What's New in Version 4.1 of the WWC Standards and Procedures Handbooks?

Dr. Jeffrey Valentine
University of Louisville

Dr. Joseph Taylor
American Institutes for Research/University of Colorado, Colorado Springs

Dr. Sarah Caverly
American Institutes for Research

Looking for the Version 4.1 Handbooks while you wait? Go here to find them:
https://ies.ed.gov/ncee/wwc/handbooks

Presenters

Joseph Taylor, Ph.D.

American Institutes for Research

–

University of Colorado, Colorado Springs

Jeff Valentine, Ph.D.

University of Louisville

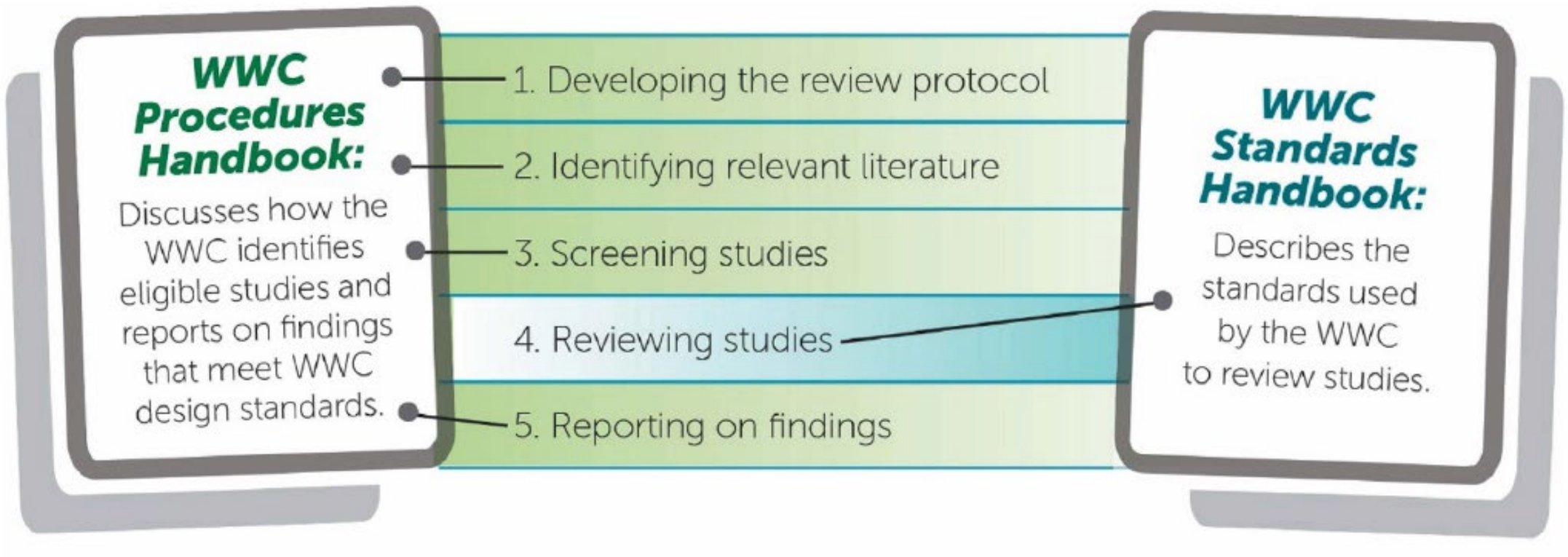
Learning Goals for the Webinar

During this webinar, you will:

- Learn in detail about four substantive changes in Version 4.1 of the WWC Procedures:
  - Estimating effect sizes when multiple effects are present in a single domain
  - Using fixed-effects meta-analysis for synthesizing effects across studies
  - A new approach to intervention report effectiveness ratings
  - Calculating effect sizes in single-case designs (SCDs).
- Learn about additional changes in Version 4.1 of the WWC Standards and Procedures Handbooks.

Revised Standards and Procedures Handbooks

The Version 4.1 Handbooks can be found here: https://ies.ed.gov/ncee/wwc/handbooks



Source: What Works Clearinghouse Procedures Handbook, Version 4.1, Figure I.1 on page 2
https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

SUBSTANTIVE CHANGE 1: ESTIMATING EFFECT SIZES WHEN MULTIPLE EFFECTS ARE PRESENT IN A SINGLE DOMAIN

Multiple Effect Sizes in a Domain

- An outcome domain is a set of conceptually and empirically related outcome measures. For example:
  - Alphabetics domain: phonemic awareness or letter identification
  - Academic achievement domain: grade-point average, standardized test scores
- The WWC averages effects within a single domain.
- Previous versions of the WWC handbooks used a standard error that assumed that the measures were perfectly correlated.

Implications of Assuming Perfect Correlation

- Assuming perfect correlation results in standard errors that are too large.
- The WWC investigated different approaches to understand the tradeoffs between Type I error rates and statistical power.
- Conclusion: The WWC will attempt to extract correlations between measures and use them in estimating domain average effect sizes.

