Analytic Technical Assistance and Development

Reporting What Readers Need to Know about Education Research Measures: a Guide

Kimberly Boller
Mathematica Policy Research

Ellen Eliason Kisker
Twin Peaks Partners, LLC

REL 2014–064

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

September 2014

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Boller, K., & Kisker, E. E. (2014). *Reporting what readers need to know about education research measures: a guide* (REL 2014–064). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Contents

Tables

Essential Information for Reporting on Quantitative Measures

Education researchers face constraints and competing demands when they write study results for publication. On the one hand, they need to succinctly describe the study's motivation, methods, and results. On the other hand, good scientific practice requires that researchers report their methods with enough details for replication. When navigating these competing demands, researchers sometimes skimp on providing details about their measures. Yet, without sufficient information about the quantitative measures and their reliability and construct validity, readers cannot evaluate the credibility of the results, and other researchers cannot replicate the study's methods. Detailed reporting on measures benefits education science, supports deeper understanding of study methods, and ultimately supports replication, a critical part of the scientific method.

Tip #1

When creating a new measure or adapting an existing measure, build in additional time and resources to analyze and report on the measure's properties and performance in the study.

This guide is designed to help researchers make sure that their research reports include enough information about study measures so that readers can assess the quality of the study's methods and results. The guide also provides examples of write-ups about measures and suggests resources for learning more about these topics. The guide assumes that researchers have (1) clearly articulated their research questions, (2) completed a rigorous review of the leading measures for assessing each necessary component of the theory of change and the relevant domains and constructs, and (3) selected measures that are aligned with the intervention's theory of change (also referred to as a logic model) and that address the study's research questions. These measures may include contextual factors, inputs to implementation, expected intervention activities, outputs, and both short-term and long-term outcomes (Lugo-Gil et al. 2011; W.K. Kellogg 2004).

The following sections in this document present five checklists to guide authors in reporting on a study's quantitative measures, provide a sample write-up to illustrate the application of the checklists, and offer a list of resources to further inform reporting of quantitative measures and publication of studies.

Checklists to guide researchers in presenting key measurement details

Use the checklists in this guide to outline study reports and assess drafts for completeness. The five checklists in this guide highlight good practices identified by the authors (informed by their roles as peer reviewers of research, providers of evaluation training and technical assistance, and developers of measures compendia) and in some research standards for reporting on measures in study reports (Boller et al. 2010; Kisker et al. 2011; What Works Clearinghouse 2013a). Researchers can use the checklists to guide report planning—gathering the information needed as background for the measure description from technical manuals or research articles; planning necessary analyses; writing a description of the measures, their properties, and any limitations; and reviewing draft reports for completeness. Depending on the type of publication that researchers are preparing and the space constraints imposed by them, researchers may decide to present the measures information in the technical appendix of a report, the body of a journal article, or supplementary material that serves as an online appendix to a journal article. For longer presentations, the information may be best organized into tables or exhibits. The checklists in this guide address the following measurement topics: (1) measure domains and descriptions; (2) data collection training and quality; (3) reference population, study sample, and measurement timing; (4) reliability and construct validity evidence, and (5) missing data and descriptive statistics. Below, we describe each checklist and present questions that researchers can ask themselves as they write their measures for research reports.

Checklist 1: Quantitative measure domains and descriptions

A good methods section describes clearly the measurement domains (for example, reading, classroom management, and teacher pedagogical knowledge) included in the study, the types of measures used (for example, tests, scales, teacher surveys), and the analytic variables created from the measures. Checklist 1 addresses:

- **Measure domains and alignment with the intervention**. Does the report describe the conceptual domains studied, individual constructs measured, and outcome and control variables assessed and used in the analysis? For intervention studies, are any of the outcome measures overly aligned with the intervention and potentially unusable in an impact analysis?[1]
- **Measure descriptions**. Does the report provide the full name of and a citation for each measure and does it describe the features of each measure, including the assessment type, data sources, and scores that can be constructed from it?
- **Inclusion of instruments in an appendix or supplementary materials**. Are copies of each measure appended to the report? If a measure is protected by copyright, is there sufficient information in the measure description for a reader to identify which items and scales the study used?

