NCEE 2010-4012

U. S. DEPARTMENT OF EDUCATION

Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions

Volume I: Measures Selection Approaches and Compendium Development Methods

April 7, 2010

Kimberly Boller
Sally Atkins-Burnett
Lizabeth M. Malone
Gail P. Baxter
Jerry West

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions

Volume I: Measures Selection Approaches and Compendium Development Methods

April 7, 2010

Kimberly Boller
Sally Atkins-Burnett
Lizabeth M. Malone
Gail P. Baxter
Jerry West

NCEE 2010-4012
U.S. DEPARTMENT OF EDUCATION

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research to document and describe outcome measures used in education evaluations. The views expressed in this report are those of the authors and do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

April 2010

Boller, Kim, Sally Atkins-Burnett, Lizabeth M. Malone, Gail P. Baxter, and Jerry West (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: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at (202) 260-9895 or (202) 205-8113.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

This report was prepared under a contract with IES. The five authors, Drs. Kimberly Boller, Sally Atkins-Burnett, Lizabeth Malone, Gail Baxter, and Jerry West are employees of Mathematica Policy Research (Mathematica). The authors and other staff of Mathematica have no financial interests that could be affected by the content of this report.

CONTENTS

Chapter **Page**

TABLES

FIGURES

EXHIBITS

FOREWORD

The National Center for Education Evaluation and Regional Assistance (NCEE) within the Institute of Education Sciences (IES) is responsible for (1) conducting evaluations of federal education programs and other programs of national significance to determine their impacts, particularly on student achievement; (2) encouraging the use of scientifically valid education research and evaluation throughout the United States; (3) providing technical assistance in research and evaluation methods; and (4) supporting the synthesis and wide dissemination of the results of evaluation, research, and products developed.

In line with its mission, NCEE supports the expert appraisal of methodological and related education evaluation issues and publishes the results through two report series: the *NCEE Technical Methods Reports,* which offer solutions and contribute to the development of specific guidance on state-of-the-art practice in conducting rigorous education research, and the *NCEE Reference Reports*, which advance the practice of rigorous education research by making focused resources available to education researchers and users of education research, to facilitate the design of future studies and help users of completed studies better understand their strengths and limitations.

Subjects selected for *NCEE Reference Reports* are those that examine and review rigorous evaluation studies conducted under NCEE to extract examples of good or promising evaluation practices. The reports present study information to demonstrate the possible range of solutions so far developed. In this way, *NCEE Reference Reports* aim to promote cost-effective study designs by identifying examples of the use of similar and/or reliable methods, measures, or analyses across evaluations. It is important to note that *NCEE Reference Reports* are not meant to resolve common methodological issues in conducting education evaluation. Rather they present information about how current evaluations under NCEE have focused on an issue or on selected measurement and analysis strategies. Compilations are cross-walks that make information buried in study reports more accessible for immediate use by the researcher or the evaluator.

This *NCEE Reference Report* is intended to help researchers select measures for future studies efficiently, assist policymakers in understanding the measures used in existing studies, facilitate comparisons of results across studies, and broaden understanding of these measures within the educational research community.

Selecting outcome measures for use in educational evaluation research is challenging. Researchers face a range of options without having the tools needed to quickly access information about existing and new measures. This report provides detailed, readily accessible, comparative information on the measures that have been used in approximately 40 NCEE evaluation studies to assess student outcomes, instructional practices, teacher pedagogical and/or content knowledge, and classroom environments.

ACKNOWLEDGMENTS

The authors would like to thank other Mathematica Policy Research staff members including Drs. Aaron Douglas and Sarah Dolfin for their careful review of the measure profiles and of the summary tables of recently developed measures. The Compendium also benefited from Dr. Phil Gleason's thorough review of Volume I and the glossary. In addition, we are indebted to Drs. John Burghardt and Irma Perez-Johnson, who provided input throughout the project and reviewed the Compendium in its draft form. The editing and production staff included Amanda Bernhardt, William Garrett, Cindy George, John Kennedy, Carol Soble, Linda Heath, Cindy McClure, Karen Groesbeck, and Jill Miller.

We would also like to acknowledge the developers of the measures included in this Compendium and are especially grateful to those who took the time to provide updated information and materials on their measures. Their involvement helped ensure inclusion of the most current information on the measures.

Finally, we would like to recognize and refer readers to three other measures compendia that were invaluable resources for several of the measures profiled in the current Compendium and that may add supplementary information to that provided here.

- *Resources for Measuring Service and Outcomes in Head Start Programs Serving Infants and Toddlers* (Kisker et al. 2003) provides descriptions of measures of child outcomes, parent/family background, and program environments as well as a model for measures profiles, for use by infant and toddler practitioners.

- *Early Childhood Measures Profiles Compendium* (Berry et al. 2004) focuses on early childhood measures of children's development and knowledge in a variety of domains.

- *Quality in Early Childhood Care and Education Settings: A Compendium of Measures* (Halle and Vick 2007) presents information about measures to observe early childhood settings (typically for infant, toddler, and preschool settings).

ABSTRACT

This report contains resources to help researchers and policymakers review measures used in NCEE evaluations of educational interventions. The measures included in the Compendium are applicable to settings for preschool through grade 12. The Compendium discusses criteria and their importance in selecting measures for assessing intervention impacts on student, teacher, and classroom outcomes, and presents profiles or table summaries of these measures. In expectation that the information in this document will be used under diverse circumstances for varied purposes, background information is presented in report format. The materials will be most useful when used in consultation with an assessment expert.

I. INTRODUCTION

The use of appropriate measures that support reliable and valid inferences is essential to ensure the rigor of evaluations of educational interventions. However, selecting measures for use in these evaluations is challenging, because researchers do not have ready access to comparative information about the focus, technical quality, and history of use of existing measures. The purpose of the Compendium is to provide this information for a compilation of identified outcome measures used in evaluations funded by the Institute of Education Sciences (IES) between 2005 and 2008. Existing resources such as the Buros Institute *Mental Measurements Yearbook* (Geisinger et al. 2007) provide information on the technical quality of measures, but accessing these reviews and searching through them can be costly and time consuming and may not help narrow the set of possible measures for further investigation. Other resources are also useful in helping researchers select measures, but they rarely address history of use in a way that allows for cross-study comparisons (Berry et al. 2004; Halle and Vick 2007; Kisker et al. 2003; McKenna and Stahl 2009; Strauss et al. 2006). Moreover, to our knowledge, there is no resource available that includes measures appropriate for assessing students in preschool through grade 12 and key characteristics of their teachers and classrooms. Nor is there a resource of measures across multiple domains used in evaluations of educational interventions funded by IES.

To help fill this gap in current resources, Mathematica Policy Research developed this two-volume Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions ("Compendium") as part of its work under the IES Analytical and Technical Support Contract. Volume I describes typical/common considerations when selecting measures and the approach used to collect and summarize information on the measures reviewed. It also provides a summary of the characteristics of these measures including domain, technical quality, and history of use in IES-funded evaluations. Volume II provides detailed descriptions of 94 measures including source information and references. The overarching goal of this Compendium is to make selecting, interpreting, and comparing measures in educational evaluations easier. More specifically, it is intended to (1) assist researchers and policymakers in understanding the measures used in recent IES-funded studies, (2) support and facilitate comparisons of results across these studies, and (3) provide information to aid researchers in selecting measures for future studies.

CRITERIA FOR INCLUDING MEASURES IN THE COMPENDIUM

The Compendium includes information on the measures used in 40 evaluation studies conducted by the Regional Educational Laboratory Program (REL) or in other recent, large evaluations funded by the National Center for Education Evaluation and Regional Assistance (NCEE)[1]. It focuses exclusively on measures used to gauge outcomes of students, teachers, and

[1] Generally, studies funded through NCEE contracts are mandated by Congress or are of national importance. REL studies are designed to address regional needs and issues. NCEE is the center through which the RELs are supported.

classrooms that were expected to change as a result of an educational intervention.[2] Measures designed to describe fidelity of implementation or capture student, teacher, or classroom characteristics that might influence responses to an intervention were excluded. The Mathematica team limited the review to the outcome measures used in REL and NCEE studies conducted between 2005 and 2008. The team only included studies that employed randomized controlled trials or regression discontinuity designs and also used outcome measures that were (1) available to other researchers and (2) had information available about psychometric properties. The Compendium includes both existing tools and study-developed measures that are available for use by other researchers. For measures with more extensive psychometric information that could not be easily conveyed in a table, the Mathematica team created profiles (compiled in Volume II). Chapter III provides more information about the formats and measures included.

The Compendium provides consistent information about the measures used in education intervention studies to help researchers to evaluate their quality and appropriateness for particular studies. That is, the Compendium uses a common format to describe the features of each measure based on information publicly available from publishers or developers. However, important caveats must be noted. First, a measure's inclusion in the Compendium should not be interpreted as an endorsement by IES or Mathematica, nor does it signify the quality of the measure. Second, for each measure, information on technical quality, cost, and ease of administration and scoring is presented in non-evaluative terms; therefore, the information presented should not be considered a formal critique of the measure. Third, project resources did not allow for an exhaustive review of all possible sources of information about a given measure, hence, the Mathematica team used a common set of resources to prepare measure profiles. Additional information may nevertheless be available from other sources. Fourth, the profiles do not provide information on the success or difficulties associated with use of the measures in a particular context (for example, region, grade, intervention type, or students' language status). Fifth, the Compendium provides links to relevant study descriptions or reports on the IES website but does not describe these studies in terms of the design, sample, intervention, research questions, or results. Finally, given the Compendium's focus on recently funded IES studies, researchers developing study plans may want to review other measures not included here.

SELECTING THE APPROPRIATE MEASURE

Researchers must weigh several factors when deciding how best to measure the outcomes needed to estimate the effects of an intervention or treatment. Factors include alignment of measures to the intervention, evidence of reliability and validity, use with diverse populations, and practical issues surrounding administration requirements, burden, and cost. Chapter II provides details about these factors; they are discussed only briefly here.

[2] Although identified as an outcome measure in the study in which it was used, the same measure might serve as mediator/moderator variable in some other intervention studies.

- **Appropriately aligned with the intervention's theory of change.** A first step in selecting a measure for an evaluation involves determining whether it captures information about outcomes that are aligned with the intervention and the theory of change. For example, if the purpose of the research is to estimate the relative effects of different elementary school mathematics curricula, the knowledge and skills targeted for assessment should represent the goals shared by elementary school mathematics curricula in general and the commonalities among different curricula. The measure must accurately capture information about the learning that results from children's exposure to each curriculum without bias toward any single curriculum or subset of curricula. So, for example, if the desired outcome of certain curricula is to develop children's mathematical reasoning, an assessment that disproportionately emphasizes precision and accuracy in computation may not be the best choice.

- **Supports reliable and valid inferences.** Researchers need measures that (1) produce a similar result with the same level of accuracy each time they are administered (reliable) and (2) accurately represent the constructs of interest (for example, beginning reading skills) that the measures purport to assess for a study's stated purpose (valid). Measures with low reliability estimates may not be sensitive to change (for example, the random error in the measure may obscure change over time), or they may result in biased results (for example, a measure may include systematic error related to the characteristics of students or classrooms). Measures that lack evidence of validity may lead to inaccurate inferences about the findings. For example, a measure of mathematics that relies heavily on language may indicate gains related to students' increased language skills rather than to increased understanding of mathematical concepts and applications. Reliability and validity are key considerations when reviewing a measure for inclusion in a study.

- **Suitable, equitable measurement of diverse populations.** When the outcomes of interest are students' academic achievement or other dimensions of student development, the measures must accurately capture the knowledge, skills, and behaviors in the population of students under study. In most cases, the study population includes students from a wide range of backgrounds. To generalize to the broader student population, the measures should have been used with different populations of school-age children, and the standardization sample should include male and female students of different races, ethnicities, and socioeconomic backgrounds. The measures must provide an accurate assessment of students in the range of achievement targeted by the study, or for the general population, throughout the achievement range. If a study focuses on the "average" student, findings from the study may not be generalized to high or low achievers. In many cases, consideration must also be given to how the measure will perform for two groups of students: (1) those with limited English skills and (2) those with disabilities. These groups have become increasingly important given that approximately 9 percent of school-age children have disabilities (U.S. Department of Education 2007), and an estimated 20 percent of the school-age population speaks a language other than English at home, and a similar percentage of children have at least one parent born outside the United States (Federal Interagency Forum on Child and Family Statistics 2008).

- **Practical considerations (administration requirements, burden, and cost).** Because research studies have resource limitations, additional considerations such as (1) staff training and skills needed for administration and scoring; (2) costs for materials, training, administration, and scoring; and (3) burden on participants are often taken into account when evaluating measures for a given study (Ross et al. 2005). These practical considerations have implications for the reliability of scores and validity of the inferences based on them. For example, poorly trained staff, administration of assessments under unusual conditions, and other sources of error can affect the reliability and validity estimates. Some measures require significantly more training than others to ensure consistent use; thus, researchers must consider the qualifications of the staff who will administer the measures and the resources available for training. In addition, researchers should estimate the burden the measure would place on participants in the evaluation. Measures of teacher instruction and classroom environments are often unobtrusive and place little or no direct burden on those observed. Other measures, such as student assessments, vary in the time required of students and ease of administration. Cost is also an important consideration. Royalty fees (if applicable), materials, training, and scoring can all involve costs. The time required for administration and scoring and the amount of training needed to ensure consistency both have cost implications. Some assessments require specialized scoring software, potentially increasing costs. Observational measures vary in the amount of coding time needed to convert raw data to usable scores. To help guide selection decisions, the Compendium includes information about the initial cost of required materials and training; the knowledge, skills, and training required for administration and scoring; and administration time.

Perhaps one of the most important decisions researchers must make is whether to use an existing, available measure; adapt an existing measure; or develop a new measure that meets the particular needs of the evaluation. As mentioned earlier, the measures selected for the Compendium reflect a mix of existing tools (commercially and noncommercially available) and study-developed measures. While there are many reasons for deciding to develop a new measure to meet the needs of a study, researchers should have a clear understanding of (1) the strengths and limitations of existing measures, (2) the resources required to create a new measure and document evidence of reliability and validity, and (3) the risks involved with using a new measure, especially in the context of a large-scale project.[3] Chapter II provides additional information about both the theoretical and practical aspects of selecting measures for studies of education interventions.

[3] Developing new measures is a complex process and one that may require expertise and resources beyond what is available for evaluations studies. See for example, Linn and Gronlund (2000); McDermott (1993); Pew Research (http://people-press.org/methodology/questionnaire) for discussions of assessment development and related issues.

OVERVIEW OF STUDIES AND MEASURES

In the seven years since the Institute of Education Sciences was established by the Educational Sciences Reform Act of 2002 (P.L. 107-279), IES has funded numerous evaluation studies that have used randomized controlled trials or regression discontinuity designs. The studies have investigated the impacts of a range of interventions and professional development approaches on student and teacher outcomes, including school organization and leadership, teacher quality and preparation, curriculum and technology, supplemental instruction and support, and support for students' social and behavioral development. The research has focused on elementary and middle school grades, although studies of prekindergarten and high school students have been conducted as well.