How Much Does It Matter?

- The importance depends on:
  1. The number of effect sizes, and
  2. The actual correlation.

- Assume that typical statistical power in WWC studies is .50 and that the WWC desires a Type I error rate of .05. If we assume a correlation of 1.0 for calculation purposes:

Number of Effect Sizes	Empirical Correlation	Approximate Actual Type I Error Rate	Approximate Implied Statistical Power
2	.80	.039	.46
2	.60	.028	.40
5	.80	.032	.43
5	.60	.017	.34

How It Works

Example: Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). *Exceptional Children, 74*(2), 155–173. https://eric.ed.gov/?id=EJ817525

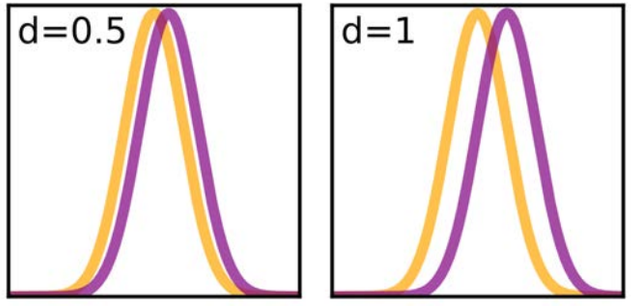
Outcome	Effect Size	Variance (*p*-value)	Combined Effect Size Variance, r = 1.0 (*p*-value)	Combined Effect Size Variance, r = .75 (*p*-value)
Subtraction fact fluency	.47	.117 (.18)	.118 (.19)	.103 (.17)
Double digit addition	.44	.118 (.21)		

- When the measures are not perfectly correlated, we can minimize the variance of the combined effect.
- Statistical power is improved when combining effects in a fixed-effects meta-analysis (discussed next).

SUBSTANTIVE CHANGE 2: USING FIXED-EFFECTS META-ANALYSIS FOR SYNTHESIZING EFFECTS ACROSS STUDIES

Changes to Synthesis Procedures

- Approach from Version 4.0 and earlier: vote counting
  - An intervention could receive a rating of "positive effects" if
    - At least two studies showed statistically significant positive effects, and
    - No studies showed statistically significant or substantively important negative effects, and
    - If at least one of the studies met WWC standards without reservations.
- Approach for Version 4.1: fixed-effects meta-analysis



Fixed-Effects Meta-Analysis

- Fixed-effects meta-analysis involves computing a weighted average effect size.
- Formally, studies are weighted by the inverse of the variance of their effect sizes.
- The weights are a function of sample size.
  - Larger studies get proportionally more weight in the analysis.

Example: Summer Counseling Interventions

- Summer counseling interventions are designed to prevent “summer melt.”
- WWC reviewed five studies of summer counseling. All studies met WWC group design standards without reservations.
- The intervention was rated as having “mixed effects” on the outcome domain of college access and enrollment:
  - Only one study was statistically significant, and
  - There were more studies with indeterminate results than with statistically significant and positive results.
- Intervention report is available at https://ies.ed.gov/ncee/wwc/InterventionReport/693.

Example: Summer Counseling Interventions

- Computing average effect size under the vote counting approach.

Study	Sample Size	Effect Size (*g*)	Statistically Significant?
Castleman et al. (2012)[1]	162	0.32	No
Castleman et al. (2014)	2,373	0.15	Yes
Castleman et al. (2015a)	1,602	-0.01	No
Castleman et al. (2015b)	3,281	0.09	No
Castleman & Page (2015)	5,059	0.06	No
Average		**+0.12**	**NA**

[1]*Example was simplified by using one effect size of 0.32 reported by Castleman et al. (2012). For all effect sizes, see pp. 21-22 of the intervention report*.

Example: Summer Counseling Interventions

- Computing weighted average effect size using fixed-effects meta-analysis

	Intervention Sample Size	Comparison Sample Size	Variance	Weight	Effect Size (*d*)	Weight * Effect Size
Castleman et al. (2012)	80	82	0.0250	39.98	0.32	12.79
Castleman et al. (2014)	886	1487	0.0018	553.74	0.15	83.06
Castleman et al. (2015a)	1074	528	0.0028	353.97	-0.01	-3.54
Castleman et al. (2015b)[1]	1761	1520	0.0678	14.75	0.09	1.33
Castleman & Page (2015)	2524	2535	0.0008	1264.18	0.06	75.85
Sums				**2226.62**		**169.49**
Weighted average (Weight * Effect Size)/Weight					**+0.08** ***p* = .0003** (169.49/2226.62 = 0.08)	