Checklist 1. Quantitative measure domains and descriptions

Item	Included in report?
Measure domains and alignment with the intervention	
Description of the relevant domains, constructs, outcomes, and control variables that were assessed	□ Yes
Description of whether any measures used during the intervention were also used to assess outcomes, and if so, which ones	□ Yes
Measure descriptions	
Full name, source (author or developer if an existing measure, development process if a new measure), and citation for existing measure	□ Yes
Assessment type (individual assessment or group self-administered assessment or test; parent, teacher, or student report or observation)	□ Yes
Number of items	□ Yes
Rating scale (scale values and their meaning)	□ Yes
Scores and subscale scores defined for the measure	□ Yes
Score type (raw score, standardized score, factor score, Item Response Theory score, other)	□ Yes
Instruments	
Copies of instruments are included in report appendixes	□ Yes

Checklist 2: Data collection training, certification, and quality monitoring

Readers need to know how the study ensured high-quality data collection. The report should provide information about any strategies used to prepare data collectors to administer measures accurately and consistently. It should also describe strategies for monitoring their work and the quality of data collected. For example, interviewers for a telephone survey may have received group training on the items and probes, followed by ongoing supervisor monitoring during survey administration. Data collectors employed to conduct direct student assessments or classroom observations may have had to meet certain requirements for education or experience, may have received training and been certified, and may have been monitored to ensure quality. Checklist 2 addresses:

- **Staff training requirements and how the study team met them**. What do the authors of the measure and others who have used it recommend for the length and intensity of training? How did the study training meet or exceed those recommendations? How long was the training, and what did it cover? Did the training address collection of data in languages other than English (if relevant)?

- **Characteristics of trainers and trainees**. What qualifications and experience did the trainers and data collectors have for fulfilling their responsibilities? How do their characteristics (such as language spoken) compare to the study population (if relevant)?
- **Certification criteria.** What standards did data collectors have to meet to be certified to collect data with study respondents? Were data collectors who administered measures in a language other than English certified to administer the measures in that language?
- **Post-training and in-field reliability testing**. For direct student assessments and classroom observations, what evidence demonstrates inter-rater reliability at the start of data collection and during the field period? How was reliability established and maintained?

Checklist 2. Data collection training, certification, and quality monitoring

Item	Included in report?
Author recommendations for adequate staff training to administer the measure and how the recommendations were met	□ Yes
Characteristics of trainers and trainees (number, experience, gender, ability to administer measure in a study-relevant language)	□ Yes
Certification criteria	□ Yes
Post-training reliability testing procedures, thresholds, and results for each data collection period	□ Yes
In-field reliability testing procedures, thresholds, and results for each data collection period	□ Yes

Checklist 3: Reference population, study sample, and measurement timing

This checklist asks researchers to provide evidence that the population for which the measure was developed and standardized is similar to the study sample. In addition, the checklist asks about the timing of data collection, which also facilitates comparisons with standardization samples (for example, if the standardization was done 10 or more years before the start of the study, the norms may be outdated and underestimate student outcomes). Measure developers often attempt to include a range of respondents in their pretesting, standardization, and norming samples to ensure that their measures are reliable and valid for use with different populations. Nationally representative norming samples increase confidence that the measure was developed with a population similar to the study population and that the norms are appropriate for the study (for example, with regard to the norming sample's race and ethnicity, socioeconomic status, and urbanicity). Readers need to know the norming sample data collection dates and the study's data collection periods so that they can assess the extent to which the norms are likely to reflect the current state of student achievement and instructional practices. Timing of the baseline and follow-up data collection relative to intervention start and end dates also informs readers about the potential for finding intervention impacts.

Checklist 3 addresses:

- **Reference population.** What are the characteristics of the population used to develop and test the measure? If the measure is normed, what are the characteristics of the norming sample, and what year did the authors conduct the norming study? How do the reference population characteristics compare to the current study sample? For example, are students of similar ages, racial and ethnic backgrounds, and parent education represented in the norming study?
- **Sample characteristics and accommodations.** What are the characteristics of the study sample? Did the research team make any accommodations for students or circumstances that were not consistent with the measure developer's standardization practices? Did the research team use appropriate norms for students that speak languages other than English?
- **Timing of measurement**. When did each wave of data collection begin (month and year)? How much time elapsed between assessments? When did the intervention begin and end?

Checklist 3. Reference population, study sample, and measurement timing

Item	Included in report?
Reference population	
Description of whether the measure is a standardized measure, and if so, the metric used	□ Yes
Description of whether the measure is normed, and if so, the characteristics of the norming population and year of norming sample data collection	□ Yes
Characteristics of the study sample compared to the norming sample	□ Yes
Sample characteristics and accommodations (to confirm alignment with standardization and norming sample)	
Age and grade of students or study respondents	□ Yes
Percentage with a disability	□ Yes
Percentage that received an accommodation for a disability during assessment administration	□ Yes
Percentage from a low-income family	□ Yes
Percentage who are English learners (percentage who speak each language)	□ Yes
Percentage assessed in English and other languages	□ Yes
Timing of measurement relative to intervention and school year	
Month and year of start and end of baseline data collection	□ Yes
Month and year of start and end of intervention studied	□ Yes
Month and year of start and end of each follow-up data collection	□ Yes