The Compendium includes measures used in 40 NCEE or REL studies to assess student outcomes or outcomes related to the teacher or classroom experiences. While the ultimate outcome of interest in 36 of these studies is student achievement, skills, or behaviors, these measures differ in what they assess and how they measure student achievement and skills in different subject areas and domains. Specifically, the student achievement/development measures are designed to assess students' academic performance, achievement, or other relevant areas of development (for example, approaches to learning/motivation, social-emotional well-being, and leadership). Some of the 36 studies have used commercially available measures such as the Peabody Picture Vocabulary Test-4 (Dunn and Dunn 2007), the Woodcock-Johnson III (Woodcock et al. 2001, 2007), or the Social Skills Rating System (Gresham and Elliott 1990). As described in Chapter IV, the overall quality of commercially available measures included in this Compendium varies with respect to the standardization sample, reliability data, and validity evidence. Sometimes researchers developed their own measures to meet the needs of a particular study. For example, the Evaluation of Reading Comprehension Interventions Study (James-Burdumy et al. 2009) developed measures of reading comprehension of expository text related to science and social studies content (Educational Testing Service 2007). The REL-West study of High School Instruction with Problem-Based Economics (Regional Educational Laboratory Program 2008a) used a rubric with categories to score students' responses to economic problem-solving tasks assessing the students' understanding, argumentation, misconceptions, and use of existing knowledge (see Table B.1 in Volume II).

The development of measures for a single study addresses the content or purpose of the study but the psychometric properties of these measures may be unknown when initially used, and limited even after the initial use (as shown in the summary tables presented in Volume II). Researchers often lack the time and resources needed to collect evidence of reliability and validity before using a newly developed measure in a study. In our experience, the collection of reliability data and evidence of validity frequently demands more time and resources than are available for the main study. This tension exemplifies a central issue in conducting rigorous evaluations—appropriately aligning measures that have evidence of reliability and validity to specific interventions and targets of behavioral change.[4]

[4] In this Compendium, appropriate alignment means that a measure is used to assess a construct explicitly targeted as an outcome of the intervention and not simply as a fidelity measure. For example, an appropriately aligned classroom practice measure used in a science curriculum evaluation would assess aspects of teaching that

Twenty-seven of the 40 REL or NCEE studies funded by IES include measures of outcomes related to students' classroom experience, teachers' knowledge, teachers' instructional practices, the quality of instruction, and the overall quality of classrooms. The teacher knowledge assessments in the Compendium include measurement of subject content knowledge and some newer areas of investigation, including pedagogical knowledge and pedagogical content knowledge. For example, the study of Professional Development Strategies in Mathematics (Regional Educational Laboratory Program 2008b), conducted by American Institutes for Research (AIR) and MDRC, included the Teacher Knowledge Inventory (TKI) as an outcome. Measures of classroom practices capture relevant dimensions of instructional and environmental quality, including teacher behaviors, classroom management, peer interactions, classroom climate, and the adequacy and supportiveness of the classroom and school settings for learning. The National Evaluation of Early Reading First (Jackson et al. 2007) used the Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms et al. 1998) to measure the language and literacy environment and overall quality of preschool classrooms. The Appalachia REL used the Early Language and Literacy Classroom Observation (ELLCO; Smith et al. 2008), along with other existing measures such as the Early Literacy Observation Tool (E-LOT; Grehan and Smith 2004) to measure the use of research-based literacy strategies in preschool and kindergarten classrooms.

Although some of the studies used commercially available measures, twelve developed their own measures or adapted existing measures for their particular purpose. For example, the National Evaluation of Reading First developed the Instructional Practice in Reading Inventory (IPRI; Dwyer et al. 2007) after reviewing existing reading practice tools and finding that none met the specific needs of the study. In some cases researchers found that as new policy priorities and research areas emerged, few appropriate measures were available. For example, when IES funded the Evaluation of Mathematics Curricula (Agodini et al. 2008; Agodini et al. 2009), the few existing choices for measures of mathematics instruction were oriented toward a particular theoretical approach (such as basic skills instruction) and therefore were not appropriate for a study of multiple curricula that varied in theoretical approach to instruction. In response, the research team developed a new observational tool and conducted its first "test" in the context of a large random assignment curriculum evaluation. Researchers, funding agencies, and policymakers often face similar decisions that require weighing a preference for the use of existing measures with available information to support their technical quality against the need for a new measure. The issue of using existing tools with a history of use emerges even more strongly for measures of teacher knowledge of content and pedagogy; few measures exist (for example, Hill et al. 2008). Thus, variability in the evidence for the soundness of many classroom and teacher measures is greater than for student outcome measures included in this Compendium (see Chapter IV).

(continued)
researchers have found to support students' science achievement, not just fidelity to one of the curricula being studied.

COMPENDIUM ORGANIZATION AND CONTENTS

The Compendium contains two volumes. Volume I details measure selection criteria, the methods used to construct the Compendium, and the information included in the measure profiles and summary tables. Volume II presents the measure profiles and summary tables as well as a glossary.

Volume I includes three chapters in addition to this introduction. Chapter II describes a general set of criteria—substantive/theoretical, operational, and cost-related—that can be used to evaluate the quality of measures and provide a basis for choosing ones that meet the goals of a study. Chapter II also presents an overview of the most common psychometric properties considered in selecting a measure. For purposes of this Compendium, the Mathematica team focused on reliability and validity, and excluded discussions of other psychometric properties such as scaling, calibration, item analysis, or test information function that may be appropriate considerations for some studies (Crocker and Algina 1986; Nunnally and Bernstein 1994; Wright and Stone 1979). Chapter III describes the methods used to identify the measures included in the Compendium, information sources, and the format of the profiles and summary tables for the selected measures described in Volume II. Chapter IV includes an overview of the measures and summarizes evidence on their technical properties; cross-measure synthesis tables IV.1, IV.2, and IV.3 provide information gleaned from the measures profiles and tables for student, teacher, and classroom outcomes, respectively. These tables allow the reader to compare measures across features such as domain, grade/age range, type of assessment, technical quality, and cost considerations.

Volume II provides an orientation to the profile format and features detailed information about 94 measures assessing student, teacher, and classroom outcomes in NCEE or REL educational intervention evaluations. The individual measure profiles in Volume II present information about what is measured and how the developers operationalize the targeted domain; method of administering the measure and time needed; personnel and training requirements; availability in languages other than English (as appropriate); cost of the measure; representativeness of the standardization samples; scores, uses, and interpretability of the measure; and evidence of reliability and validity. For newly developed study-specific measures, a summary table presents a description of what is assessed, administration times, and, if available, evidence of reliability and validity.

II. CRITERIA FOR GUIDING MEASURE SELECTION

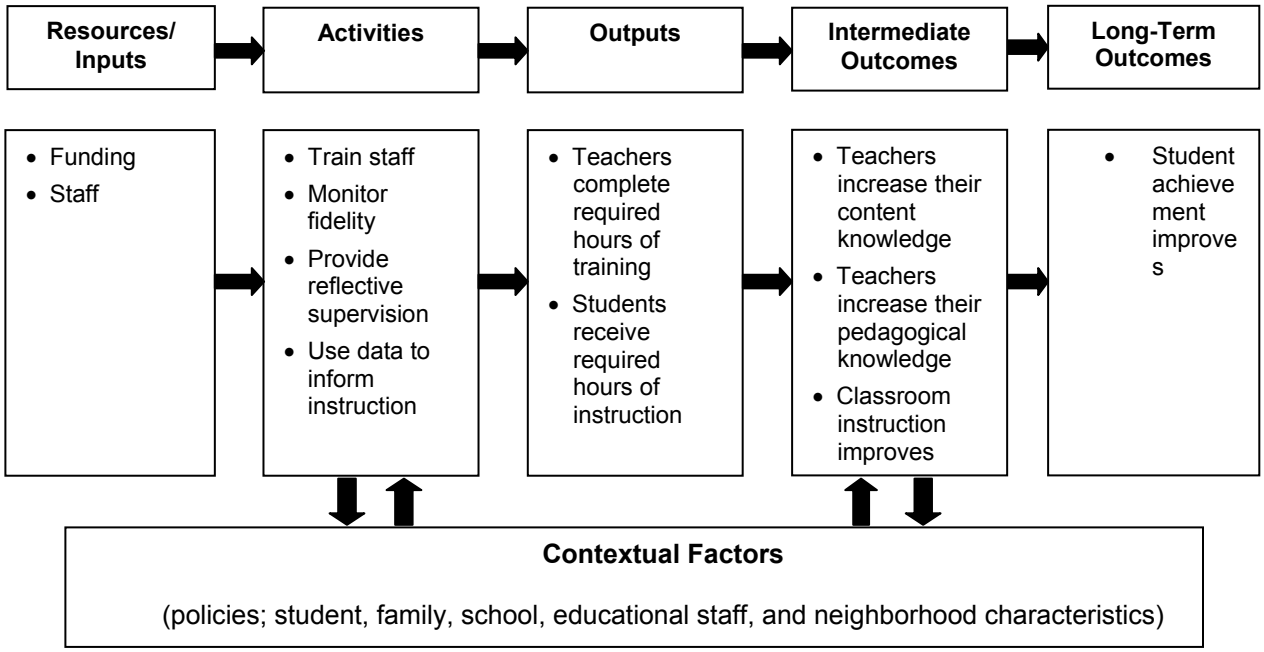
Selecting measures for evaluations of educational interventions requires weighing a variety of theoretical and practical matters. This chapter describes the range of issues that researchers and their funding agencies typically consider, including the alignment of measures with the intervention's theory of change, appropriateness for the study population, adequacy of the measures' psychometric properties (specifically, evidence of reliability and validity in prior use as well as information about the sensitivity of the measure), a history of use in studies of similar size and scope (including sensitivity to change within the prescribed time period), and burden and cost.

ALIGNMENT WITH INTERVENTION'S THEORY OF CHANGE

Researchers designing an evaluation typically begin by stating the research questions and hypotheses they plan to test. Clearly describing the intervention's logic, or theory of change, is an important next step that facilitates the development of a measurement plan aligned with the research questions and the expected outcomes. The logic model then takes the assumptions and inputs (resources, staff, and other supports) that drive an intervention and links them to activities required to meet specific goals, outputs, and hypothesized intermediate and long-term outcomes (for example, W.K. Kellogg Foundation 2001). Figure II.1 provides an example of a basic educational intervention logic model, which can be a powerful tool to help intervention developers, implementers (for example, school administrators and staff), and evaluators forge a common understanding of how the intervention is expected to work.

FIGURE II.1

BASIC EDUCATIONAL INTERVENTION LOGIC MODEL



A logic model can also facilitate the development of a credible and comprehensive measurement plan aligned with the intervention's goals, activities, and hypothesized outcomes. A strong plan would include indicators and outcome measures for each main component of the intervention's logic model, thus facilitating analyses that may help make sense of evaluation findings and produce insights to inform program refinement. For example, by measuring the number of teacher training hours completed for a new curriculum and comparing it to those recommended by the curriculum authors/publisher, evaluators could document whether the intervention demonstrated fidelity to the training model. If the number of teacher training hours is below the recommended level and there are minimal or no impacts, a second generation of programs might aim to improve fidelity of training (for example, by keeping the number of teacher training hours consistent with the recommended level).[1]

Aligning measures with the expected outcomes of the intervention involves selecting measures that assess aspects of change arising from teacher/school and student exposure to the intervention. For example, a science curriculum focused on using the scientific method as an approach to asking questions about the natural world might affect both scientific knowledge and broad problem-solving strategies but is unlikely to change literacy skills. Close alignment between an intervention and assessment of its domain-specific outcomes is warranted when the intervention is focused on changing outcomes in only one domain (for example, literacy or mathematics).

At the same time, it may be important to include measures of possible unintended or adverse effects. The concern is that schools or teachers focused on improving one academic area or set of behaviors (particularly in elementary school) may inadvertently reduce the emphasis placed on another area. This potential displacement of attention, classroom instructional time, student reinforcement, or teacher motivation may warrant allowing additional time for evaluation data collection to measure outcomes in areas other than those directly targeted by the intervention.

Although selecting outcome measures that are well aligned with the intervention is critical in the design of a meaningful evaluation, researchers must be careful not to overalign their measures to the specific intervention (such as by using intervention-specific activities as the outcome measure). For example, in a study of a vocabulary intervention, an assessment that included only the words taught in the intervention would be overaligned. Similarly, if an assessment included a task that was used to teach the students in the intervention group but that would be otherwise unfamiliar to most students, the assessment would not accurately assess the knowledge of the students in the comparison group. This could lead to finding group differences that are caused by the focus of the assessment rather than by actual differences in knowledge.

When an intervention is more global or comprehensive, the challenge is to select a parsimonious set of outcome measures that assess a range of potential impacts (for example, a

[1] Although many of the examples provided in this chapter describe selection of measures for curriculum-level interventions, other types of educational interventions may include more global changes in schools. These can include implementation of charter schools or alternative routes to teacher certification. Research questions and measurement approaches for these types of interventions often focus on the impacts of such changes on students' academic performance.

reading comprehension intervention may improve student achievement in literacy, science, and social studies). As described by Schochet (2008), the tendency in intervention research is to measure and report on many outcomes rather than to focus on a small set of domains and measures. This can be problematic because so many statistical tests and comparisons are then made between the treatment and control/comparison groups that impacts may be found by chance. Researchers and funding agencies must strive to identify a small number of key outcomes a priori, and measure and analyze them for impacts.

When assessing a measure's alignment with the intervention's theory of change, evaluators must consider the length of time it will take to observe hypothesized changes in outcomes, and design the evaluation and measurement plan accordingly. For example, depending on the complexity of a new curriculum and its training requirements, changes in teacher knowledge and practices may reasonably be expected within the first six months to one year of implementation, whereas student academic outcomes may take a full year or more to improve. Drawing upon the research literature about similar interventions and observed impacts, researchers must work to align a study's measurement schedule to the expected timing of the changes.

In studies of educational interventions, however, program implementation and measurement schedules are often driven by the school calendar. Often, teachers are expected to learn a new skill or curriculum during a summer workshop (possibly augmented by professional development in-service training during the school year and/or coaching) and implement it with fidelity and good quality in the school year. Using this design, students are assessed at baseline (before or as soon as possible after the start of the school year) and then again at the end of the school year. Measures that assess teacher knowledge or classroom practices expected to change because of the intervention may be administered according to the same timing. Designers and evaluators of the intervention must consider whether this timing is optimal, especially when an intervention requires substantial teacher knowledge and practice with students. At least two of the REL randomized control trials include teacher training/coaching a year prior to student entry into the study (for example, Midwest REL's Measures of Academic Progress and Central REL's Classroom Assessment for Student Learning Studies). Careful consideration of the timing and intensity of teacher training and professional development supporting high-quality implementation, as well as the required "practice" time teachers need to demonstrate fidelity, are important drivers of the timing of any assessment of outcomes.