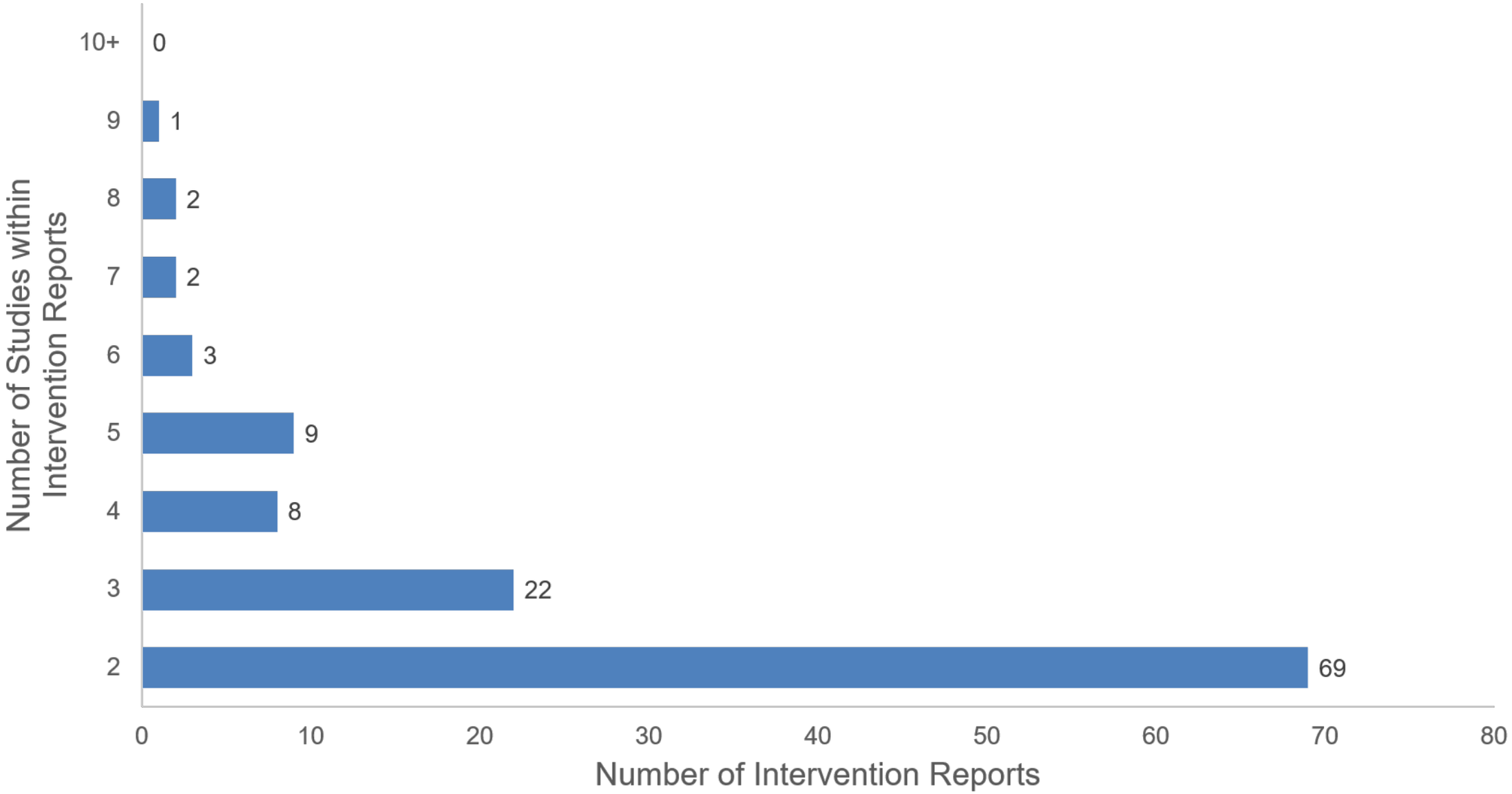
[1]This study needed a cluster correction as described in Procedures Handbook Version 4.1. The study's weight is based on an effective sample size of 59.

Example: Summer Counseling Interventions

- The intervention would be rated as having “positive effects" using the new synthesis procedure (discussed next)*:*
  - The mean effect size from a fixed-effects meta-analysis is positive and statistically significant, and
  - More than 50% of the fixed-effects meta-analytic weight comes from studies that *Meet WWC Standards Without Reservations.*

Why Not Random-Effects Meta-Analysis?

- Random-effects meta-analysis is less applicable for the WWC:
  - The random-effects model requires 10–15 studies per synthesis to obtain precise estimates; the WWC often has 3 or fewer studies per synthesis.
  - If the estimate is poor, it could lead the WWC to make too many Type I errors or too many Type II errors.



Note: This chart reflects the number of studies included within Intervention Reports that synthesized at least two studies that Meet WWC Standards *(n* = 116).

SUBSTANTIVE CHANGE 3: A NEW APPROACH TO INTERVENTION REPORT EFFECTIVENESS RATINGS

Intervention Report Effectiveness Rating

- The WWC replaced criteria for determining evidence of effectiveness from multiple studies from vote-counting to a fixed meta-analytic estimate.
- It was necessary to align the intervention reports' effectiveness rating with the new approach.
- The intervention effectiveness ratings are now used as factors for assigning levels of evidence for practice guide recommendations.