Checklist 4: Reliability and construct validity evidence

No methods section is complete without a presentation of reliability and construct validity evidence. *Reliability* refers to the consistency and stability of measures. *Construct validity* refers to the degree to which a measure accurately assesses what it is designed to measure for its intended purpose.[2] Researchers need to provide the appropriate types of reliability and construct validity evidence for each measure in their study and enable readers to apply commonly used heuristics for determining the adequacy of the evidence (Boller et al. 2011). This applies to all scales and assessments, including those with a track record of use in research and practice as well as new and adapted measures. (The measures compendia shown in Table 2 identify sources of information on measures and their psychometric properties and also provide definitions for the terms in Checklist 4.) Checklist 4 addresses:

- **Reliability evidence**. For scales or test scores, what is the internal consistency reliability? For test scores, what is the evidence for alternate forms reliability? For observational or direct study assessments, what is the evidence of inter-rater reliability (gathered after data collector training and again during the field period)? What is the test-retest reliability?
- **Construct validity evidence**. What is the evidence that the measure assesses the intended construct and not something else? What is the evidence that the measure is predictive of the same or a related construct collected at a later point in time?
- **Other evidence of measure quality**. What other evidence is there of the measure's reliability, construct validity, and psychometric quality? Does the measure have a track record of success in other similar studies? If it is a new or adapted measure, was pretesting used to assess its psychometric properties? If not, has the team engaged an expert to help assess the measure's properties?

Tip #2

If available, present reliability and validity evidence from the measure's authors as well as evidence based on current study findings. Establish what threshold will determine whether a measure has adequate reliability and validity.

Many researchers use 0.70 as the threshold for internal consistency reliability, a measure of whether items in a scale are correlated and assess the same construct.

Checklist 4. Reliability and construct validity evidence

Item	Included in report?
Reliability evidence	
Internal consistency reliability	☐ Yes
Alternate forms reliability	☐ Yes
Inter-rater reliability	☐ Yes
Test-retest reliability	☐ Yes
Construct validity evidence	
Content validity	☐ Yes
Substantive validity	☐ Yes
Structural validity	☐ Yes
Generalizability validity	☐ Yes
External validity (convergent, discriminant, predictive)	☐ Yes
Consequential validity	☐ Yes
Other evidence	
Any other evidence of measure quality	☐ Yes

Checklist 5: Missing data, descriptive statistics, and the key strengths and limitations of constructed variables

To set the stage for reporting on study findings, authors should provide information about the variables constructed from the raw data, including specifications of how variables were constructed, levels of missing data and how they were handled, descriptive statistics, and the results of sensitivity analyses for assessing the robustness of study findings to alternative ways of defining measures or handling missing data. Checklist 5 addresses:

- **Extent of and approaches used to account for missing data**. For how many respondents is each measure missing? How systematic is the missingness (missing completely at random, missing at random, or not missing at random; Puma et al. 2009)? For measures that are multi-item scales, how many items are missing for each respondent? What rules did the research team use to handle missing data and either impute or remove measures or individuals from the analyses?
- **Descriptive statistics**. What are the group means, variance, and possible and actual ranges of the measures? Did the team make any transformations to the scores on the basis of their distribution (for example, if the score on a measure did not have a normal distribution)?
- **Key strengths and limitations of constructed variables**. Is there evidence that the study's approach to collecting data using the measure and analyzing the variables constructed from it was successful? What are the main problems with the constructed variables based on the measure that may affect the results and conclusions from the study? What could be done to mitigate those issues in the future?

Tip #3

Space allocated for describing measures and their properties may be limited by the publisher or sponsor. Whatever the constraints, researchers should conduct the types of analyses described here and document how well the measures work.

- **Many journals allow authors to submit supplemental material for readers to access online.**
- **Government and foundation reports may include a technical appendix detailing the measures and their properties.**

Checklist 5. Missing data, descriptive statistics, and key strengths and limitations of constructed variables

Item	Included in report?
Extent and handling of missing data	
Extent of within-scale missingness and how it was handled	☐ Yes
Extent of case-level missingness and how it was handled	☐ Yes
Sensitivity test findings to demonstrate measure robustness to alternate specifications and approaches to handling missing data	☐ Yes
Descriptive statistics for baseline, pre-intervention analysis sample, and post-intervention analysis sample measures (overall and by treatment group)	
Analytic variable specifications (especially for tests and scales)	☐ Yes
Means and standard deviations	☐ Yes
Sample sizes	☐ Yes
Minimum and maximum possible values	☐ Yes
Minimum and maximum actual values observed and percentage of sample members with minimum and maximum values	☐ Yes
Key strengths and limitations of measures and their data sources	
Strengths and limitations of each measure	☐ Yes
Strengths and limitations of each data source	☐ Yes

Example of reporting on measures

This section presents an example of parts of a methods section for a *fictitious* study that uses a fictitious classroom quality measure. It includes a description of the measure [Checklist 1]; details about how observers were trained to collect data, including tests of inter-rater reliability [Checklist 2]; additional reliability and construct validity evidence collected by the research team [Checklist 4]; and information about missing data and descriptive statistics summarizing the data that were collected using the measure [Checklist 5]. The measure is not normed, but the example includes a description of the similarity of the schools in the study to the schools in published studies by the measure developers [Checklist 3]. Not all items in the checklists are applicable in this example; footnotes are added to identify how elements of the checklists are addressed in the example.