APPROPRIATENESS FOR THE STUDY POPULATION

Variation in the characteristics of participants in educational intervention research (school districts, schools, administrative staff, teachers, students, and families) has implications for measures selection. Researchers want to select and use measures that have a demonstrated history of assessing outcomes accurately, fairly, and consistently in populations similar to those that will participate in the planned study. Authors and publishers want to provide evidence that their measure is appropriate for a range of potential study participants, and produces results that can be used to generalize about performance across samples. Examples of sample characteristics researchers must consider when reviewing the appropriateness of a measure include gender, age, grade level, primary language, disability status, race/ethnicity, and socioeconomic background.

In addition, researchers may want to consider whether the measure was developed with samples from both urban and rural areas and from different regions of the country. If a measure has norms, that means that some attempt has been made to gather information about how the measure performs when used with a sample representing a given population. When reviewing norming sample information and evidence for reliability and validity across sample characteristics, researchers consider the relative strengths of each measure for a given purpose. For example, a reading assessment in English may not be appropriate for some students who are primarily Spanish speakers, but when measuring the outcomes of an English reading instruction program for these same students, an assessment in Spanish may not be appropriate. In other words, an assessment is not reliable and valid in and of itself; rather, the use of the assessment must be reliable and valid (Nunnally and Bernstein 1994). Selecting assessments that are likely to support reliable and valid inferences requires consideration of the appropriateness of the measure for the target population and the purpose of the study.

ADEQUACY OF PSYCHOMETRIC PROPERTIES

The field of measurement includes a variety of indicators used to describe the psychometric properties of a measure (that is, how well the measure performs). The two main constructs used for evaluating the psychometric or technical adequacy of a measure are validity and reliability. Validity refers to the extent to which the results of a measure serve their intended use. Reliability is concerned with the consistency of results when the same measurement procedure is applied more than once. Consistency of measurement provides a necessary but not sufficient condition for validity and it indicates the degree to which test results are dependable (that is, free from error). (See Volume II, Appendix E glossary for definitions of key terms.)

Professional organizations and experts have written extensively about gathering and interpreting evidence of the psychometric properties of measures. Among the most influential works is the volume *Standards for Educational and Psychological Testing*, developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). These standards provide guidance for gathering and reporting evidence to support interpretation. Although there are no hard and fast rules for interpretation, the highest-quality measures provide an accumulation of evidence to indicate that the assessment consistently measures what it purports to measure. In what follows, we provide some common categorizations of reliability and validity, and methods for gathering and interpreting evidence for each. Our intent is to explain the indicators most commonly reported in the measures reviewed. This discussion is not meant to be comprehensive. Volumes have been written about psychometrics and different approaches to evaluating the properties of a measure (Crocker and Algina 1986; Nunnally and Bernstein 1994). Additional characteristics of a measure may be important for a particular use. For example, researchers may want to know more about the point-biserial correlations and the size of the item gradients if a measure is going to be used to make fine discriminations. The measures that we reviewed did not regularly provide this level of detail about the characteristics of the items.

Reliability

Reliability refers to the accuracy (consistency and stability) of measures. Consistency can be defined across two forms of the same measure (alternate forms reliability), across items within a measure (internal consistency reliability), across raters or observers (inter-rater reliability) or across occasions (test-retest reliability). Reliability coefficients report the degree to which a test is free from error that affects test scores or performance. These coefficients range from 0 to 1.0, with a higher value reflecting greater dependability and less error. A number of factors may affect the reliability estimate, including homogeneity of test takers (reliability estimate will be lower than if the sample was more heterogeneous) and length of the test (longer tests are generally more reliable than shorter tests comprised of items with similar characteristics).[2] Researchers generally apply common "rules of thumb" when judging the reliability of a measure for a proposed purpose. For example, researchers and assessment developers often require that assessment and screening tools have evidence of reliability values of 0.70 or higher to support inferences about the measure (Bacon 2004; Cohen 1977; Litwin 2003; Nunnally 1978), however, the minimal level of reliability that is recommended differs according to the type of inference that will be made about the results. Nunnally and Bernstein (1994) state that reliabilities of 0.80 are sufficient for research examining mean between group differences, but inadequate for making decisions about individuals. For making decisions about individuals, reliability of 0.80 to 0.90 or higher is recommended depending on the stakes involved[3] (Nunally and Bernstein 1994; Salvia and Ysseldyke 2004). Internal consistencies of 0.80 to 0.90 and test-retest reliabilities of 0.70 are recommended by some researchers (Andrews et al. 1994) as minimal acceptable standards for outcome studies.

Four approaches to measuring reliability are:

- ***Alternate form reliability.*** Publishers may provide two or more versions of the same measure so that the same skills or behaviors can be assessed multiples times (as in pre-post or longitudinal studies with the same group of students). Using alternate forms reduces concerns that students' scores may change due to "learning the test" or practice effects from repeated administration of the same items. To demonstrate that the two forms of the measure are essentially equivalent, a group of students completes both (the time between administrations may vary). Alternate form reliability is demonstrated if the scores on the two forms are highly correlated.

- ***Internal consistency reliability.*** Methods of estimating reliability that require only a single test administration are referred to as measures of internal consistency or homogeneity. They are based on estimates of how well items within a test measure the same domain or construct. Three classical measures of internal consistency[4] are

[2] See Traub (1994) for a discussion of factors affecting the reliability coefficient.

[3] If the assessment is used for screening purposes, 0.80 is acceptable. For other individual level decisions, .90 is the recommended minimum level.

[4] Newer measures that utilize item response theory (IRT) evaluate the dimensionality of a test and how precise the measurement is, using fit statistics that indicate how well the model fits the data and how well the measure (and the items in the measure) discriminate among individuals. IRT offers more information about measures and the

split half, Cronbach's coefficient alpha (coefficient α), and Kuder-Richardson formula (KR-20) (Sattler 2001)). Split-half reliability refers to the correlation between the odd- and even-numbered items in an assessment. The resulting correlation is the reliability of each half of the test; the Spearman-Brown Prophecy formula can be used to provide an estimate of the reliability of the whole test. Coefficient α and KR-20 provide an estimate of what the average reliability would be if all possible ways of splitting the test into halves were used. KR-20 can be used with dichotomously scored items (such as right/wrong or true/false), whereas coefficient α can be used with items that have multiple response options such as rating scales.

- ***Inter-rater reliability.*** When assessments involve some level of subjective judgment, inter-rater reliability provides an estimate of the consistency of different scorers or observers. Lack of consistency results in error of measurement and reduces the reliability coefficient. Inter-rater reliability can be estimated in a number of ways, including percentage agreement, Pearson Product Moment correlation, intraclass correlation, and kappa coefficient (Cohen 1960; Frick and Semmel 1978; Shrout and Fleiss 1979). Inter-rater reliability is particularly relevant for measures that require an observer to score another person's behavior or complete a rating or checklist describing the behavior observed. To use observational assessments in evaluation, researchers and assessment developers must be sure that ratings can be made consistently, that is, that two people observing the same event would agree on the extent to which something did happen. Measures that require complex scoring or high levels of inference also require more time to train observers to reach acceptable levels of reliability. One index of consistency is the extent to which two trained assessors or observers obtain the same information. It may be reported either as the correlation between the scores or ratings obtained by the two observers, or as the percentage of items on which the two agree. In research, inter-rater agreement is often used as a certification criterion for assessors/observers. For their data to be considered reliable, at the end of training and during in-field data collection, data collectors must meet the inter-rater reliability standards set by the study.

  Inter-rater reliability is particularly important for measures that assess complex behaviors such as teaching. As described by Raudenbush and Sadoff (2008), low inter-rater reliability is evidence of one specific source of error that may contribute to the small size of the associations between quality measures and student outcomes. When reporting inter-rater reliability, authors and users of classroom observation measures may report only inter-rater exact agreement—the proportion of instances on which an observer agrees perfectly with a trainer or gold standard coder. For some classroom observation measures, however, only observer agreement within one rating scale point is used as the criterion (for example, ECERS-R). When evaluating a

(continued)
items that comprise them than is found with classical approaches to the development of measures (Embretson and Hershberger 1999).

measure, researchers must decide the level of agreement needed to meet the reliability targets for their study and provide the desired level of precision of measurement.

- ***Test-retest reliability.*** A measure is reliable to the extent it yields the same result on two different occasions (consistent over time). Test-retest reliability involves testing the same group of individuals (or observing the same teacher or classroom) at least twice within a specified period of time. The time between testing is dependent on the stability of what is being measured. The reliability coefficient is then obtained by correlating the scores from the two administrations. These reliability estimates may also be referred to as a coefficient or measure of stability. The higher the test-retest reliability, the more stable the assessment tool is considered to be. Longer periods between administrations of the same assessment typically will reduce reliability, partly because the individual's situation (for example, skill level) can be expected to change.

Each method for estimating reliability described above accounts for a different source of measurement error (forms, items, raters, or occasions). In contrast, generalizability (G) theory estimates all sources of measurement error simultaneously and evaluates the generalizability (reliability) obtainable under different conditions. For example, the researcher can estimate the reliability with two raters and three observations of teaching, and explore the changes in reliability with an increase or decrease in raters and/or observations. Does the reliability change if the observations are conducted twice instead of three times? Although the details are beyond the scope of this report, the interested reader is referred to Shavelson and Webb (1991) for an introduction or Brennan (1983) for a technical discussion.

Validity

The concept of validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. Because the central issue is how well an instrument measures what it purports to measure, one validates the use to which a measure is put rather than the instrument itself. Tests or measures may be valid for one purpose or with one group of respondents but not another. Evidence may be accumulated in a variety of ways. We focus here on three aspects or types of validity—content, construct, and predictive.[5]

- ***Content validity*** refers to how well the content of the measure samples the situations or subject matter about which conclusions are to be drawn. This indicator of validity provides information about whether the measure includes items that are relevant to and representative of the construct it is supposed to assess. Typically, there are no statistics associated with content validity. Instead, it is based on the professional judgment of experts who review the items to verify that the measure represents the

[5] Current conceptions of validity view all forms of validity as special instances of construct validity (Messick 1993). Here in this discussion, we consider criterion-related validity as a form of construct validity. Some authors present predictive and concurrent validity as two types of criterion-related validity (Crocker and Algina 1986).

domain that the developer intended and that the items, when appropriate for some measures, provide variety and a range of difficulty.

- ***Construct validity*** refers to the degree to which an assessment measures the theoretical construct it is intended to measure and confirms that inferences based on the assessment are relevant to the construct. Several approaches are used to provide evidence of construct validity. First, using correlational analysis, researchers look for a positive relationship with other measures of that construct or a similar construct to provide evidence of convergent validity, and weak or negative relationships with measures of dissimilar constructs as evidence of divergent or discriminant validity (Campbell and Fiske 1959). Second, researchers look for a relationship between the measure of interest and a different measure or criterion. For example, one could compare an achievement test score with performance on a task that is an independent measure of some skill related to the construct, such as receiving a passing grade. If higher test scores are associated with passing grades, the test demonstrates criterion-related validity. Third, factor analysis can be used to examine whether the observed dimensionality of the measure is consistent with the theoretical dimensions of the construct, and to examine the strength of the associations among the different items and dimensions (Crocker and Algina 1986).

- ***Predictive validity*** refers to the extent to which performance on one measure predicts future performance on another measure or criterion. To measure predictive validity, researchers look at the association between two measures that are separated by some period of time. If, for example, a measure of vocabulary in kindergarten is highly correlated with an assessment of reading ability in grade 2, the vocabulary assessment demonstrates evidence of predictive validity; kindergarten vocabulary predicts second-grade reading achievement. In some cases, researchers use performance at some later point in time as the criterion. For example, researchers might show a positive correlation between performance on a kindergarten counting task and grade 2 mathematics report card grades as evidence of predictive validity. In general, predictive validity is weaker for assessments administered to children younger than age 6 (Kim and Suen 2003; LaParo and Pianta 2000).

HISTORY OF USE IN STUDIES OF SIMILAR SIZE AND SCOPE

To increase comparability with other national studies and educational evaluations, as well as to avoid the costs of measurement development, researchers and their funding agencies should look for measures used in other studies with a demonstrated ease of administration and evidence of technical quality. In preparation for conducting large or resource-intensive studies, our experience suggests that research teams should conduct an extensive review of measurement domains, constructs, and specific measures that have been used to assess outcomes of similar interventions. In addition, they will often contact researchers who most recently used a given measure to learn about any problems with training assessors or collecting or analyzing the data. Most research teams would only consider developing a new measure over one used previously and successfully in comparable studies if all this information has been reviewed, other criteria described above are considered, and a new measure offers improvements in important areas

(such as alignment with the intervention, stronger evidence of reliability, or greater sensitivity to change[6]).

In the face of an identified gap in the available measurement tools, a research team may work closely with its funding agency and other research teams to select from options that do not have a history of use in studies of similar size and scope. These may be more "experimental" measures, or ones designed and used by academic researchers in small-scale studies. Although these measures may be challenging to implement in a large-scale data collection effort, they may be the most appropriate method for assessing a critical outcome. As described in Chapters III and IV, the NCEE and REL study measures in this Compendium include more than 30 newly developed measures. Study teams embarked on what in some cases amounted to long and intensive measurement development efforts because the existing measures were deemed inadequate to meet study needs. Given the complexity of issues and the time and resources needed to design technically sound measures, research teams should exercise caution in opting to design their own measure rather than use existing measures with an established history of use.

REASONABLE BURDEN AND COST

NCEE and REL studies are often large, multisite projects that require centralized training of data collectors as well as complex data collection plans that include student assessments and/or measures of teacher knowledge and observations of classroom practices. The burden on school staff, students, and parents (if they are included as study informants) may be high, ranging from 30 to 60 minutes for each student assessment, teacher, or parent questionnaire to three or more hours for a classroom observation. Excessive burden may increase the risk of respondents refusing to participate in later data collection activities. Managing and reducing study burden as much as possible is critical to ensuring good response rates over time (Groves and Couper 1998; Hoogendoorn and Sikkel 1998). It is especially important if the study includes longitudinal followup of students into later grades. After bearing the costs of locating students who have moved, researchers need to minimize all other potential sources of attrition, including respondent frustration with the length of assessments and interviews. Federally funded contract research requires clearance by the Office of Management and Budget to guard against undue burden on participants.

In addition to burden, cost considerations are also an important factor in selecting measures. Careful attention to the psychometric properties of measures can reduce the costs of a study. When the reliability of a test is high, smaller sample sizes are needed to detect effects (Bacon 2004). For example, to use a test with reliability of 0.70 and a pretest that is correlated with the post-test at $r = 0.50$, the research team would need a sample size of approximately 300 students in the treatment and control groups. Detecting the same effect size using a test with a reliability of 0.50 would require approximately 400 students in each group (Bacon 2004). Any decrease in

[6] This is a consideration especially when researchers want to assess change in students who would typically fall in the tail of a distribution where measurement error is typically greater. Use of measures that are adaptive and target the students' ability levels avoids ceiling and floor problems and decreases measurement error (Embretson and Hershberger 1999).

sample size can save money in recruiting participants, obtaining permissions, and handling administration and data entry. Similarly, generalizability studies can be used to estimate the number of raters, items, observations, or occasions needed to obtain a particular level of reliability (generalizability). Researchers can consider costs associated with various design tradeoffs such as decreasing the number of raters and increasing the number of observations. In this way, a study can be designed to minimize costs while achieving target levels of reliability.