Intervention Report Effectiveness Rating

Version 4.0

Table IV.3. Criteria Used to Determine the WWC Rating of Effectiveness for an Intervention

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.	• Two or more studies show statistically significant positive effects, at least one of which meets WWC group design standards without reservations, AND • No studies show statistically significant or substantively important negative effects.

Version 4.1

Table IV.3. What Works Clearinghouse characterization of findings in intervention reports

Characterization	Criteria
Positive effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and positive; AND • More than 50.0 percent of the fixed-effects meta-analytic weight comes from studies that are rated *Meets WWC Standards Without Reservations.*

Intervention Report Effectiveness Rating

Version 4.0

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.	• At least one study shows statistically significant or substantively important positive effects, AND • Fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects, AND • No studies show statistically significant or substantively important negative effects.

Version 4.1

Potentially positive effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meet WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and positive; AND • 50.0 percent or less of the fixed-effects meta-analytic weight comes from studies that are rated *Meets WWC Standards Without Reservations*. OR • One study is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations;* AND • The study has a statistically significant and positive effect.

Synthesizing Evidence From Multiple Studies

Version 4.0

No discernible effects: No affirmative evidence of effects.	• None of the studies show statistically significant or substantively important effects, either positive or negative.
Mixed effects: Evidence of inconsistent effects.	EITHER both of the following: • At least one study shows statistically significant or substantively important positive effects, AND • At least one study shows statistically significant or substantively important negative effects, BUT no more such studies than the number showing statistically significant or substantively important positive effects. OR both of the following: • At least one study shows statistically significant or substantively important effects, AND • More studies show an indeterminate effect than show statistically significant or substantively important effects.