Tip #4

Use the checklists in this document plus a good example of a methods and results section as guides to select and use a measure and to document its use.

Measure description

We conducted the September 2012 pre-intervention and April 2013 post-intervention observations of 40 grade 1 through grade 3 classrooms by using the Education Quality-Inquiry (EQUAL-I; Jones et al. 2005).[3] The EQUAL-I is a 2-hour observation that focuses on the quality of teacher-child interaction during math and science instruction in grade 1 through grade 3 classrooms.[4] It measures process quality through 15 items in three areas: (1) Teacher Facilitation of Inquiry, (2) Student Engagement in Inquiry-Based Learning, and (3) Use of Technology for Inquiry-Based Learning. Each area consists of five rating scales defined by observable indicators along a 5-point scale, with ratings reflecting scores in the low (1–2), moderate (3), and high (4–5) ranges of quality. For the EQUAL-I, observers look for evidence of specific indicators as they rate each scale.[5] The three resulting subscale scores and the overall total score are the simple mean of each rating scale and the total.[6] Few observations had missing values (some data were missing for 6 out of the 120 total classrooms); at most, one rating scale was missing per subscale and for the total score. We did not observe any systematic patterns of missingness. Therefore, we imputed the mean of the nonmissing items.[7] EQUAL-I has no norming or nationally representative comparison sample, but the published studies by the measure authors and others were conducted with schools and classrooms that serve children with similar characteristics (schools are of a similar size [290–420]), in a mix of urban and semi-urban settings, with a similar proportion of children in poverty (28 to 35 percent eligible for free and reduced price lunch) (Jones and Miller 2008).[8]

Training and inter-rater reliability results

EQUAL-I has been used in education research for the past five years (more than 10 studies have used it as an intermediate outcome measure), and its properties have been documented in articles published by its authors. Training includes certification by the measure authors directly or through a training of trainers model (Jones et al. 2005). The training is classroom- and video-based with inter-rater reliability tests conducted following two days of in-class training and practice. The post-training certification test includes rating of six classroom videotapes of 30 minutes each. Each test video was consensus coded by the authors and serves as the gold standard against which trainees are assessed.[9] To be certified, observers must achieve exact agreement with gold standard codes on 80 percent of the ratings on five of the six videotapes.[10,11]

Baseline training and certification. At baseline, we engaged the measure authors to train 6 observers (4 female and 2 male; all new to conducting the EQUAL-I), and 5 passed the post-training certification requirements. After additional classroom training and reassessment one week later using a second set of six test videotapes, the sixth observer was certified.[12] On average, exact agreement of the observers with the gold standard at the end of training on all six test videotapes was 85 percent. Two weeks after the start of data collection, observers took a second reliability test with another set of six test tapes, and all met or exceeded the certification requirements.[13]

In all, three videotaped inter-rater reliability tests (with six tapes for each test) were administered during the 8-week baseline data collection period. The average across the six observers across all of the rating items was 82 percent exact agreement with the gold standard. We computed inter-rater reliability for the subscale and total scores and found that all observers met the standard of at least 80 percent agreement with the gold standard across all of the tests during the data collection period.[14]

Post-intervention follow-up training and certification. At the post-intervention follow-up, we retrained five of the six original observers for four hours and tested them on a new set of six videotapes. Average post-retraining exact agreement with the gold standard on all six test tapes was 88 percent.[15] In-field tests were conducted twice during the 6-week follow-up data collection period, and average exact agreement with the gold standard tapes was 81 percent for the first in-field test and 87 percent for the second.[16]

Additional reliability and construct validity evidence

Internal consistency reliability. We computed scale scores as specified by the measure authors and found acceptable internal consistency reliability (Cronbach's standardized coefficient alpha of 0.85, 0.82, 0.79, and 0.93 for the EQUAL-I Total Score, Teacher Facilitation of Inquiry Subscale Score, Student Engagement Subscale Score, and the Use of Technology Subscale Score, respectively) that was similar to reports in the manual (Jones et al. 2005; Table 1). We also computed and analyzed alpha separately for classroom observations conducted in schools serving higher proportions (22 of 40) versus lower proportions (18 of 40) of children receiving free and reduced-price lunch and found similar patterns across these groups, which provides evidence for the reliability of the measures across school populations and contexts.[17] The sample size, unweighted means, standard deviations, coefficient alphas, and ranges for the subscale scores and the total for each data collection period are presented in Table 1 and the appendix.[18]