III. DEVELOPMENT OF PROFILES AND SUMMARY TABLES FOR THE COMPENDIUM

This chapter describes the methods used to summarize major characteristics of measures of student achievement/development, teacher knowledge, and classroom practices and settings. The following section describes the information sources consulted on the selected measures as well as the Compendium's presentation format. In addition, an overview of the key content of the profiles is presented to ensure the use of common terminology, concepts, and language across measures.

METHOD

The number of measures assessing student achievement and development, teacher knowledge, and classroom practices and settings from preschool through secondary school is too extensive to permit a review of all available measures. Therefore, as described in Chapter I, the Mathematica team first established inclusion criteria, focusing on outcomes of educational intervention evaluations used in an NCEE or REL study. Next, the team used similar types of sources (for example, administration or technical manuals) to guide the collection of information on the selected measures, thus ensuring consistency in the level of information across reviews. As discussed later in this chapter, the use of similar sources presented a challenge for measures that are not commercially available, which often have no formal manual. Based on the available evidence of reliability and validity, measures summaries took one of two formats that are described in more detail in this chapter: (1) a multipage profile or (2) a summary table.

Application of the criteria and available information resulted in the inclusion of 94 measures in the Compendium. Fifty-nine measures had enough information to support creating individual profiles; the others had less information available and thus are described only in summary tables. Tables III.1 through III.3 list the selected measures, the NCEE or REL study using them, and the format of the measure review (profile or summary table). For ease of locating a specific type of measure, the Mathematica team grouped measures within an outcome type (student, teacher, classroom) by category. For example, student achievement/development measures are grouped into categories such as Comprehensive Achievement Tests, Literacy, Reading, and so on (the left-hand column of Table III.1). These categories are not intended to reflect underlying constructs of the measures; they serve only to facilitate the search process for researchers interested in a specific outcome category.

TABLE III.1

STUDENT ACHIEVEMENT/DEVELOPMENT MEASURES
INCLUDED IN THE COMPENDIUM

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Comprehensive Cognitive and Achievement Tests	Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)	Program for Infant/Toddler Care (REL-West)	Profile
	Kaufman Test of Educational Achievement, Comprehensive Form, Second Edition (KTEA-II)	Accelerating Language Development (REL-Southeast)	Profile
	Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) and Achievement Level Tests (ALT)	Effects of Success in Sight (REL-Central)	Profile
		Professional Development Strategies in Math	
	Stanford Achievement Test Series, Tenth Edition (Stanford-10)	Reading First	Profile
		Enhanced Academic Instruction in After-School Programs	
		Alabama Math, Science, and Technology Initiative (REL-Southeast)	
		Principles-Based Professional Development (REL-Pacific)	
		DC Opportunity Scholarship	
		Reading and Mathematics Software Products	
	TerraNova 3	Odyssey Math® (REL-Mid-Atlantic)	Profile
		Connected Mathematics 2 (REL-Mid-Atlantic)	
		Small Group Mathematics (REL-Southwest)	
		Different Routes to Certification	
	Woodcock Johnson-III Normative Update (WJ III NU)	Opening the World of Learning (REL-Appalachia)	Profile
		Program for Infant/Toddler Care (REL-West)	

TABLE III.1 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Literacy, Reading	6+1 Trait Writing Scoring Guide (Rubrics)	6+1 Trait Writing Model (REL-Northwest)	Profile
	AIMSweb Oral Reading Fluency	Closing the Reading Gap	Profile
	Dynamic Indication of Basic Early Literacy Skills (DIBELS), Sixth Edition	Opening the World of Learning (REL-Appalachia)	Profile
		Comprehensive Teacher Induction	
		Enhanced Academic Instruction in After-School Programs	
		Project ELLA	
	Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4)	Thinking Reader (REL-Northeast & Islands)	Profile
		Principles-Based Professional Development (REL-Pacific)	
		Collaborative Strategic Reading (REL-Southwest)	
	Group Reading Assessment and Diagnostic Evaluation (GRADE)	Reading Comprehension	Profile
		Adolescent Literacy Across the Curriculum (REL-Midwest)	
		Collaborative Strategic Reading (REL-Southwest)	
		Enhanced Reading Opportunities	
		Closing the Reading Gap	
	Indicadores Dinámicos del Éxito en la Lectura (IDEL), Seventh Edition	Project ELLA	Profile
	Metacognitive Awareness of Reading Strategies Inventory (MARSI)	Thinking Reader (REL-Northeast & Islands)	Profile
	Phonological Awareness Literacy Screening (PALS) PreK, PALS-K, PALS 1-3	Opening the World of Learning (REL-Appalachia)	Profile
	Science Reading Comprehension Assessment	Reading Comprehension	Profile
	Social Science Reading Comprehension Assessment	Reading Comprehension	Profile
	Stanford Diagnostic Reading Test, Fourth Edition (SDRT 4)	Project CRISS Reading Program (REL-Northwest)	Profile

TABLE III.1 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
	Test of Preschool Early Literacy (TOPEL; formerly PreCTOPP)	Early Reading First	Profile
		Even Start Classroom Literacy	
	Test of Silent Contextual Reading Fluency (TOSCRF)	Reading Comprehension	Profile
	Test of Silent Word Reading Fluency (TOSWRF)	Reading First	Profile
	Test of Word Reading Efficiency (TOWRE)	Reading and Mathematics Software Products	Profile
		Closing the Reading Gap	
	Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU)	Closing the Reading Gap	Profile
Vocabulary, Communication	Expressive One-Word Picture Vocabulary Test, Third Edition (EOWPVT)	Early Reading First	Profile
	Expressive Vocabulary Test, Second Edition (EVT-2)	Accelerating Language Development (REL-Southeast)	Profile
	Lexical diversity	Accelerating Language Development (REL-Southeast)	Table
	MacArthur-Bates Communicative Development Inventories (CDI)	Program for Infant/Toddler Care (REL-West)	Profile
	Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)	Opening the World of Learning (REL-Appalachia)	Profile
		Accelerating Language Development (REL-Southeast)	
		Program for Infant/Toddler Care (REL-West)	
		Even Start Classroom Literacy	
	Preschool Individual Growth and Developmental Indicators (IGDI)	Even Start Classroom Literacy	Profile
	Preschool Language Scale, Fourth Edition (PLS-4)	Early Reading First	Profile
	Test of Language Development—Primary, Fourth Edition (TOLD-4)	Even Start Classroom Literacy	Profile
Language Proficiency	IDEA Oral Language Proficiency Test (IPT I-Oral English)	English Language Learner Training and Materials (REL-Central)	Profile
	IDEA Oral Language Proficiency Test, Third Edition (IPT I-Oral Spanish)	Project ELLA	Profile

TABLE III.1 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
	PreLAS 2000	Early Reading First	Profile
Mathematics	Algebra End-of-Course Assessment	Reading and Mathematics Software Products	Profile
	Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS-K) Mathematics Assessment	Math Curricula	Profile
	Test of Early Mathematics Ability, Third Edition (TEMA-3)	Small Group Mathematics (REL-Southwest)	Profile
Science	Assessing Teacher Learning about Science Teaching (ATLAST) test of force and motion	Science Professional Development (REL-West)	Profile
Social Studies	Student performance assessment tasks (UCLA/CRESST)	Problem-Based Economics (REL-West)	Table
	Test of Economic Literacy, Third Edition (TEL-3)	Problem-Based Economics (REL-West)	Profile
Approaches Toward Learning, Motivation	Motivation for Reading Questionnaire (MRQ)	Thinking Reader (REL-Northeast & Islands)	Profile
	Patterns of Adaptive Learning Scales (PALS)	Classroom Assessment for Student Learning (REL-Central)	Profile
	The Research Assessment Package for Schools – Student Self Report (RAPS-S)	Classroom Assessment for Student Learning (REL-Central)	Profile
	Self- and task-perception questionnaire	Connected Mathematics 2 (REL-Mid-Atlantic)	Profile
	Student questionnaire of economic interest and attitudes	Problem-Based Economics (REL-West)	Table
	Student Time-on-Task and Engagement with Print (STEP)	Reading First	Table
Social-Emotional Well-Being	Social Competence and Behavior Evaluation, Preschool Edition (SCBE)	Early Reading First	Profile
	Social Skills Rating System (SSRS)	Lessons in Character Education (REL-West)	Profile
	Student questionnaire of behaviors and violence	School-Based Violence Prevention	Table
Other/Multidomain	Character traits and behavior questionnaire	Lessons in Character Education (REL-West)	Table
	Student questionnaire of reading behavior and attitudes	Enhanced Reading Opportunities	Table

TABLE III.1 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
	Student questionnaire of substance use	Mandatory Random Student Drug Testing	Table
	Student questionnaire on behavior and school	Student Mentoring Program	Table

Note: The Compendium includes these student achievement/development measures because of their use as an outcome in a recent NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental designs.

[a] Studies used either the most current or a previous version of a measure based on the timing of data collection. The Compendium reviewed the most recently published version of a measure as of November 2008. Study names are short forms (see Appendix F for a cross-walk to full study names).

[b] Format refers to how the Compendium presents the available information in Volume II. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information.

TABLE III.2

TEACHER KNOWLEDGE MEASURES INCLUDED IN THE COMPENDIUM

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Reading Knowledge (content and/or pedagogical)	Reading Content and Practices Survey (RCPS)	Professional Development Interventions on Early Reading	Table
	Teacher impact questionnaire of ELL instructional pedagogy	Principles-Based Professional Development (REL-Pacific)	Table
Mathematics Knowledge (content and/or pedagogical)	Pedagogical Content Knowledge Assessment (PCK)	Math Curricula	Profile
	Teacher Knowledge Inventory (TKI)	Professional Development Strategies in Math	Table
Science Knowledge (content and/or pedagogical)	Assessing Teacher Learning about Science Teaching (ATLAST) test of force and motion	Science Professional Development (REL-West)	Profile
Social Studies Knowledge	Test of Economic Literacy, Third Edition (TEL-3)	Problem-Based Economics (REL-West)	Profile
Pedagogical Knowledge	Test of assessment knowledge	Classroom Assessment for Student Learning (REL-Central)	Table
Multidomain Pedagogical Content Knowledge	Diagnostic Classroom Observation Tool (DCO; formerly VCOT)	Different Routes to Certification	Profile
		Comprehensive Teacher Induction	
	Reformed Teaching Observation Protocol (RTOP)	Alabama Math, Science, and Technology Initiative (REL-Southeast)	Profile

Note: The Compendium includes these student achievement/development measures because of their use as an outcome in a recent NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental designs.

[a] Studies used either the most current or a previous version of a measure based on the timing of data collection. The Compendium reviewed the most recently published version of a measure as of November 2008. Study names are short forms (see Appendix F for a cross-walk to full study names).

[b] Format refers to how the Compendium presents the available information in Volume II. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information.

TABLE III.3

CLASSROOM PRACTICES AND SETTINGS MEASURES
INCLUDED IN THE COMPENDIUM

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Comprehensive Classroom Practices	Authentic Instructional Practices Classroom Observation Form	Alabama Math, Science, and Technology Initiative (REL-Southeast)	Profile
	Center for the Improvement of Early Reading Achievement (CIERA) classroom observation scheme for classroom literacy instruction	Thinking Reader (REL-Northeast & Islands)	Profile
		Formative Assessment (REL-Midwest)	
	Classroom Characteristics (CC) form	Math Curricula	Table
	Diagnostic Classroom Observation Tool (DCO, formerly VCOT)	Different Routes to Certification	Profile
	Early Childhood Environment Rating Scale-Revised Edition (ECERS-R)	Early Reading First	Profile
	Early Reading Professional Development (PD) Classroom Observation	Professional Development Interventions on Early Reading	Profile
	Infant/Toddler Environment Rating Scale, Revised Edition (ITERS-R)	Program for Infant and Toddler Caregivers (REL-West)	Profile
	Reformed Teaching Observation Protocol (RTOP)	Alabama Math, Science, and Technology Initiative (REL-Southeast)	Profile
	School Observation Measure (SOM)	Hybrid Algebra I (REL-Appalachia)	Profile
	Sheltered Instruction Observation Protocol (SIOP)	Quality Teaching for English Learners (REL-West)	Profile
		Principles-Based Professional Development (REL-Pacific)	
	Teacher questionnaire of attitudes and behaviors	Formative Assessment (REL-Midwest)	Table
	Teacher questionnaire of classroom quality and instructional practices	Different Routes to Certification	Table
	Teacher questionnaire of educational practices	Effects of Success in Sight (REL-Central)	Table

Table III.3 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Reading Practices	Classroom observations of instructional quality	Adolescent Literacy Across the Curriculum (REL-Midwest)	Table
	Classroom observation of literacy teaching practices	Accelerating language Development (REL-Southeast)	Table
	Early Language & Literacy Classroom Observation (ELLCO) Pre-K and K-3 Tools	Opening the World of Learning (REL-Appalachia)	Profile
	Expository Reading Comprehension Classroom Observation (ERCCO)	Reading Comprehension	Table
		Collaborative Strategic Reading (REL-Southwest)	
	Instructional Practice in Reading Inventory (IPRI)	Reading First	Table
	Lexical diversity	Accelerating language Development (REL-Southeast)	Table
	Literacy Observation Tools (E-LOT, LOT, and A-LOT)	Opening the World of Learning (REL-Appalachia)	Profile
	Observation Measure of Language and Literacy Instruction (OMLIT)	Even Start Classroom Literacy	Profile
	Teacher Behavior Rating Scale (TBRS)	Early Reading First	Profile
	Teacher Interaction and Language Rating Scale	Accelerating language Development (REL-Southeast)	Profile
	Teacher questionnaire on reading instructional practices	Reading First	Table
Mathematics Practices	Algebra I Quality Assessment (AQA)	Hybrid Algebra I (REL-Appalachia)	Table
	Algebra I teacher questionnaire	Hybrid Algebra I (REL-Appalachia)	Table
	Classroom observation of math practices	Professional Development Strategies in Math	Table
	Observation of Math Instruction (OMI) form	Math Curricula	Table
School Engagement or Climate	Student and parent questionnaires of school climate	DC Opportunity Scholarship	Table
	Student questionnaire of behaviors and violence	School-Based Violence Prevention	Table
	Teacher questionnaire of school climate	Lessons in Character Education (REL-West)	Table
	Teacher questionnaire on safety and victimization	School-Based Violence Prevention	Table

Table III.3 (*continued*)

Category	Measure	NCEE or REL Study Use[a]	Format[b]
Other/Multidomain	Caregiver Interaction Scale (CIS)	Program for Infant and Toddler Caregivers (REL-West)	Profile
	Student questionnaire of economic interests and attitudes	Problem-Based Economics (REL-West)	Table
	Teacher questionnaire of instructional practices	Alabama Math, Science, and Technology Initiative-AMSTI (REL-Southeast)	Table
	Teacher questionnaire of instructional practices and self-efficacy	Principles-Based Professional Development (REL-Pacific)	Table
	Teacher questionnaire of practices and economic attitudes	Problem-Based Economics (REL-West)	Table

Note: The Compendium includes these student achievement/development measures because of their use as an outcome in a recent NCEE or REL study evaluating an educational intervention using randomized controlled trials or quasi-experimental designs.