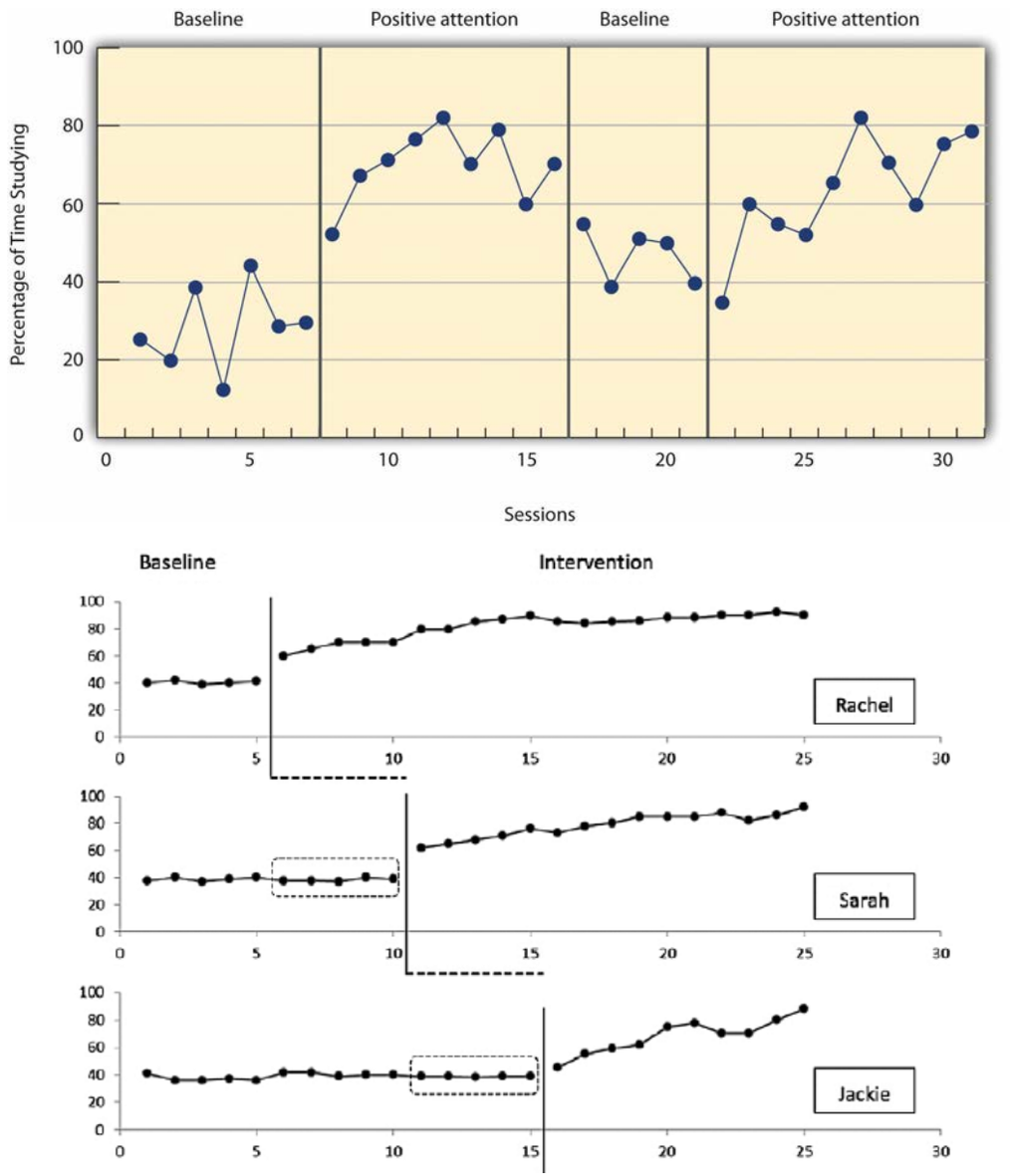
Version 4.1

Uncertain effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is not statistically significant. OR • One study is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The study does not have a statistically significant effect.

SUBSTANTIVE CHANGE 4: CALCULATING EFFECT SIZES IN SINGLE-CASE DESIGNS

Effect Sizes for SCDs

SCD researchers use experiments where an outcome measure is assessed repeatedly within and across different phases that are defined by the presence or absence of an intervention.



In SCD research, a case, such as a student or classroom, is the unit of intervention administration and data analysis.

Effect Sizes for SCDs: Prior Practice

- Extracting an Effect: Reviewers conducted visual analysis of graphical findings, assessing for demonstration of an intervention effect.

- Synthesizing Effects: Effects were only synthesized if the studies within an intervention or domain met the 5-3-20 rule.
  - 5-3-20 rule. SCD studies contributed to the WWC effectiveness rating only if the studies met three criteria: at least 5 studies met WWC SCD standards with or without reservations, the studies had to be conducted by at least 3 different research teams with no overlapping authorship at 3 different institutions, and the combined number of cases (i.e., participants, classrooms) had to total at least 20.
  - Often, the set of SCDs used to study an intervention did not meet all three criteria and did not contribute to the effectiveness rating.
  - When the 5-3-20 rule was satisfied for a given intervention-domain combination, review teams tallied and computed the proportion of SCD intervention effects that were positive or negative. SCD effects were not synthesized with effects from group design studies.

Effect Sizes for SCDs: Prior Practice

- Characterizing Intervention Effectiveness: Intervention effectiveness ratings were based on the proportion of SCD effects that clearly demonstrated a positive or negative effect.

- These practices do not align with the new WWC synthesis approach for group designs, which no longer relies on vote counting but on a fixed-effects meta-analysis using the effect size estimate from each study.

Effect Sizes for SCDs: A New Approach

- Methodological advances in the last decade provide an effect size (and standard error) option for SCDs that is designed to facilitate combination of SCD effects with effects from group designs (e.g., Hedges, Pustejovsky, & Shadish, 2012; Pustejovsky, Hedges, & Shadish, 2014).

- This effect size is referred to as the *design-comparable effect size* (D-CES).

- Study requirements for the D-CES under current (4.1) WWC Procedures:
  - Design: treatment reversal (AB^k), multiple baseline, multiple probe
  - Sample size: 3 or more cases.