Author-reported test-retest reliability. The EQUAL-I authors reported adequate test-retest reliability as part of the studies completed on the measure (average correlations of .82 across two days for 25 classrooms; Jones et al. 2005), which provided sufficient evidence to support our approach of conducting a single observation per classroom. There are no alternate forms of this measure.[19]

Author-reported and study-observed construct validity evidence. Jones and colleagues (Jones et al. 2005 and Jones et al. 2010) reported that a panel of 10 early elementary school teaching experts participated in reviewing the EQUAL-I scales and recommended changes in how rating scales were structured and in the way that key behaviors were defined. Those changes were made before pretesting and finalizing the measure. The experts agreed that there was sufficient content validity for a measure designed to assess support in the classroom for inquiry-based teaching and learning in early elementary school. Our application of this measure to lessons focused on mathematics and science is consistent with the overall goal of the measure and its content.[20] The authors reported convergent validity between the EQUAL-I Total Score in the fall and student mathematics achievement measured in spring of the same school year (correlations between quality and student outcomes for 25 classrooms and 75 students in grade 3 of .35, p<.05).[21]

We also assessed convergent validity by correlating the quality ratings with another measure of classroom quality: minutes of instructional time.[22] All of the EQUAL-I subscale scores and the total scale score were positively correlated (p<.05) with instructional time at baseline (0.32, 0.44, 0.51, and 0.49 for the EQUAL-I Total Score, Teacher Facilitation of Inquiry Subscale Score, Student Engagement Subscale Score, and the Use of Technology Subscale Score, respectively).[23]

Tip #5

In reports and journal articles, consider putting measurement details—such as reliability and validity information and the timing of baseline and follow-up assessments—into a table or exhibit.

Table 1. Example of a descriptive statistics table: Education Quality-Inquiry total and subscale ranges, means, and alphas (treatment and control classrooms combined)

	Possible range		Reported range (% with that value)				
Measure	Min.	Max.	Min.	Max.	Mean/ percentage	Standard deviation	Cronbach's alpha
EQUAL-I total score	1	5	2.00 (8)	5.00 (14)	3.60	0.96	0.85
Teacher facilitation of inquiry	1	5	2.00 (15)	5.00 (1)	3.24	0.87	0.82
Student engagement in inquiry-based learning	1	5	2.00 (12)	4.00 (22)	3.8	0.98	0.79
Use of technology for inquiry-based learning	1	5	2.00 (11)	5.00 (5)	3.56	0.95	0.93
Sample size	40						

Note: The total score includes 15 items, and each of the three subscales includes 5 items.

Source: Spring 2013 Year 1 Post-Intervention Observations.

Additional resources to inform measures reporting

In addition to using this guide to inform measures reporting, researchers can consult two other types of sources: (1) existing guidelines for reporting on measures and (2) measures compendia. First, resources that contain guidelines and recommendations for measure reporting also describe good practices to use in the description of measures. For example, these resources include a description of measures information that is required for reviews conducted by the What Works Clearinghouse. Second, measures compendia often provide summaries of the author-reported and recent research findings about a measure's reliability and construct validity as well as citations for where to find more information about the measure. Researchers can draw on these resources as they plan and write their reports. (Table 2 lists examples of these resources and gives a short description of each.)

Table 2. Measures reporting resources

Resource	Description
Measure reporting guidelines and recommendations	
Higgins J.P.T., Green S. (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Retrieved on May 27, 2014, from http://www.cochrane-handbook.org	This handbook defines outcomes, summarizes approaches to handling missing data, and lists the information needed in reports on outcomes and on sources of bias that may affect study findings and subsequent systematic reviews.
U.S. National Institutes of Health. *ClincalTrials.gov Protocol Data Element Definitions (Draft).* Retrieved on September 26, 2013, from http://prsinfo.clinicaltrials.gov/definitions.html	ClinicalTrials.gov defines primary and secondary outcome measures that researchers must report on to register a study into its international database of clinical studies. The data element definitions document also specifies requirements for reporting on missingness and timing of follow-up data collection.
What Works Clearinghouse. *Evidence Review Protocols*. Retrieved on August 2, 2013, from http://ies.ed.gov/ncee/wwc/Publications_Reviews.aspx?f=Publication%20and%20Review%20Types,5;#pubsearch	The What Works Clearinghouse posts review protocols for each topical area. Protocols describe the outcomes included in the review and the standards for reliability and reporting on outcome measures in both group design studies (randomized controlled trials, regression discontinuity designs, and quasi-experimental designs) and single-case design research.
What Works Clearinghouse. *What Works Clearinghouse Reporting Guide for Study Authors*. Retrieved on July 25, 2013, from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=235	This What Works Clearinghouse Guide summarizes how the systematic reviews are done and what reviewers are trained to assess and report on, including outcome measure reporting for both the treatment and control groups.
What Works Clearinghouse. *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0).* Retrieved on May 27, 2014, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf	The What Works Clearinghouse Procedures and Standards Handbook describes the criteria used and the information needed to assess studies included in What Works Clearinghouse reviews.
Measures compendia	
Boller, K., Atkins-Burnett, S., Malone, L.M., Baxter, G.P., and West, J. *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions*. Volume I, Measures Selection Approaches and Compendium Development Methods. NCEE 2010-4012. Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, April 2010. Malone, M., Cabili, C., Henderson, J., Esposito, A.M., Coolahan, K., Henke, J., Asheer, S., O'Toole, M., and Boller, K. *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions*. Volume II, Technical Details, Measure Profiles, and Glossary (Appendices A-G). NCEE 2010-4012. Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, April 2010.	This compendium provides suggestions for how to assess the quality of measures, a glossary with definitions of measure-related terms, and profiles of more than 90 measures that researchers have used in NCEE evaluations. Volume I explains why it is important for researchers to select and use measures with strong psychometric properties, and Volume II presents the measure profiles.