[a] Studies used either the most current or a previous version of a measure based on the timing of data collection. The Compendium reviewed the most recently published version of a measure as of November 2008. Study names are short forms (see Appendix F for a cross-walk to full study names).

[b] Format refers to how the Compendium presents the available information in Volume II. A profile includes an overview and narrative. The table format summarizes important information about measures recently developed, generally those for studies with less technical information.

Information Sources

Profiles of selected measures summarize information obtained from several types of sources.[1] In most cases, the primary information source was the commercial publisher's information. If the instrument was developed for a particular study and/or not commercially available, the Mathematica team used the relevant journal article, book chapter, or NCEE or REL study report describing the measure and its psychometric properties. The team reviewed existing early childhood measures compendia (Berry et al. 2004; Kisker et al. 2003; Holland and Vick 2007) as well as critiques of measures such as the Buros Institute's *Mental Measurements Yearbook* (Geisinger et al. 2007) and incorporated relevant comments from them. To keep the measures review manageable, the Mathematica team did not perform a complete review of the available literature on a particular measure or its use. Rather, the team used a consistent set of resources for all measures. Each source used in developing profiles and the summary table is described below. The specific sources consulted for each profile are referenced in the profile.

- **Commercial publisher information** included both instrument materials and the publisher's website. Administration manuals typically outline the measure and its uses, administration instructions, scoring instructions, and interpretation and implications of results.[2] The accompanying technical manuals sometimes described the norming or development samples as well as research that provides evidence of reliability and validity. The Mathematica team obtained copies of forms or materials for administration in order to clarify assessment procedures, content, and the number of items, and consulted publisher websites for updated information on costs and contact information. At times, websites and manuals differed in the number of students included in the standardization process and its results. In such cases, the Compendium references the information in the printed manual rather than the information on the publisher's website, but provides publisher website URLs for the interested reader. If information on the website was unclear or absent (for example, qualifications needed to obtain the instrument), the Mathematica team made direct calls to the author/publisher.

- **Journal articles and books** were the main information sources for some measures, particularly those developed by academic researchers. Such measures are often published for a particular scholarly field in peer-reviewed journals and books that describe the measure, the development and theoretical support of its content, and the research sample, and sometimes provide factor analysis (exploratory or confirmatory) to support scale structure, along with other evidence of reliability and validity. These sources often include copies of the forms or items.

[1] To be included in the Compendium, information sources for a measure had to be available to the Mathematica team by mid-November 2008. For some measures, especially existing noncommercial tools and recently developed measures, it was difficult to obtain information because of problems such as data not being available from ongoing analyses, publication delays, or lack of formal documentation.

[2] This Compendium uses the terms *administration manual* or *technical manual* to encompass all such supporting information.

- **NCEE or REL study reports** provided published information on measures for several large-scale efforts. The reports discuss the measure, reliability information, and some validity evidence for both existing measures and study-developed measures. For the latter, if the report did not provide needed materials, the Mathematica team requested copies of forms or background documentation to review the breadth of content coverage and the number of items. Personal communications provided additional information in cases with limited documentation. When necessary, team members corresponded directly with a researcher or developer to obtain information about a measure, including evidence of reliability and validity, training requirements, and costs.

- **Existing measures compendia** in the early childhood area (generally, birth through age 8) provided useful information. When information was available about a measure under review, the Mathematica team turned to the existing compendia. For example, Mathematica previously compiled a compendium for the Administration for Children and Families to support practitioners and researchers in selecting and using measures appropriate for assessing infant and toddler child outcomes, parent/family well-being, and program environment quality (Kisker et al. 2003). The team adapted that profile template for the current project. In addition, Child Trends produced two compendia that provided a useful resource—one on early childhood measures of children's development and knowledge in a variety of domains (Berry et al. 2004) and a second on early childhood settings focused on classroom observations (typically for infant, toddler, and preschool settings) (Halle and Vick 2007).

- **The Buros Institute*'s Mental Measurements Yearbook*** (MMY) provided information compiled from independent reviews by leading academics, practitioners, or psychometricians of measures, indicating their major strengths and weaknesses (Geisinger et al. 2007). The yearbooks confirmed or augmented information collected from a thorough review of the manuals available from publishers or authors.

Review Format

Measure reviews took one of two formats: (1) a multipage profile or (2) a summary table. Profiles contain a description of administration, scoring, and evidence of reliability and validity for existing tools and for study-developed measures with sufficient psychometric information (that is, those with information beyond a single reliability estimate). A profile contains two components—a one-page summary and a multipage narrative (described in Volume II, Appendix A).

The summary tables provide brief descriptions of recently developed measures for which little or no information on reliability or validity was available. Twenty-two NCEE or REL studies developed such measures to focus on particular student, teacher, or classroom outcomes when available tools did not meet study needs. The Compendium includes these measures so that future studies may build on this current work. The summary table presents brief descriptions of the measures and includes information on domain, type of assessment, and grade/age range, as well as the availability of evidence of reliability and validity (see Exhibit III.1).

EXHIBIT III.1

TEMPLATE FOR COMPENDIUM PROFILE SUMMARY PAGE

Authors:		***Type of Assessment:*** ***Domain:***
Publisher:		***Grade/Age Range:*** **Administration Interval:**
Material, Training, and Scoring Costs:		**Personnel and Training Requirements** **Credentials Required for Use:** **Personnel for Administration:** **Training for Administration:**
Languages:		**Alternate Forms:**
Representativeness of Norming Sample:		**Summary** ***Initial Material Cost:*** **Time to Administer:** ***Ease of Administration and Scoring:*** ***Reliability:*** ***Predictive Validity:*** ***Construct/Concurrent Validity:*** ***Norming Sample Characteristics:***

Note: The italicized headings are included in the Chapter IV summary tables and are described here in Chapter III.

CONTENTS

This section provides an overview of eight key elements for the cross-measure summary presented in Chapter IV, derived from the profiles and summary tables: (1) domain, (2) grade/age range, (3) type of assessment, (4) initial material cost, (5) reliability, (6) validity, (7) norming sample, and (8) ease of administration and scoring. As shown in Exhibit III.1, all eight elements are covered in the profile summary sheet, with validity split across two separate categories for predictive and concurrent validity evidence. The summary tables include information on domain, grade/age range, reliability, and validity specifically, and a description section to summarize the other elements when that information was available. Below we describe each of the eight elements, presenting them in the order found in the Chapter IV tables that immediately follow this chapter. Full technical information for all profile components is included in Volume II, Appendix A. For further clarification on terms, see the glossary (Volume II, Appendix E).

Domain

The content covered by the measures fell into several domains applicable to student achievement or development, teacher knowledge, and classroom practices and settings. Table III.4 provides descriptions of common domains used to group measures. As part of the cross-measure synthesis in Chapter IV, some domains were either further specified or collapsed according to frequency. For example, the "teacher subject content knowledge" domain contains subject-specific labels for areas such as mathematics, reading, science, and social studies. The "other" domain for student achievement/development reflects the label for cross-measure synthesis. Similarly, for classroom practices and settings, the "social context" domain includes school climate, school engagement, and motivation for teaching. Volume II, Table A.1 provides a full list of domains and descriptions used to summarize and present the collected measures information.

Grade/Age Range

The grade and/or age range for which a measure is appropriate guides measure selection. Although one measure may be able to assess a broad range of ages and grades, others may have several forms for assessing various ages and grades.

Type of Assessment

Measures typically in the compendium are one of three types: (1) direct assessment, in which the student or teacher completes a series of items administered individually or to a group; (2) observation, in which a trained individual observes the teacher or other aspects of the classroom and rates or scores the behaviors of interest; and (3) report or ratings of one's own or another's behavior or knowledge.

TABLE III.4

DESCRIPTION OF THE STUDENT, TEACHER, AND CLASSROOM DOMAINS ASSESSED BY MEASURES IN THE COMPENDIUM

Domain	Description
Student Achievement/Development	
Reading	Early literacy skills—such as phonological awareness, letter recognition and naming, and print concepts—as well as reading vocabulary, decoding, phonics, reading fluency, and various comprehension skills
Language arts/language proficiency	Assessment of expressive and receptive language skills (oral and written); areas such as writing, oral skills, editing skills, grammar, syntax, vocabulary, and morphology may be included in these assessments along with English- or Spanish-language proficiency
Mathematics	Skills and topics related to counting, calculation, problem solving, geometry, and algebra
Science	Fields of earth, space, physical, and life sciences
Social studies	Topics across a range of disciplines, such as history, culture, geography, and economics
Approaches to learning/motivation	Executive functioning, attention, cognitive flexibility, curiosity, and engagement in learning
Social-emotional	Social skills, emotional well-being, and problem behaviors
Other	Areas such as general knowledge, motor development, and substance use
Teacher Knowledge	
Subject content knowledge	Knowledge about the content, topics, constructs, and procedures for a specific subject, noting the content subject area
Pedagogical content knowledge	Knowledge about how to teach the content of a particular subject or domain of learning, noting the particular content subject area
Other	Includes pedagogical knowledge about how to teach in general or about assessment
Classroom Practices and Settings	
Classroom quality	Aspects and quality level of physical, social, and temporal environments to include effective classroom management, use of routines, time use, interactions, materials, and space; the profile notes in parentheses if a particular measure assesses teacher-student interactions or the classroom environment. Teacher-student interactions include areas such as positive support, warmth, negative interactions, and punitiveness; environment rates the materials and space available for learning.
Instructional practices	The practices, activities, and strategies employed in the teaching of students, including both teacher-initiated instructional practices and feedback and noting whether the measure is comprehensive or subject-specific Comprehensive measures cover all subject areas when giving a rating; subject-specific refers to instructional practices that address a particular domain of learning.
Social context	Includes school climate (for example, sense of safety or positive regard for members of the community), school engagement, and motivation for teaching (enthusiasm, self-efficacy)

Initial Material Cost

Financial resources available for studies may vary. To aid readers, the Mathematica team assigned a rating based on the cost of an assessment kit or, in the absence of a kit, the cost of forms and manuals and any easily identifiable training or scoring costs as required by the authors or publishers: 1 = less than $100; 2 = $100 to $200; 3 = $201 to $500; and 4 = more than $500. Some measures do not present a fixed price for materials because the cost may depend on particular study parameters (for example, number of data collection waves or number of students/teachers to be assessed). In these instances, the Mathematica team rated cost as "to be determined based on negotiations with the publisher," or TBD. Because costs change, these estimates are meant only to give a broad sense of cost variation; those contemplating use of a particular measure should consult the publisher for exact costs to meet study-specific needs.

Reliability

Reliability indicates the consistency and stability of a measure. Other things being equal, the higher the reliability, the lower the measurement error and the better the measure is for the purposes for which it was designed. For direct assessments of knowledge and reports of behavior, the reliability category is based on internal consistency estimates that reflect the extent to which items in the measure capture the same construct. For observation tools, the reliability category is based on inter-rater reliability estimates of consistency that provide evidence that the tool captures the same information across observers or raters. When selecting measures, the level of reliability required depends largely on how confident one needs to be about the decision being made; greater confidence requires higher reliability. The Mathematica team chose the threshold of 0.70 for internal consistency reliability, following the prevalent rule of thumb in the field used by researchers and assessment developers (Bacon 2004; Cohen 2007; Litwin 2003; Nunnally 2008).[1] Therefore, reliability ratings for the Chapter IV cross-measure synthesis are reported as 1 (none described), 2 (all or most under 0.70), or 3 (all at or above 0.70). However, it is important to note that higher estimates are desirable; measurement experts recommend 0.80 when the measure will be used for group-level decisions and .90 when the results of the measure will be used for high-stakes decisions affecting individuals (for example, special education placement) (Nunnally and Bernstein 1994).

Validity

Indicators of validity help determine whether a measure assesses what it is supposed to measure for the intended purpose. The Mathematica team categorized validity as "available" or "not available," indicating whether information exists about the relationship between the measure and another measure or criterion administered at the same time (concurrent) or later (predictive). The profiles provide the concurrent and predictive validity information separately, but for ease of presentation in the Chapter IV summary tables we indicate only that one of these

[1] Kappas are generally used to compute inter-rater reliability for categorical measures; intra-class correlations are used for continuous measures (Cohen 1960; Shrout and Fleiss 1979).

two types of validity is available. The interested reader may then review the available information in light of his or her current study's purpose and context.

Norming Sample

Knowing whether a measure has norms and, if so, whether the norming sample was nationally representative or representative of the students or teachers under study is an important consideration when you want to be able to draw conclusions about the progress or status of students relative to students in the nation. The Compendium rates the norming sample as 1 (none described), 2 (older than 10 years or not nationally representative), or 3 (normed within past 10 years and nationally representative). The Mathematica team selected a criterion of 10 years because many publishers or developers re-norm or standardize their measure that often to ensure that it is representative and current.

Rather than compare a student's test results to those of a reference group (norming sample), criterion-referenced tests (CRT) judge student performance in terms of how well students have learned or mastered a body of knowledge. On standardized CRTs such as those administered to large groups of students, a passing or cut score is set by a committee of experts. Some published instruments provide a criterion-referenced interpretation of results (for example the percentage of students who are able to identify the vowel sounds in words) instead of, or in addition to, a norm-referenced interpretation.

Ease of Administration and Scoring

Measures differ in the knowledge, skills, and training needed by the person administering the assessment or performing the observation, with requirements ranging from clerical skills to specialized training. The Mathematica team included an estimate of the level of training needed to learn about, conduct, and score the instrument: 1 (not described), 2 (self-administered or administered and scored by someone with basic clerical skills), 3 (administered and scored by a highly trained individual), or 4 (administered or scored by a clinician, specialist, or the publisher). Some measures may have requirements that stipulate different levels of training for administration and scoring; in such cases, the higher rating applies.

SUMMARY

Together, the features of measures described above represent important areas for consideration when reviewing a measure for use in a given study. The information collection method employed as part of the Compendium's development called for the use of a consistent set of sources to locate information and a standard format for summarizing the information across measures. The current chapter presents an overview of the collected information, while Volume II, Appendix A, fully describes the content of the profiles and summary tables of recently developed measures. Volume II also contains separate appendixes of the individual profiles and tables for measures of student achievement/development, teacher knowledge, and classroom practices and settings. Each profile also contains a link to the NCEE or REL study that used the measure.