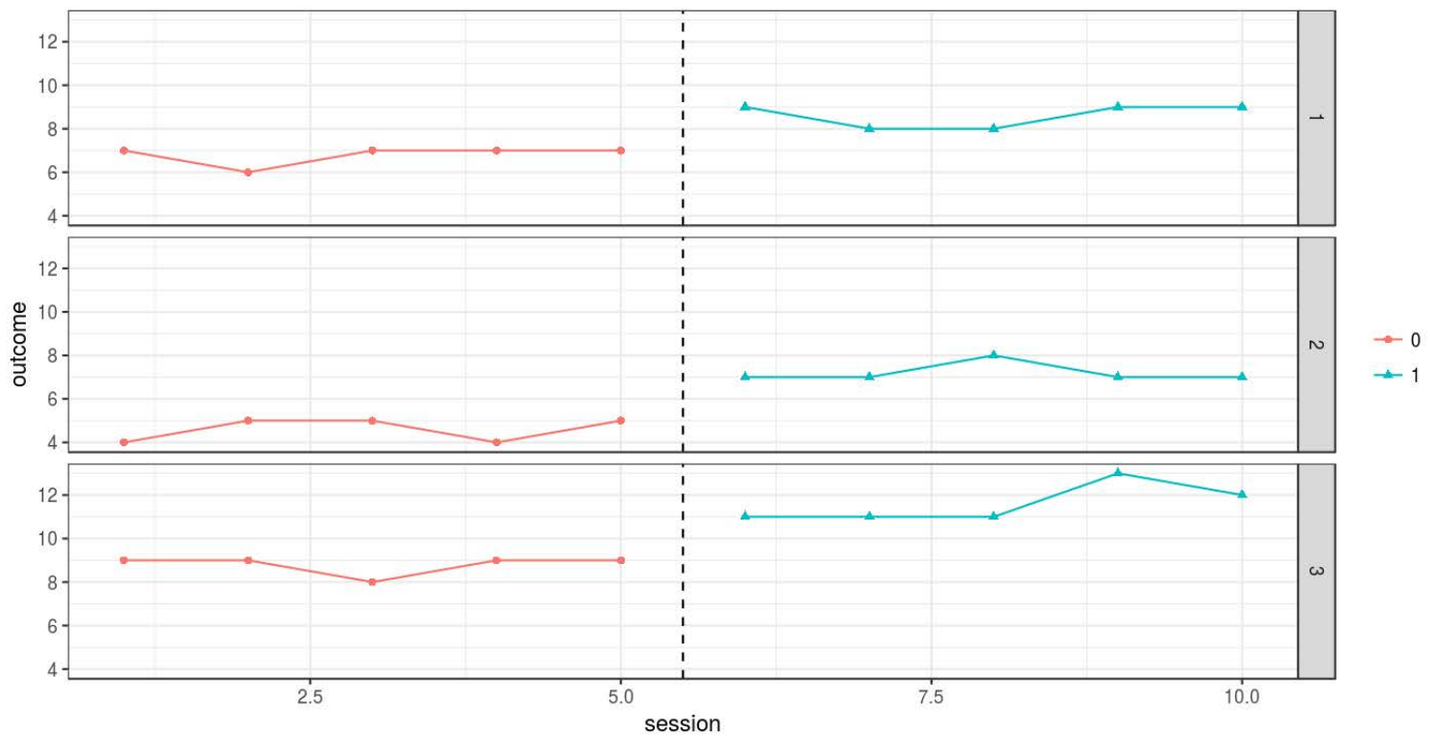
- The 5-3-20 rule has been eliminated.

Effect Sizes for SCDs: A New Approach

- In order to be rated *Meets WWC SCD Standards*, study authors are required to submit raw data in graphical or tabular form.

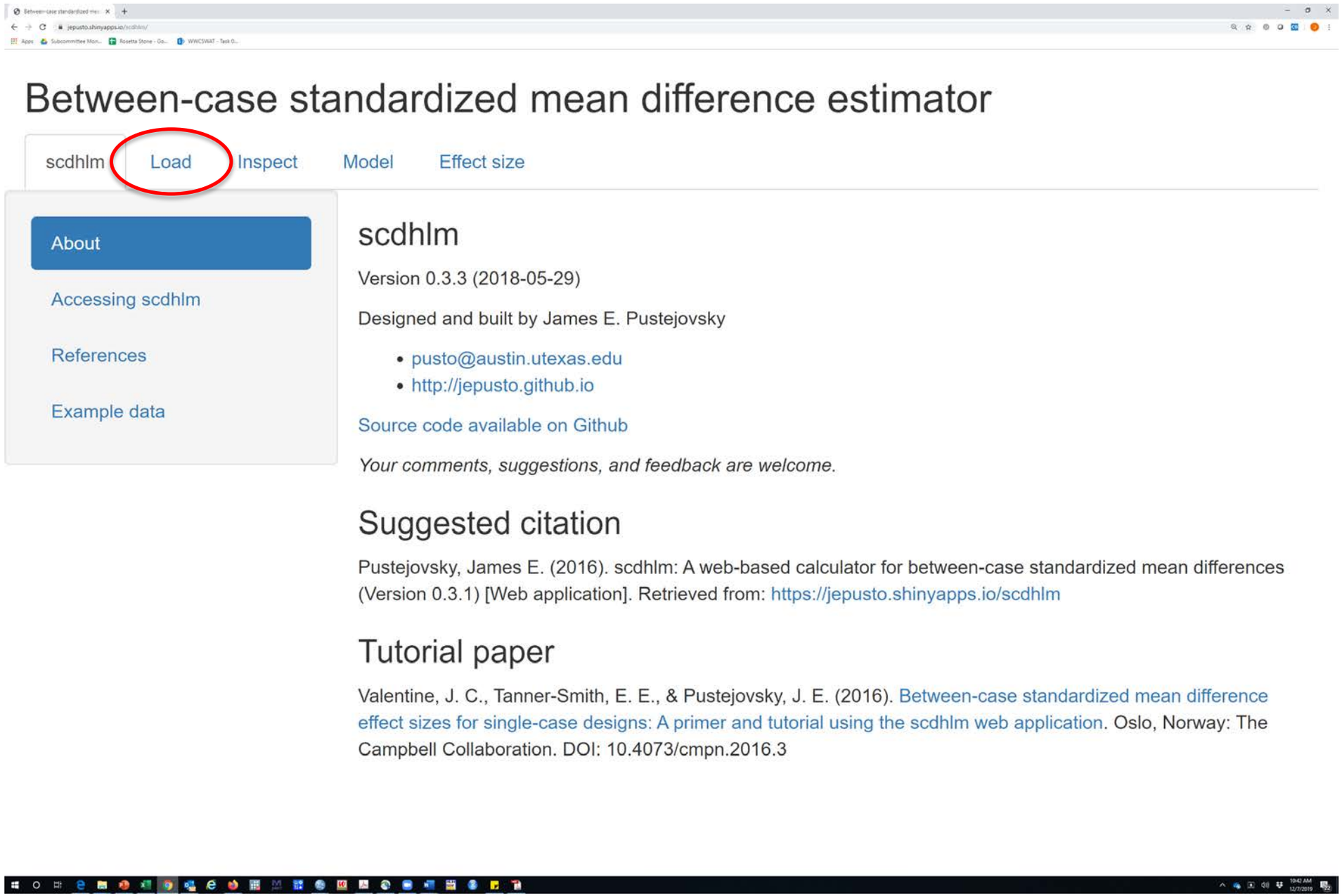
- Minimum required numeric variables in long format:
  - Case ID
  - Measurement session (i.e., occasion)
  - Phase indicator (e.g., baseline or intervention)
  - Outcome