Denham, S.A., Ji, P., and Hamre, B. *Compendium of Preschool Through Elementary School Social-Emotional Learning and Associated Assessment Measures*. Chicago: University of Illinois at Chicago, 2010.	This compendium profiles 65 measures of social-emotional learning and their psychometric properties. The measures assess aspects of the school context, social-emotional learning competencies, and academic-related social-emotional learning competencies.
Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., and Mooney, K. "Measuring Student Engagement in Upper Elementary Through High School: A Description of 21 Instruments." *Issues & Answers Report*, REL 2011–No. 098. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast, 2011.	This compendium defines the construct of student engagement and presents profiles of 21 measures. The authors provide basic information about psychometric properties and also cite references for how to learn more.
Halle, T., Vick Whittaker, J. E., and Anderson, R. (2010). *Quality in Early Childhood Care and Education Settings: A Compendium of Measures*, Second Edition. Washington, DC: Child Trends. Prepared by Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.	This compendium summarizes the psychometric properties of 50 measures of early care and education quality. The profiles provide information on how the measures are scored and references for learning more about the measures and their properties.
Wilson-Ahlstrom, A., Yohalem, N., DuBois, D., and Ji, P. "From Soft Skills to Hard Data: Measuring Youth Program Outcomes." Washington, DC: The Forum for Youth Investment, 2011.	This compendium presents eight measures in four outcome areas relevant to youth programs. The report presents basic information about the measures and their technical properties.

Appendix. Measures reporting checklist for researchers

Item	Included in report?
Checklist 1: Quantitative measure domains and descriptions	
Measure domains and alignment with intervention	
Description of the relevant domains, constructs, outcomes, and control variables that were assessed	□ Yes
Description of whether any measures used during the intervention were also used to assess outcomes, and if so, which ones	□ Yes
Measure descriptions	
Full name, source (author or developer if an existing measure, development process if a new measure), and citation for existing measure	□ Yes
Assessment type (individual assessment or group self-administered assessment or test, parent, teacher, or student report or observation)	□ Yes
Number of items	□ Yes
Rating scale (scale values and their meaning)	□ Yes
Scores and subscale scores defined for the measure	□ Yes
Score type (raw score, standardized score, factor score, Item Response Theory score, other)	□ Yes
Instruments	
Instruments are included in report appendixes	□ Yes
Checklist 2: Data collection training, certification, and quality monitoring	
Author requirements for adequate staff training to administer the measure and how they were met	□ Yes
Characteristics of trainers and trainees (number, experience, gender, ability to administer measure in a study-relevant language	□ Yes
Certification criteria	□ Yes
Post-training reliability testing procedures, thresholds, and results for each data collection period	□ Yes
In-field reliability testing procedures, thresholds, and results for each data collection period	□ Yes

Checklist 3: Reference population, study sample, and measurement timing	
Reference population	
Description of whether the measure is a standardized measure, and if so, the metric used	☐ Yes
Description of whether the measure is normed, and if so, the characteristics of the norming population and year of norming sample data collection	☐ Yes
Characteristics of the study sample compared to the norming sample	☐ Yes
Sample characteristics and accommodations (to confirm alignment with standardization and norming sample)	
Age and grade of students or study respondents	☐ Yes
Percentage with a disability	☐ Yes
Percentage that received an accommodation for a disability during assessment administration	☐ Yes
Percentage from a low-income family	☐ Yes
Percentage who are English learners (percentage who speak each language)	☐ Yes
Percentage assessed in English and other languages	☐ Yes
Timing of measurement relative to intervention and school year	
Month and year of start and end of baseline data collection	☐ Yes
Month and year of start and end of intervention studied	☐ Yes
Month and year of start and end of each follow-up data collection	☐ Yes
Checklist 4: Reliability and construct validity evidence	
Reliability evidence	
Internal consistency reliability	☐ Yes
Alternate forms reliability	☐ Yes
Inter-rater reliability	☐ Yes
Test-retest reliability	☐ Yes
Construct validity evidence	
Content validity	☐ Yes