IV. CHARACTERISTICS OF THE COMPENDIUM MEASURES ACROSS DOMAINS

This chapter summarizes the measures included in the Compendium. Three sections with accompanying tables support at-a-glance comparisons of how the measures vary across important dimensions. The sections group measures of (1) student outcomes, (2) teacher knowledge, and (3) classroom practices and settings. The accompanying tables present summary information about the domain of assessment (some measures assess more than one domain), grade/age range of the students with whom the measure has been used and for whom it is appropriate, assessment type (for example, adaptive, direct, self-report), purchase cost of the assessment and supporting materials, reliability estimates, availability of validity evidence, characteristics of the norming sample, and ease of administration and scoring.

SUMMARY OF STUDENT OUTCOME MEASURES

Profiles of 43 measures of student outcomes, along with descriptions of nine other recently developed measures, are provided in Table IV.1 and Volume II, Appendix B. For the newer measures, only limited psychometric data are currently available. The range of outcomes and the domains assessed by the 52 measures vary according to the research questions explored in studies funded in recent years. Indicative of the relative emphasis of NCEE-funded evaluations on literacy interventions, 24 of the student outcome measures in the Compendium include an assessment of achievement in reading, and 20 assess language arts or language proficiency. Nine measures include an assessment of mathematics. Fewer measures include assessments of science and social studies (for each, N = 5). Eight measures assess some aspect of student motivation, approaches to learning, or executive functioning. Three measures with full profiles and three additional recently developed measures address the assessment of social and emotional development or behavior. Of the 52 measures, 39 assess a single domain and 13 assess multiple domains, such as reading and language arts.

Five of the Compendium measures focus on a specific grade level, while all other measures address a range of ages or grades. Eight measures assess students from preschool through adulthood, with an additional six measures targeting students in kindergarten or grade 1 through grade 12 and two others that are appropriate for grades 2 or 3 through grades 11 or 12. Six measures are designed for preschool only. Three measures assess preschool through grade 3, two measures assess preschool or kindergarten through grade 1, and one measure was designed for grades 1 through 3. The grade span for assessing the middle grades varies; one measure covers grades 3 through 6, one measure covers grades 3 through 8, one indicates "middle school students," two assess grades 4 through grade 8 or 9, two assess grades 4 through 6, and one assesses grades 4 and 5.

Sixteen of the 52 student outcome measures in the Compendium (30 percent) have nationally representative samples and have been normed within the past 10 years. Norming studies are expensive but provide important information. Researchers use norms to compare assessment results from the study sample of students with a national sample. In addition,

TABLE IV.1

KEY CHARACTERISTICS OF STUDENT ACHIEVEMENT/DEVELOPMENT MEASURES BY CATEGORY

	Domain														
Instrument Name	**R**	**LA/P**	**M**	**Sci**	**SS**	**AM**	**SE**	**Oth**	**Grade/Age Range**	**Assessment Type**	**Initial Material Cost[a]**	**Reliability[b]**	**Validity[c]**	**Norming Sample[a]**	**Ease of Administration and Scoring[a]**
Comprehensive Cognitive and Achievement Tests															
Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)		X					X	X	1–42 months	A-D, R	4	3	A	3	3
Kaufman Test of Educational Achievement, Comprehensive Form, Second Edition (KTEA-II)	X	X	X						4.5–25 years	A-D	3	3	A	3	3
Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) and Achievement Level Tests (ALT)	X	X	X	X					Grades 2–11	A-D	TBD	3	A	1	2
Stanford Achievement Test Series, Tenth Edition (Stanford-10)	X	X	X	X	X				K–Grade 12	D	2	3	A	3	3
TerraNova 3	X	X	X	X	X				K–Grade 12	D	3	3	A	3	4
Woodcock Johnson-III Normative Update (WJ III NU)	X	X	X	X	X	X		X	2 years–Adult	A-D	4	3	A	3	3
Literacy, Reading															
6+1 Trait Writing Scoring Guide (Rubrics)		X							Grades 3–12	D	1	3	A	1	3
AIMSweb Oral Reading Fluency	X								Grades 1–8	D	3	1[d]	A	1	3
Dynamic Indication of Basic Early Literacy Skills (DIBELS), Sixth Edition	X								K–Grade 6	D	1	1[d]	A	1	3

TABLE IV.1 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	R	LA/P	M	Sci	SS	AM	SE	Oth							
Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4)	X								K–Adult	D	2	3	A	3	3
Group Reading Assessment and Diagnostic Evaluation (GRADE)	X								Pre-K–Postsecondary	A-D	2	3	A	3	2
Indicadores Dinámicos del Éxito en la Lectura (IDEL), Seventh Edition	X								K–Grade 3	A-D	1	1[d]	A	1	3
Metacognitive Awareness of Reading Strategies Inventory (MARSI)	X								Grade 6–College	R	1	3	A	1	2
Phonological Awareness Literacy Screening (PALS) PreK, PALS-K, PALS 1–3	X	X							Pre-K–Grade 3	D	1	3	A	1	3
Science Reading Comprehension Assessment	X								Grade 5	D	TBD	3	A	2	4
Social Science Reading Comprehension Assessment	X								Grade 5	D	TBD	3	A	2	4
Stanford Diagnostic Reading Test, Fourth Edition (SDRT 4)	X								K–Grade 13 (first semester of college)	D	2	3	A	2–3	3
Test of Preschool Early Literacy (TOPEL; formerly PreCTOPP)	X	X							3–5 years	A-D	3	3	A	3	3
Test of Silent Contextual Reading Fluency (TOSCRF)	X								7–18 years	D	3	1[d]	A	2	3
Test of Silent Word Reading Fluency (TOSWRF)	X								6–17 years	D	2	1[d]	A	3	3

TABLE IV.1 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	R	LA/P	M	Sci	SS	AM	SE	Oth							
Test of Word Reading Efficiency (TOWRE)	X								6–24 years	D	2	1[d]	A	2	3
Woodcock Reading Mastery Test-Revised/Normative Update (WRMT-R/NU)	X								5–75+ years	A-D	3	3	A	2	3
Vocabulary, Communication															
Expressive One-Word Picture Vocabulary Test—Third Edition (EOWPVT)		X							2–18 years	A-D	2	3	A	3	3
Expressive Vocabulary Test, Second Edition (EVT-2)		X							2.5–90 years	A-D	3	3	A	3	3
Lexical diversity (Accelerating Language Development REL-Southeast)[e]		X							K–Grade 1	D	n.a.	1	NA	n.a.	n.a.
MacArthur-Bates Communicative Development Inventories (CDI)		X							8–37 months	R	2	3	A	2	3
Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)		X							2–90 years	A-D	3	3	A	3	3
Preschool Individual Growth and Developmental Indicators (IGDI)	X	X							3–5 years	D	1	2	A	1	2
Preschool Language Scale, Fourth Edition (PLS-4)	X	X							0–6 years	A-D	3	3	A	3	3
Test of Language Development-Primary, Fourth Edition (TOLD-4)		X							4–8 years	A-D	3	3	A	3	3

TABLE IV.1 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	R	LA/P	M	Sci	SS	AM	SE	Oth							
Language Proficiency															
IDEA Oral Language Proficiency Test (IPT I-Oral English)		X							K–Grade 6	A-D	2	3	A	2	2
IDEA Oral Language Proficiency Test, Third Edition (IPT I-Oral Spanish)		X							K–Grade 6	A-D	2	3	A	2	2
PreLAS 2000	X	X	X					X	Pre-K–Grade 1	D	3	3	A	2	3
Mathematics															
Algebra End-of-Course Assessment			X						Grades 6–12	D	TBD	3	NA	1	2
Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS-K) Mathematics Assessment			X						K–Grade 5	A-D	NA	3	A	3	4
Test of Early Mathematics Ability, Third Edition (TEMA-3)			X						3–8 years	A-D	3	3	A	3	3
Science															
Assessing Teacher Learning about Science Teaching (ATLAST) Test of Force and Motion				X					Middle school students	D	1	3	NA	1	4
Social Studies															
Student performance assessment tasks (UCLA/CRESST; Problem-Based Economics REL-West)[e]					X				Grade 12	D	n.a.	1	NA	n.a.	n.a.

TABLE IV.1 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	R	LA/P	M	Sci	SS	AM	SE	Oth							
Test of Economic Literacy, Third Edition (TEL-3)					X				Grades 9–12	D	1	3	A	2	2
Approaches Toward Learning, Motivation															
Motivation for Reading Questionnaire (MRQ)						X			Grades 3–6	R	1	2	A	1	1
Patterns of Adaptive Learning Scales (PALS)						X			Grades 4–9	R	1	3	A	1	3
The Research Assessment Package for Schools—Student Self-Report (RAPS-S)						X			Grades 3–8	R	1	3	A	1	3
Self- and Task-Perception Questionnaire						X			Grades 5–12	R	1	3	A	1	1
Student questionnaire of economic interest and attitudes (Problem-Based Economics REL-West)[e]						X			Grade 12	R	n.a.	3	NA	n.a.	n.a.
Student Time-On-Task and Engagement with Print (STEP; Reading First)[e]						X			Grades 1–3	O	n.a.	3	NA	n.a.	n.a.
Social-Emotional Well-Being															
Social Competence and Behavior Evaluation, Preschool Edition (SCBE)							X		30–78 months	R	1	3	A	2	2
Social Skills Rating System (SSRS)							X		3–18 years	R	2	2–3	A	2	2

TABLE IV.1 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	R	LA/P	M	Sci	SS	AM	SE	Oth							
Student questionnaire of behaviors and violence (School-Based Violence Prevention)[e]							X		Grades 6–8	R	n.a.	3	NA	n.a.	n.a.
Other/Multidomain															
Character traits and behavior questionnaire (Lessons in Character Education REL-West)[e]							X		Grades 4–5	R	n.a.	3	NA	n.a.	n.a.
Student questionnaire of reading behavior and attitudes (Enhanced Reading Opportunities)[e]	X							X	Grade 9	R	n.a.	3	NA	n.a.	n.a.
Student questionnaire of substance use (Mandatory Random Student Drug Testing)[e]								X	Grades 9–12	R	n.a.	1	NA	n.a.	n.a.
Student questionnaire on behavior and school (Student Mentoring Program)[e]						X	X		Grades 4–8	R	n.a.	3	NA	n.a.	n.a.

Source: The information included in this table was drawn from the manuals or other resources available from the authors and publishers of the measures. Individual users may have different experiences with locating information.

[a] n.a. for *not applicable* refers to recently developed measures summarized in a table (Volume II, Table B.1). The measures are so recent that at the time this report was prepared, published information was not yet available or the study report had little to no psychometric data such that information on cost, norming samples, or administration was limited and/or not meaningful for evaluating such a measure. Thus, the information is not applicable for these subsets of measures.

[b] Ratings refer to total test scores or scores commonly reported; individual subscales may differ, as noted in the specific profile. In addition, ratings may reflect availability based on a previous version, as noted in the specific profile.

[c] Validity ratings reflect the availability of predictive or construct/concurrent validity. Ratings may also reflect availability based on a previous version, as noted in the specific profile.

TABLE IV.1 (*continued*)

[d] For the reading fluency measures, developers typically do not conduct estimates of internal consistency reliability. Some developers note such an estimate is not appropriate for timed fluency measures; other researchers indicate the potential to do so. The calculation can be more complex or less feasible due to the level of missing data in that students vary in how far (number of lines, number of passages) they get and not every student would receive the same "items." Some researchers may then suggest alternate form or test-retest reliability estimates as indicators of internal consistency; however, the possibility of practice effects exists. The current Compendium selected internal consistency reliability for determining the summary reliability rating, and thus, the reading fluency measures received a rating based on the presence (or absence) of that information.

[e] These recently developed study measures, for which limited psychometric information is available, may be found in Volume II, Table B.1.

KEY

Domain
R = Reading
LA/P = Language Arts/Language Proficiency
M = Mathematics
Sci = Science
SS = Social Studies
AM = Approaches to Learning/Motivation
SE = Social-emotional
Oth = Other

Assessment Type
A-D =Adaptively administered direct assessment
D = Direct Assessment
O = Observation
R = Parent/Teacher/Student Report

Initial Material Cost
n.a. = Not applicable
TBD = To be determined upon negotiation with the publisher
1 = under $100
2 = $100 to $200
3 = $201 to $500
4 = more than $500

Reliability
1 = None described
2 = All or mostly under 0.70
3 = All at or above 0.70.

Validity
NA = Not Available
A = Available

Norming Sample
n.a. = Not applicable
1 = None described
2 = Older than 10 years or not nationally representative
3 = Normed within past 10 years and nationally representative

Ease of Administration and Scoring
n.a. = Not applicable
1 = Not described
2 = Self-administered or administered and scored by someone with basic clerical skills
3 = Administered and scored by a highly trained individual
4 = Administered or scored by a clinician or specialist or the publisher

publishers often use the norming sample or portions of it to conduct substudies related to special populations (for example, students with developmental disabilities or delays, English language learners, and younger children) and to establish concurrent validity by assessing some children with two different measures in the same domain.[1] A review of the profiles suggests that the age of the norms is a particular challenge in efforts to locate an available measure with current norms in both Spanish and English. It is also important to note that some measures without current norms are undergoing new standardization as reported in the profiles. For example, the Social Skills Rating System was recently revised and restandardized (and renamed the Social Skills Improvement System [Gresham and Elliott 2008]).

The type of assessment ranges from adaptively administered direct student assessments to observations and ratings by parents or teachers, or student self-reports. Most of the measures of student outcomes (N = 37) are direct assessments that typically address cognitive (for example, Bayley-III and Woodcock Johnson-III) or academic content, such as reading, math, or social studies. The adaptive tests in the Compendium are individually administered and often use floor and ceiling rules to identify items of appropriate difficulty for measurement. One measure—Northwest Evaluation Association (NWEA) Measures of Academic Progress and Achievement Level Tests (NWEA undated)—is a fully adaptive computer-administered measure. Adaptive measures involve fewer problems with floor and ceiling effects and thus are able to provide reliable estimates for students of varying ability levels, including those at the ends of the age/grade distribution.

The measures of social and emotional development, behavior, motivation, and attitudes typically use parent, teacher, or self-report measures. Twelve of the measures include a parent, teacher, or self-rating/report. One measure of student engagement is a direct observation of student behavior. Overall, direct assessments and student observations are more costly to administer than parent, teacher, or student self-reports. In general, direct assessment is preferred in domains that include knowledge and academic skills. Researchers use other assessment types when outcomes target student behavior and attitudes, and may obtain information from several informants.

The initial cost of the measures of student outcomes varies from less than $100 to more than $500. The most expensive measures are published assessments that are designed and normed to measure multiple aspects of development or achievement across years. In general, commercially available measures of cognition and achievement that are based on large nationally representative normative samples are more expensive than most measures of social-emotional development (see Table IV.1 and Volume II, Appendix B). In addition, the authors/publishers of these more expensive measures typically provide considerable psychometric information, including convergent validity data and comparisons of results from children in different ability groups.