Graphical or Tabular Data Format



Case	Occasion	Phase	Outcome
1	1	0	7
1	2	0	6
1	3	0	7
1	4	0	7
1	5	0	7
1	6	1	9
1	7	1	8
1	8	1	8
1	9	1	9
1	10	1	9
2	1	0	4
2	2	0	5
2	3	0	5
2	4	0	4
2	5	0	5
2	6	1	7
2	7	1	7
2	8	1	8
2	9	1	7
2	10	1	7
3	1	0	9
3	2	0	9
3	3	0	8
3	4	0	9
3	5	0	9
3	6	1	11
3	7	1	11
3	8	1	11
3	9	1	13
3	10	1	12

A Tool for Computing the D-CES



ADDITIONAL CHANGES IN THE WWC STANDARDS AND PROCEDURES, VERSION 4.1

Changes Applied to Procedures *and* Standards Handbooks

- The “substantively important” designation has been removed.
  - The handbooks no longer include the concept of “substantively important” for effects of 0.25 SD or larger that are not statistically significant.

- Standard error formulas are given for all included effect sizes.
  - All effect size formulas in the handbooks now include the formulas to compute the corresponding standard errors.
  - These standard errors will be used in the estimate of the fixed-effects meta-analysis.

Changes to Procedures

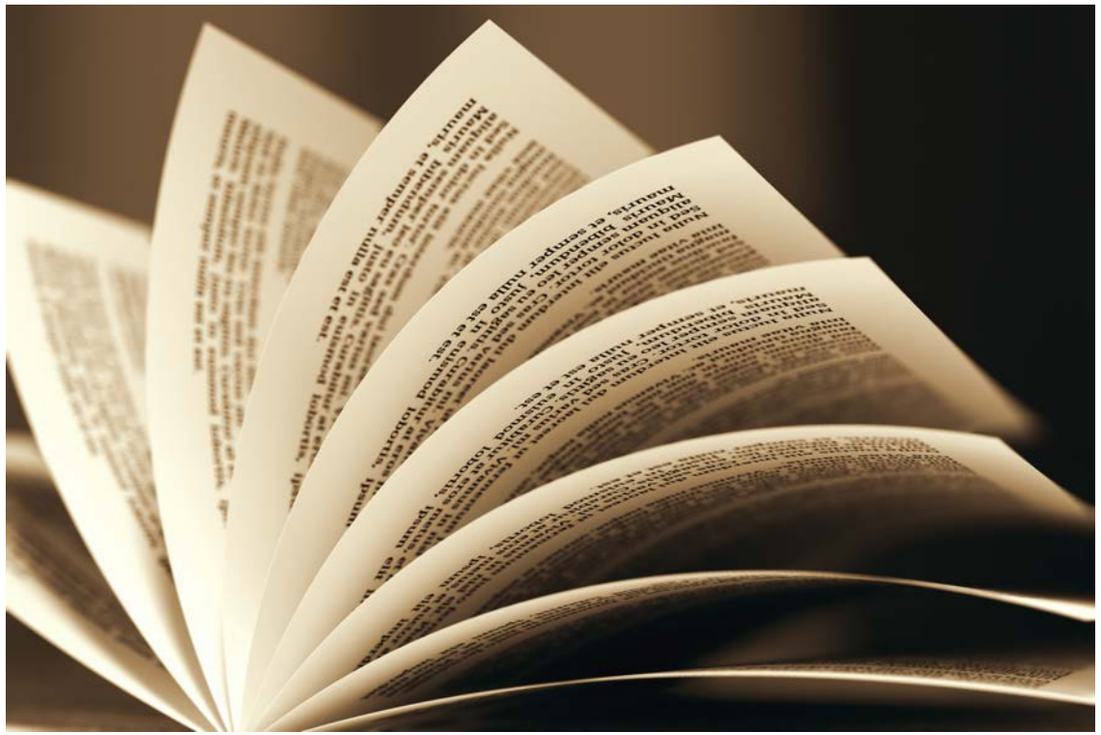
- New text distinguishes the Standards for Excellence in Education Research (SEER) Principles from the work of the WWC.
  - For additional information about SEER Principles, visit https://ies.ed.gov/seer.asp.
- Language has been added about the Every Student Succeeds Act Tiers of Evidence to align with updates to the WWC website.