Substantive validity	☐ Yes
Structural validity	☐ Yes
Generalizability validity	☐ Yes
External validity (convergent, discriminant, predictive)	☐ Yes
Consequential validity	☐ Yes
Other evidence	
Other information about measure quality	☐ Yes
Checklist 5: Missing data, and descriptive statistics, and key strengths and limitations of constructed variables	
Extent and handling of missing data	
Extent of within-scale missingness and how it was handled	☐ Yes
Extent of case-level missingness and how it was handled	☐ Yes
Sensitivity test findings to demonstrate measure robustness to alternate specifications and approaches to handling missing data	☐ Yes
Descriptive statistics for baseline, pre-intervention analysis sample, and post-intervention analysis sample measures (overall and by treatment group)	
Analytic variable specifications (especially for tests and scales)	☐ Yes
Means and standard deviations	☐ Yes
Sample sizes	☐ Yes
Minimum and maximum possible values	☐ Yes
Minimum and maximum actual values observed and percentage of sample members with minimum and maximum values	☐ Yes
Key strengths and limitations of measures and their data sources	
Strengths and limitations of the measures	☐ Yes
Strengths and limitations of the data sources	☐ Yes

Notes

1. An outcome measure is overly aligned with the intervention if it gives an unfair advantage to the intervention group. (What Works Clearinghouse 2013a). For example, measures used as part of the intervention or measures that include items derived from intervention materials seen by the intervention group but not the comparison group would be overly aligned.

2. The unified view of validity holds that there are six aspects of construct validity: content, substantive, structural, generalizability, external, and consequential. The content aspect includes evidence of content relevance and representativeness. The substantive aspect refers to substantive theories and process modeling to identify processes to be revealed in assessment tasks along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. The structural aspect appraises the fidelity of the scoring structure to the substantive theory of the construct domain. The generalizability aspect examines the extent to which assessment scores are generalizable across tasks and contexts. The external aspect refers to the extent to which the assessment scores' relationships with other measurements and nonassessment behaviors reflect the expected relations implicit in the theory of the construct being assessed. Finally, the consequential aspect appraises the intended and unintended consequences of score interpretation and use in both the short- and long-term. Messick (1995) discusses these aspects of construct validity and the types of evidence that inform each aspect.

3. Checklist 1: Full name and source of measure.

4. Checklist 1: Assessment type.

5. Checklist 1: Number of items, rating scale.

6. Checklist 1: Score type.

7. Checklist 5: Extent of missingness and how it was handled.

8. Checklist 3: Whether the measure is normed, characteristics of the study sample compared to the norming sample.

9. Checklist 2: Post-training reliability thresholds.

10. Checklist 2: Author requirements for adequate staff training.

11. Checklist 2: Certification criteria.

12. Checklist 2: Post-training reliability testing procedures and results.

13. Checklist 2: Characteristics of trainees.

14. Checklist 2: In-field reliability testing procedures, thresholds, and results.

15. Checklist 2: Post-training reliability testing procedures, thresholds, and results.

16. Checklist 2: In-field reliability testing procedures, thresholds, and results.

17. Checklist 4: Internal consistency reliability.

18. Checklist 4: Internal consistency reliability; Checklist 5: Means and standard deviations, sample sizes, minimum and maximum possible values.

19. Checklist 4: Test-retest reliability, alternate forms reliability.

20. Checklist 4: Construct validity.

21. Checklist 4: Construct validity.

22. Lack of resources constrained our ability to assess another aspect of construct validity, discriminant validity, that requires administration of measures of instructional quality that are not expected to be correlated with the EQUAL-I.

23. Checklist 4: Construct validity.

Bibliography

Andrews, G., Peters, L., and Teesson, M. (1994). *The Measurement of Consumer Outcomes in Mental Health*. Canberra, Australia: Australian Government Publishing Services.

Bacon, D. (2004). "The Contributions of Reliability and Pretests to Effective Assessment." *Practical Assessment, Research & Evaluation, 9*(3).

Berry, D. J., Bridges, L. J., and Zaslow, M. J. (2004). *Early Childhood Measures Profiles*. Washington, DC: Child Trends.