[1] Table IV.1 indicates "not applicable" (noted as n.a.) if a measure is so recently developed that information about it is not yet published. If a norming sample is not available or not described, the assessment earns a rating of 1. If a norming sample is older than 10 years or not nationally representative, it earns a rating of 2.

The majority of the measures of student outcomes have evidence of reliability at or above 0.70. Measures of reading fluency did not estimate internal consistency and so are rated as "none described" on the tables. Some researchers argue that internal consistency estimates are not appropriate for timed fluency tests because they do not have distinct items (Mather et al. 2004; Young 2005). All of the fluency measures provided evidence of test-retest or alternate form reliability. In addition to the fluency measures, three measures did not provide estimates of internal consistency and three measures had at least one subtest that fell below 0.70. Some evidence of validity was available for almost all the direct assessments (three direct assessments did not provide evidence of convergent validity), although only eight of the parent, teacher, or student report measures (53 percent) provided evidence of convergent validity.

The requirements for administration and scoring vary across assessments. More specifically, self-report measures completed by students or teachers require little oversight, whereas assessments of student achievement and cognition require those administering the measures to undergo training to ensure adherence to standard rules of administration. Additional training is needed for most adaptive assessments (other than computer-administered ones) to ensure administration of the correct items. For most of the student outcome measures included in the Compendium, scoring and interpretation are more complex than administration. Requirements for scoring the student outcome measures in the Compendium vary from a simple sum of the correct items to a complex conversion of a raw score into a standardized score using a computer scoring program that takes into account the student's date of birth, gender, and other demographic characteristics.[2] Similarly, interpretation of scores may require special skills or credentials.

SUMMARY OF TEACHER KNOWLEDGE MEASURES

Teacher knowledge is a more recent measurement focus in studies of curricular and educational interventions (Hill et al. 2008). Similar to examination of classroom quality and instructional practice, examination of teacher knowledge is often used as a mediator in exploring the effect of an intervention. In recent years there has been an increase in the assessment of teacher knowledge or practice as the initial outcome of an intervention. This approach views the role of the teacher as proximal to student outcomes, and the conduit through which some interventions affect student achievement (Hill et al. 2005; Sawadee et al. 2002).

The nine teacher knowledge assessments in the Compendium (Table IV.2 and Volume II, Appendix C) focus on content knowledge (N = 5) or pedagogical content knowledge (PCK; N = 6); four measures assess both. PCK was introduced to educational researchers in Shulman's

[2] For measures that require scoring by a publisher or a psychometrician, the Mathematica team assigned a rating of 4 in the category of ease of administration and scoring, even if the measure is self-administered. If norms tables are available for use in scoring a measure, the Mathematica team assigned a rating of 3. If the publisher offers a computer program that provides scores for the user, the team assigned a rating of 2. A rating of 2 may also indicate that no norms are available and the score is a sum of individual items.

TABLE IV.2

KEY CHARACTERISTICS OF TEACHER KNOWLEDGE MEASURES BY CATEGORY

	Domain														
Instrument Name	**MCK**	**R CK**	**Sci CK**	**SS CK**	**M PCK**	**R PCK**	**Sci PCK**	**Oth**	**Grade/Age Range**	**Assessment Type**	**Initial Material Cost [a]**	**Reliability [b]**	**Validity [c]**	**Norming Sample [a]**	**Ease of Administration and Scoring [a]**
Reading Knowledge (content and/or pedagogical)															
Reading Content and Practices Survey (RCPS; Professional Development Interventions on Early Reading)[d]		X				X			Grade 2	D	n.a.	2	NA	n.a.	n.a.
Teacher impact questionnaire of ELL instructional pedagogy (Principles-Based Professional Development REL-Pacific)[d]								X	Grades 4–5	D	n.a.	1	NA	n.a.	n.a.
Mathematics Knowledge (content and/or pedagogical)															
Pedagogical Content Knowledge Assessment (PCK)	X				X				K–Grade 6	D	TBD	3	A	1	3
Teacher Knowledge Inventory (TKI; Professional Development Strategies in Math)[d]	X				X				Grade 7	D	n.a.	1	NA	n.a.	n.a.
Science Knowledge (content and/or pedagogical)															
Assessing Teacher Learning about Science Teaching (ATLAST) Test of Force and Motion			X				X		Middle school teachers	D	1	3	NA	1	4
Social Studies Knowledge															
Test of Economic Literacy, Third Edition (TEL-3)[e]				X					Grades 9–12	D	1	1	NA	1	2
Pedagogical Knowledge															
Test of Assessment Knowledge (Classroom Assessment for Student Learning REL-Central)[d]								X	Grades 4–5	D	n.a.	1	NA	n.a.	n.a.
Multidomain Pedagogical Content Knowledge															
Diagnostic Classroom Observation Tool (DCO, formerly VCOT)					X	X	X		K–Grade 12	O	1	3	A	1	3

Table IV.2 (*continued*)

Instrument Name	Domain								Grade/Age Range	Assessment Type	Initial Material Cost [a]	Reliability [b]	Validity [c]	Norming Sample [a]	Ease of Administration and Scoring [a]
	MCK	R CK	Sci CK	SS CK	M PCK	R PCK	Sci PCK	Oth							
Reformed Teaching Observation Protocol (RTOP)					X		X		K– Graduate programs	O	1	3	A	1	3

Source: The information included in this table was drawn from the manuals or other resources available from the authors and publishers of the measures. Individual users may have different experiences with locating information.

[a]n.a. for not applicable refers to recently developed measures summarized in a table (Volume II, Table C.1). These measures are so recent that at the time this report was prepared, published information was not available or the study report had little to no psychometric data such that information on cost, norming samples, or administration was limited and/or not meaningful for evaluating such a measure. Thus, the information is not applicable for these subsets of measures.

[b]Ratings refer to total test scores or scores commonly reported; individual subscales may differ, as noted in the specific profile. In addition, ratings may reflect availability based on a previous version, as noted in the specific profile.

[c]Validity ratings reflect the availability of predictive or construct/concurrent validity. Ratings may also reflect availability based on a previous version, as noted in the specific profile.

[d]These measures may be found in Volume II, Table C.1 on recently developed study measures with limited availability of psychometric information.

[e]Reliability, validity, and norming were conducted with samples of students, not teachers, but the measure has been used with both groups in some studies.

KEY

Domain
MCK = Math Content Knowledge
R CK = Reading Content Knowledge
Sci CK = Science Content Knowledge
SS CK = Social Studies Content Knowledge
M PCK = Math Pedagogical Content Knowledge
R PCK = Reading Pedagogical Content Knowledge
Sci PCK = Science Pedagogical Content Knowledge
Oth = Other

Assessment Type
D = Direct Assessment
O = Observation
R = Parent/Teacher/Student Report

Initial Material Cost
n.a.= Not applicable
TBD = To be determined upon negotiation with the publisher
1 = under $100
2 = $100 to $200
3 = $201 to $500
4 = >$500

Reliability
1 = None described
2 = All or mostly under 0.70
3 = All at or above 0.70

Validity
NA = Not Available
A = Available

Norming Sample
n.a.= Not applicable
1 = None described
2 = Older than 10 years or not nationally representative
3= Normed within past 10 years and nationally representative

Ease of Administration and Scoring
n.a.= Not applicable
1 = Not described
2 = Self-administered or administered and scored by someone with basic clerical skills
3 = Administered and scored by a highly trained individual
4 = Administered or scored by a clinician or specialist or the publisher

48

seminal 1986 address to the American Educational Research Association. Shulman (1986) differentiates PCK from subject matter or content knowledge, stating that it "goes beyond knowledge of subject matter per se to the dimension of subject matter knowledge for teaching" (p. 9). PCK includes measurement of what teachers know about how students develop knowledge in a given domain, how teachers interpret student errors, optimal ways of representing information to students, and pedagogical practices specific to the curricular area. The Compendium includes measures of PCK used in the primary and middle grades in a variety of content areas, including reading, mathematics, and science.

The Compendium also includes two measures of teacher knowledge that are distinct in their focus from other measures of academic content knowledge or PCK. One measure assesses teacher knowledge of strategies for teaching English language learners, while the other addresses teacher knowledge of assessment. Both measures were developed to assess knowledge of teachers in the middle grades.

All of the teacher knowledge measures except the RTOP and VCOT are direct assessments in paper-and-pencil format that are easy to administer and are sometimes included in teacher surveys. However, the scoring of the measures is more complex. Most use item response theory (IRT; see Chapter II) and thus have a higher rating on ease of administration and scoring; this may also increase scoring costs if researchers are not familiar with psychometric software and analysis.

Given that measures of teacher content knowledge and PCK are relatively new (developed within the last decade), evidence of reliability and validity is limited. Only three of the nine measures report reliability estimates at or above 0.70 and provide some validity evidence. None of the measures describes a norming sample, though some information is available for the pilot samples of those measures.

SUMMARY OF CLASSROOM PRACTICES AND SETTINGS MEASURES

Research studies such as Raudenbush (2008) examine classroom practices and settings as more proximal indicators of the effects of educational interventions than student outcomes. Measurement examines overall quality, instructional practices, learning environment, and classroom climate. The focus of a given measure is sometimes content-specific and sometimes general or global. Assessment of classroom practices and settings is conducted in a variety of ways, including live observation with varying coding techniques; later coding of videotaped classrooms (allowing for coding in several ways with different instruments); and parent, student, and teacher reports. Observers may use rubrics with clear behavioral descriptions that anchor the points on the rubric, tallies of behaviors in a specified time sample, or frequency or quality ratings.

The Compendium includes profiles of 15 measures that address some aspect of the classroom (Table IV.3 and Volume II, Appendix D). The table of recently developed measures (Volume II, Appendix Table D.1) presents an additional 22 measures with limited psychometric data available, bringing the total number of classroom measures to 37 (Table IV.3). All but four address some aspect of instructional practices and/or classroom quality. The remaining four examine some aspect of the social context, such as school climate, student engagement,

TABLE IV.3

KEY CHARACTERISTICS OF CLASSROOM PRACTICES AND SETTINGS MEASURES

	Domain												
Instrument Name	**CQ**	**CI**	**RI**	**MI**	**Oth IP-C**	**Oth SC**	**Grade/Age Range**	**Assessment Type**	**Initial Material Cost[a]**	**Reliability[b]**	**Validity[c]**	**Norming Sample[a]**	**Ease of Administration and Scoring[a]**
Comprehensive Classroom Practices													
Authentic Instructional Practices Classroom Observation Form	X	X	X	X		X	K–Grade 12	O	1	3	A	1	3
CIERA classroom observation scheme for classroom literacy instruction	X		X			X	K–Grade 6	O	TBD	3	A	1	3
Classroom Characteristics (CC) form (Math Curricula)[d]	X					X	Grades 1–3	O	n.a.	1	NA	n.a.	n.a.
Diagnostic Classroom Observation Tool (DCO, formerly VCOT)	X		X	X	X		K–Grade 12	O	1	3	A	1	3
Early Childhood Environment Rating Scale-Revised Edition(ECERS-R)	X						2.5–5 years	O, R	1	3	A	1	3
Early Reading Professional Development (PD) Classroom Observation			X				Grade 2	O	TBD	3	A	1	3
Infant/Toddler Environment Rating Scale-Revised Edition (ITERS-R)	X						0–30 months	O, R	1	3	A	1	3
Reformed Teaching Observation Protocol (RTOP)				X	X		K–Graduate programs	O	1	3	A	1	3
School Observation Measure (SOM)	X	X				X	K–Grade 12	O	1	3	A	1	3
Sheltered Instruction Observation Protocol (SIOP)	X	X					K–Grade 12	O	1	3	A	1	3
Teacher questionnaire of attitudes and behaviors (Formative Assessment REL-Midwest)[d]		X					Grades 4–5	R	n.a.	1	NA	n.a.	n.a.
Teacher questionnaire of classroom quality and instructional practices (Different Routes to Certification)[d]	X		X	X		X	K–Grade 5	R	n.a.	1	NA	n.a.	n.a.

Table IV.3 (*continued)*

Instrument Name	Domain						Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	CQ	CI	RI	MI	Oth IP-C	Oth SC							
Teacher questionnaire of educational practices (Effects of Success in Sight REL Central)[d]		X				X	Grades 3–5	R	n.a.	1	NA	n.a.	n.a.
Reading Practices													
Classroom observations of instructional quality (Adolescent Literacy Across the Curriculum REL-Midwest)[d]			X				Grades 9–12	O	n.a.	1	NA	n.a.	n.a.
Classroom observation of literacy teaching practices (Accelerating Language Development REL-Southeast)[d]			X				K–Grade 1	O	n.a.	1	NA	n.a.	n.a.
Early Language & Literacy Classroom Observation (ELLCO) Pre-K and K-3 Tools	X		X				Pre-K–Grade 3	O	1	3	A	1	3
Expository Reading Comprehension Classroom Observation (ERCCO; Reading Comprehension; Collaborative Strategic Reading REL-Southwest)[d]			X				Grade 5	O	n.a.	3	NA	n.a.	n.a.
Instructional Practice in Reading Inventory (IPRI; Reading First)[d]			X				Grades 1–2	O	n.a.	3	NA	n.a.	n.a.
Lexical diversity (Accelerating Language Development REL-Southeast)[d]	X						K–Grade 1	O	n.a.	1	NA	n.a.	n.a.
Literacy Observation Tools (LOT; E-LOT, LOT, and A-LOT)	X		X			X	Pre-K–Grade 12	O	4	3	A	1	4
Observation Measure of Language and Literacy Instruction (OMLIT)	X		X				Early childhood classrooms	O	4	3	A	1	3
Teacher Behavior Rating Scale (TBRS)	X	X					Preschool	O	TBD	3	A	1	3
Teacher Interaction and Language Rating Scale			X				2–4 years	O	1	3	A	1	3

Table IV.3 (*continued)*

Instrument Name	Domain						Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	CQ	CI	RI	MI	Oth IP-C	Oth SC							
Teacher questionnaire on reading instructional strategies (Reading First)[d]			X				Grades 1–3	R	n.a.	3	NA	n.a.	n.a.
Mathematics Practices													
Algebra I Quality Assessment (AQA; Hybrid Algebra I REL-Appalachia)[d]				X			Grade 9	O	n.a.	1	NA	n.a.	n.a.
Algebra I teacher questionnaire (Hybrid Algebra I REL-Appalachia)[d]				X		X	Grade 9	R	n.a.	1	NA	n.a.	n.a.
Classroom observation of math practices (Professional Development Strategies in Math)[d]				X		X	Grade 7	O	n.a.	1	NA	n.a.	n.a.
Observation of Math Instruction (OMI) form (Math Curricula)[d]				X			Grades 1–3	O	n.a.	1	NA	n.a.	n.a.
School Engagement or Climate													
Student and parent questionnaires of school climate (DC Opportunity Scholarship)[d]						X	Grades 4–12	R	n.a.	3	NA	n.a.	n.a.
Student questionnaire of behaviors and violence (School-Based Violence Prevention)[d]						X	Grades 6–8	R	n.a.	1	NA	n.a.	n.a.
Teacher questionnaire of school climate (Lessons in Character Education REL-West)[d]						X	Grades 2–5	R	n.a.	3	NA	n.a.	n.a.
Teacher questionnaire on safety and victimization (School-Based Violence Prevention)[d]						X	Grades 6–8	R	n.a.	3	NA	n.a.	n.a.
Other/Multidomain													
Caregiver Interaction Scale (CIS)	X						Caregivers/teachers of preschool-age children	O	1	3	A	1	3

Table IV.3 (*continued)*

Instrument Name	Domain						Grade/Age Range	Assessment Type	Initial Material Cost[a]	Reliability[b]	Validity[c]	Norming Sample[a]	Ease of Administration and Scoring[a]
	CQ	CI	RI	MI	Oth IP-C	Oth SC							
Student questionnaire of economic interest and attitudes (Problem-Based Economics REL-West)[d]					X		Grade 12	R	n.a.	3	NA	n.a.	n.a.
Teacher questionnaire of instructional practices (Alabama Math, Science, and Technology Initiative-AMSTI REL-Southeast)[d]				X	X		Grades 4–8	R	n.a.	1	NA	n.a.	n.a.
Teacher questionnaire of instructional practices and self-efficacy (Principles-Based Professional Development REL-Pacific)[d]	X		X			X	Grades 4–5	R	n.a.	1	NA	n.a.	n.a.
Teacher questionnaire of practices and economic attitudes (Problem-Based Economics REL-West)[d]					X	X	Grade 12	R	n.a.	3	NA	n.a.	n.a.

