hold, the probabilities of two samples meeting the criteria would be even smaller.

References

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. Journal of School Psychology, 47, 187–214.

Government Accountability Office (2010). *Department of Education: Improved dissemination and timely product release would enhance the usefulness of the What Works Clearinghouse* (GAO-10-644). Washington, D.C. GAO.

Iversen, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85*(1), 112–126.

McMillan, J.H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment, Research & Evaluation, 12* (15). http://pareonline.net/pdf/v12n15.pdf

Shadish, W.R.. Cook, T.D. & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

What Works Clearinghouse (2013a). About us. Washington, D.C.: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/aboutus.aspx, retrieved September 9, 2013.

What Works Clearinghouse (2013b). *WWC intervention report, Reading Mastery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved December 13, 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_ReadingMastery_081208.pdf

What Works Clearinghouse (2013c). *WWC intervention report, Reading Recovery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved December 13, 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_readrecovery_071613.pdf

What Works Clearinghouse (2013d). WWC procedures and standards handbook (Version 3.0). Washington, D.C.: Institute of Education Sciences. Retrieved June 28, 2013, from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19