Boller, K., Atkins-Burnett, S., Malone, L. M., Baxter, G. P., and West, J. (2010). *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions. Volume I, Measures Selection Approaches and Compendium Development Methods*. NCEE 2010-4012. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED511790

Campbell, D. T. and Fiske, D. W. (1995). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin*, *56*(2):81–105.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*, *Revised Edition.* New York: Academic Press.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Philadelphia: Harcourt Brace Jovanovich College Publisher. http://eric.ed.gov/?id=ED312281

Denham, S.A., Ji, P., and Hamre, B. (2010). *Compendium of Preschool Through Elementary School Social-Emotional Learning and Associated Assessment Measures*. Chicago: University of Illinois at Chicago.

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B, and Mooney, K. (2011). *Measuring Student Engagement in Upper Elementary Through High School: A Description of 21 Instruments*. (Issues & Answers Report, REL 2011–No. 098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. http://eric.ed.gov/?id=ED514996

Halle, T., Vick Whittaker, J. E., and Anderson, R. (2010). *Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Heitjan, D. F., and Basu, S. (1996). "Distinguishing 'Missing at Random' and 'Missing Completely at Random.'" *The American Statistician*, *50*(3):207–213.

Higgins, J.P.T., and Green S. (editors) (2011). *Cochrane Handbook for Systematic Reviews of Interventions,* Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Institute of Education Sciences. (2013). *NCEE Guidance for REL Study Proposals, Reports, and Other Products*. Washington, DC: U.S. Department of Education.

Kisker, E., Boller, K., Cabili, C., Nagatoshi, C., Kamler, C., Johnson, C.J., et al. (2011). *Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

LeBel, E. P., and Paunonen, S. V. (2011). "Sexy but Often Unreliable: Impact of Unreliability on the Replicability of Experimental Findings Involving Implicit Measures." *Personality and Social Psychology Bulletin*, *37*:570–583.

LeBel, E. P., and Peters, K. R. (2011). "Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of psi as a Case Study of Deficiencies in Modal Research Practice." *Review of General Psychology*, *15*(4):371–379.

Litwin, M. S. (2003). *How to Assess and Interpret Survey Psychometrics,* 2nd Edition. Thousand Oaks, CA: Sage Publications.

Lugo-Gil, J., Sattar, S., Ross, C., Tout, K., Kirby, G., and Boller, K. (2011). "The Quality Rating and Improvement System (QRIS) Evaluation Toolkit." OPRE report no. 2011-31. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Malone, L. M., Cabili, C., Henderson, J., Mraz Esposito, A., Coolahan, K., Henke, J., et al. (2010). *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions. Volume II: Technical Details, Measure Profiles, and Glossary (Appendices A–G)*. NCEE 2010–4013. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED511792

Messick, S. (1995). "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist*, 50:741–749. http://eric.ed.gov/?id=EJ517194

Nunnally, J. C. (1978). *Psychometric Theory, 2nd Edition.* New York: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory, 3rd Edition.* New York: McGraw-Hill.

Person, A. E., Moiduddin, E., Hague-Angus, M., and Malone, L. M. (2009). *Survey of Outcomes Measurement in Research on Character Education Programs.* (NCEE 2009-006). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED511774

Puma, M. J., Olsen, R. B., Bell, S. H., and Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://eric.ed.gov/?id=ED511781

Rubin, D.B. (1976). "Inference and Missing Data." *Biometrika*, 63:581–592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schochet, P. Z. (2009). "An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations." *Evaluation Review, 33*(6):539–67. http://eric.ed.gov/?id=EJ862322

Spies, R. A., Carlson, J. F., and Geisinger, K. F. (eds.) (2010). *The Eighteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

U.S. National Institutes of Health. *ClincalTrials.gov Protocol Data Element Definitions (Draft).* Retrieved on September 26, 2013, from http://prsinfo.clinicaltrials.gov/definitions.html.

What Works Clearinghouse. (2013a). *Evidence Review Protocols*. Retrieved on August 2, 2013, from http://ies.ed.gov/ncee/wwc/Publications_Reviews.aspx?f=Publication%20and%20Review%20Types,5;#pubsearch.

What Works Clearinghouse. (2013b). *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0).* Retrieved on May 27, 2014, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf.

What Works Clearinghouse. (2013c). *What Works Clearinghouse Reporting Guide for Study Authors*. Retrieved on July 25, 2013, from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=235.

Wilson-Ahlstrom., A., Yohalem, N., DuBois, D. and Ji, P. (2011). *From Soft Skills to Hard Data: Measuring Youth Program Outcomes*. Washington, DC: The Forum for Youth Investment.

W. K. Kellogg Foundation. (2004). *Evaluation Handbook.* Battle Creek, MI: Author. Retrieved on July 28, 2013, from http://www.wkkf.org/knowledge-center/resources/2010/w-k-kellogg-foundation-evaluation-handbook.aspx.