Changes to Procedures (cont.)

- Procedures for prioritizing interventions for review are now more flexible: topic area protocols may identify criteria for selecting studies and determining eligibility.

- Education Resources Information Center (ERIC) should be used as the initial source of studies for WWC reviews, with topic area protocols specifying additional databases.

Changes to Procedures (cont.)



- When multiple versions of a manuscript are available, the most recent version submitted to ERIC is now preferred, followed by the most recent published version if no version is available on ERIC.
- Master's theses are now eligible for WWC reviews.
- Abstracts are now required to be screened for eligibility by two trained WWC staff instead of one.

Changes to Procedures (cont.)

- New language clarifies that author queries are intended to obtain information necessary to arrive at a study rating. For studies that meet standards, missing study context information will also be queried.
- New language explains that studies may be re-reviewed if they were previously reviewed under version 2.0 of the WWC Standards or under a separate review protocol.

Changes to Procedures (cont.)

- New language explains how to calculate regression discontinuity effect sizes.
- More detailed language about calculating Hedges' g and the Cox index has been added.
- New text notes that the WWC will not compute a standard error for the difference-in-differences estimate with a dichotomous outcome.

Changes to Standards

- SCD standards are no longer designated as “pilot” standards.
- WWC now provides examples of confounds in SCDs that would cause a study to be rated *Does Not Meet WWC SCD Standards.*
- WWC now allows study authors greater flexibility in accounting for missing baseline data in uncompromised randomized controlled trials.
- Language stating that joiners pose no risk of bias when they are excluded from the analytic sample has been removed.

Summary

In this webinar:

- We described in detail four substantive updates to the WWC Procedures:
  - Estimating effect sizes when multiple effects are present in a domain
  - Using fixed-effects meta-analysis for synthesizing effects across studies
  - A new approach to intervention report effectiveness ratings
  - Calculating effect sizes in SCDs.
- We briefly described additional noteworthy changes in Version 4.1 of the WWC Standards Handbook and Procedures Handbook.

❖Questions?

Resources

- WWC Handbooks and Other Resources: https://ies.ed.gov/ncee/wwc/Handbooks
- Webinars Related to WWC Handbook Content: https://ies.ed.gov/ncee/wwc/Handbooks#webinars

Have questions? Contact us:

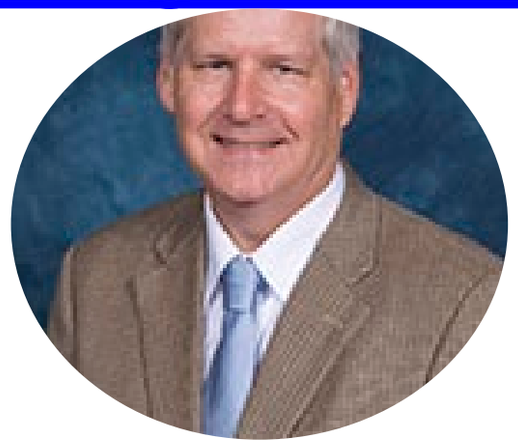
https://ies.ed.gov/ncee/wwc/help

Joseph Taylor, Ph.D.

American Institutes for Research

–

University of Colorado, Colorado Springs



Jeff Valentine, Ph.D.

University of Louisville

Sarah Caverly, Ph.D.

American Institutes for Research

References and Further Reading

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*, 224–239. Retrieved December 18, 2019, from https://doi.org/10.1002/jrsm.1052.

Pustejovsky, J. E. (2016). *scdhlm: A web-based calculator for between-case standardized mean differences* (Version 0.3.1) [Web application]. Retrieved December 18, 2019, from https://jepusto.shinyapps.io/scdhlm.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368–393. Retrieved December 18, 2019, from https://doi.org/10.3102/1076998614547577.

Valentine, J. C., Tanner-Smith, E. E., & Pustejovsky, J. E. (2016). *Between-case standardized mean difference effect sizes for SCDs: A primer and tutorial using the scdhlm web application.* Oslo, Norway: The Campbell Collaboration. Retrieved December 18, 2019, from https://doi.org/10.4073/cmpn.2016.3.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2018). *Transition to College intervention report: Summer counseling.* Retrieved December 18, 2019, from https://ies.ed.gov/ncee/wwc/InterventionReport/693.

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2020). *WWC Version 4.1 Standards and Procedures Handbooks.* Retrieved January 8, 2020, from https://ies.ed.gov/ncee/wwc/Handbooks.