Source: The information included in this table was drawn from the manuals or other resources available from the authors and publishers of the measures. Individual users may have different experiences with locating information.

[a] n.a. for *not applicable* refers to recently developed measures summarized in a table (Volume II, Table D.1). These measures are so recent that at the time this report was prepared, published information was not available or the study report had little to no psychometric data such that information on cost, norming samples, or administration was limited and/or not meaningful for evaluating such a measure. Thus, the information is not applicable for these subsets of measures.

[b] Ratings refer to total test scores or scores commonly reported; individual subscales may differ, as noted in the specific profile. In addition, ratings may reflect availability based on a previous version, as noted in the specific profile.

[c] Validity ratings reflect the availability of predictive or construct/concurrent validity. Ratings may also reflect availability based on a previous version, as noted in the specific profile.

[d] These measures may be found in Volume II, Table D.1 on recently developed study measures with limited availability of psychometric information.

Table IV.3 (*continued)*

KEY

Domain
CQ = Classroom Quality
CI = Comprehensive Instructional Practices
RI = Reading Instructional Practices
MI = Math Instructional Practices
Oth IP-C = Other Instructional Practice Content Area
Oth SC = Other Social Context

Assessment Type
D = Direct Assessment
O = Observation
R = Parent/Teacher/Student Report

Initial Material Cost
n.a. = Not applicable
TBD = To be determined upon negotiation with the publisher
1 = under $100
2 = $100 to $200
3 = $201 to $500
4 = >$500

Reliability
1 = None described
2 = All or mostly under 0.70
3 = All at or above 0.70.

Validity
NA = Not Available

A = Available

Norming Sample
n.a. = Not applicable
1 = None described
2 = Older than 10 years or not nationally representative
3 = Normed within past 10 years and nationally representative

Ease of Administration and Scoring
n.a. = Not applicable
1 = Not described
2 = Self-administered or administered and scored by someone with basic clerical skills
3 = Administered and scored by a highly trained individual
4 = Administered or scored by a clinician or specialist or the publisher

motivation, self-efficacy, or violence prevention. Eleven measures include assessment of these domains while evaluating instructional practices, classroom instructional environment, productive use of time, and/or other aspects of quality.

The domain of interest varies across the classroom measures. Fifteen measures provide information about reading instructional practices. Nine yield information about mathematics instructional practices and three provide estimates of both reading and mathematics instruction. Six measures assess instructional practices that apply across curricular areas (comprehensive instructional practices) and sometimes look at the application of a theoretical approach to instruction (see, for example, the Reformed Teaching Observation Protocol). Fifteen measures examine aspects of the environment and classroom interactions that are considered important for improving student outcomes, such as positive behavior management, productive use of time, and supportive relationships.

The majority of available measures focus on the early childhood and elementary years, with far fewer designed for middle school and secondary grades. Fourteen measures are intended for or have been used only in early childhood classrooms, from preschool through grade 3. Researchers used 13 measures (including the early childhood measures used with primary grades) in kindergarten through grade 5. Four measures have been used only in the middle grades; six measures were created for grades 7 through 12, with an additional one designed for grades 4 through 12. Six measures address the full range from kindergarten through grade 12. Measures for use across the full grade range typically assess instructional practices that are either cross-curricular or pertain to language and literacy.

Twenty-two of the measures of classroom settings involve only an observation, and 13 collect information only from parent, student, or teacher reports. Two of the early childhood measures collect both observation data and teacher reports. The parent and student report measures focus primarily on aspects of school and classroom climate, whereas the teacher report measures vary in focus. Some measures ask teachers to report on the frequency of use of different, often content-specific instructional practices and strategies (N = 3); others ask about classroom and school climate. Some measures also ask teachers to report on self-efficacy in teaching or their attitudes toward different instructional approaches or practices.

The classroom observation measures use a variety of approaches, including time sampling, event sampling, rubrics, and rating scales. Some measures use a sum of frequency counts; others code a specific aspect of quality of instruction or interactions with students. The complexity of the coding or rubric differs across measures. Some codes, for example, are worded objectively and require the observer to note the presence or absence of specific materials in the environment. Other coding systems involve complex judgments on the part of the observer, such as assessing the quality of feedback, and thus require a high level of observer training in order to ensure the collection of reliable data.

Classroom observation data are much more costly to collect than teacher, student, or parent reports because of the costs associated with training and the number of raters/observers needed to conduct multiple visits to each classroom. The initial cost of obtaining those measures (without training) is low, except for ones that require developer-delivered training as a condition of use. However, training costs for any classroom observation increase the expense of data collection, and the level or extent of the training needed to attain reliability on observation measures varies

with the complexity of the coding. The most costly aspect of classroom observation is the visit to the classroom. Classroom observations often require several hours in a classroom either in a single day or during several visits. Time sampling measures, such as the Literacy Observation Tools (LOT), report the number of observations needed to attain different levels of reliability. Other observation measures recommend the amount of time required based on the fielding of the observation.

Six studies report internal consistency reliability greater than 0.70 for the survey measures (teacher, parent, or student reports). The other survey measures are new and their available documentation does not include estimates of reliability. Fifteen measures, all classroom observations, have some validity evidence, usually a relationship with a student outcome.

Classroom observation measures report inter-rater reliability in a variety of ways, but most often as percentage agreement among observers. Some report agreement with a gold standard rater (usually the developer) or between paired raters. A few measures (for example, the Diagnostic Classroom Observation Tool) report adjacent agreement (that is, "within 1" point) rather than exact agreement with ratings or rubrics. Some measures report agreement only before observers go into the field and do not include the frequency of agreement among raters in the field. In practice, raters may "drift" from the prescribed use of scales and begin to use scales in idiosyncratic ways that introduce a "rater effect" into the data. As noted, the complexity of the coding or rubric and the amount of guidance provided to raters pose challenges for reaching rater agreement. More complex coding systems and higher inference codes can increase the cost of training and the difficulty in achieving inter-rater reliability, but may also capture nuances of classroom interaction that are difficult to obtain in well-anchored behavioral descriptions. Some classroom observation measures in the Compendium have extensive training materials that include videotapes for both training and reliability testing (for example, the Sheltered Instruction Observation Protocol); some measures offer no training support (for example, the Authentic Instructional Practices Classroom Observation Form).

Studies report inter-rater reliability estimates that exceed 0.70 for 17 of the classroom observation measures; however, the estimates may represent agreement between raters within adjacent categories (within 1 point) rather than exact agreement with more sophisticated estimates.[3] Eight of the 17 measures report agreement only, 5 report interrater reliability and 3 report agreement and reliability, and 1 did not report either. The recently developed measures (those without profiles) do not yet have available psychometric information. Measures with lower inter-rater agreement, particularly when evaluating adjacent agreement, contain rater effects that decrease the ability to detect differences in groups and, depending on how raters are assigned to classrooms, may bias the results of any analysis of the data.

SUMMARY OF THE COMPENDIUM MEASURES

The Compendium of Student, Teacher and Classroom Measures Used in NCEE Evaluations of Educational Interventions is intended to help researchers efficiently select measures for future studies, assist policymakers in understanding the measures used in existing studies, facilitate

[3] Kappa coefficients and intraclass correlations are examples of more sophisticated ways than agreement to document inter-rater reliability.

comparisons of results across studies, and broaden understanding of these measures within the educational research community. The 63 measures profiled in the Compendium, combined with the 31 recently developed measures, represent a small fraction of the measures available in the field. For example, the *Mental Measurements Yearbook* (Geisinger et al. 2007) provides reviews and information for almost 4,000 tests, many of which are designed for school-age children, but some of which are more than 20 years old and no longer in print. In contrast, the measures in the Compendium are those that researchers selected for recent studies or developed after reviews of available assessments yielded few or unsatisfactory options. As such, they reflect current thought about the domains of interest in recent studies of NCEE educational interventions.

Of the 94 measures included in the Compendium[4], 52 are directed toward students, 9 to teachers and 37 to classroom practices or settings. More measures are available for the elementary grades than for the secondary grades, with 43 measures that can be used to assess outcomes in K-8 but only 15 that can be used in grades 9-12. Assessments of language (English) and literacy are more common than are assessments of other content areas, particularly social studies and the sciences. Twenty-two of the 37 classroom measures were newly developed to meet the needs of a specific project, in contrast to 9 of the 43 student measures and none of the 9 teacher measures. Some of the newly developed measures draw on earlier research by combining items or subscales from existing measures. The decision to create a new measure may be prompted by the sense that existing measures are not well aligned with the intervention being examined or the specific research questions addressed by a study.

Although 17 measures in the Compendium indicated the availability of a Spanish version with documented psychometric properties, most assessments lack a Spanish version with norms and properties comparable to the English versions[5]. Sometimes the norms are based on samples of children from Spanish-speaking countries (for example, Woodcock Johnson-III) rather than samples of children in the United States who come from homes where Spanish is the primary language. In addition, evidence suggests that a single assessment in either English or Spanish underestimates children's linguistic and cognitive skills, particularly in the early years (Bedore et al. 2005; Pavlenko 1999; Umbel et al. 1992). Recent research supports the use of alternative assessment and scoring procedures for bilingual children in order to quantify their skills appropriately (see Bedore et al. [2005] and Brownell [2000] for descriptions of conceptual scoring, an approach designed to credit children for mastery of concepts in either English or Spanish). Approaches for addressing language diversity are important issues when an evaluation team seeks to use either outcome data from all students to assess intervention impacts on achievement or other outcome domains that require a direct assessment or student self-report. If equivalent forms of a measure are not available in different languages, the pooling of data across assessments may result in misleading conclusions about program impacts.

[4] There are 67 profiles and 31 table entries. Two teacher profiles also appear as student profiles and two classroom profiles also appear as teacher profiles. Therefore, a total of 94 measures are included in the compendium.

[5] Of the 17 measures with a Spanish version, 4 have norming samplings based on U.S. children and 3 used norming samplings from U.S. and other countries combined. Others do not report norming sample.

By comparing the type of assessment and other key information, the summary tables in this chapter can serve as a first step in locating a measure that addresses a specific domain. The standard format used in presentation of information in the Compendium profiles should also help facilitate the comparison of the measures used in NCEE studies.

REFERENCES

Agodini, Roberto, John Deke, Sally Atkins-Burnett, Barbara Harris, and Robert Murphy. "Design for the Evaluation of Early Elementary School Mathematics Curricula." Submitted to the Institute of Education Sciences, U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, Inc., January 2008.

Agodini, Roberto, Barbara Harris, Sally Atkins-Burnett, Sheila Heaviside, Timothy Novak, and Robert Murphy. "Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools." No. (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *The Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association, 1999.

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

Annie E. Casey Foundation. 2008 KIDS COUNT Data Book. (2008) Retrieved from: http://datacenter.kidscount.org/databook/2008/Default.aspx.

Bacon, Donald. "The Contributions of Reliability and Pretests to Effective Assessment." *Practical Assessment, Research & Evaluation,* vol. 9, no. 3, 2004,

Bedore, Lisa, Elizabeth Pena, Melissa Garcia, and Celina Cortez. "Conceptual Versus Monolingual Scoring: When does it make a Difference?" *Language, Speech, and Hearing Services in Schools,* vol. 36, 2005, pp. 188-200.

Berry, Daniel J., Lisa J. Bridges, and Martha J. Zaslow. "Early Childhood Measures Profiles." Washington, DC: Child Trends, 2004.

Brennan, Robert L. "Elements of Generalizability Theory." Iowa City, IA: American College Testing Program, 1983.

Brownell, Rick. *Expressive One Word Picture Vocabulary Tests– Spanish Bilingual Edition (EOWPVT-SBE).* Novato, CA: Academic Therapy Publications, 2000.

Campbell, Donald T. and D. W. Fiske. "Convergent and Discriminant Validation by the Multi-Trait-Multimethod Matrix." *Psychological Bulletin,* vol. 56, 1959, pp. 81-105.

Cohen, J. "A Coefficient for Agreement for Nominal Scales." *Educational and Psychological Measurement,* vol. 20, 1960, pp. 37-46.

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

Crocker, L. and J. Algina. *Introduction to Classical and Modern Test Theory.* Orlando, FL: Holt, Rinehart and Winston, Inc., 1986.

Dunn, Lloyd M. and Douglas M. Dunn. *Peabody Picture Vocabulary Test—Fourth Edition Manual.* Minneapolis, MN: Wascana Limited Partnership, 2007.

Dwyer, M.C., Smith, W.C., Dixon, L.Q., and Gamse, B.C. "The Development of the Instructional Practice in Reading Inventory." *Presented at the annual meeting of the American Educational Research Association.* Chicago, 2007.

Educational Testing Service. "Mathematica Reading Comprehension Assessments 2007 Technical Report." Princeton, NJ: Educational Testing Service, December 2007.

Embretson, Susan E. "Issues in the Measurement of Cognitive Abilities." In *The New Rules of Measurement: What Every Psychologist and Educator Should Know,* edited by Susan E. Embretson and Scott L. Hershberger. Mahwah, NJ: Lawrence Erlbaum Associates, 1999.

Embretson, Susan E. "The New Rules of Measurement." *Psychological Assessment,* vol. 8, 1995, pp. 341-349.

Embretson, Susan E. and Scott L. Hershberger. *The New Rules of Measurement: What Every Psychologist and Educator Should Know.* Mahwah, NJ: Lawrence Erlbaum Associates, 1999.

Embretson, Susan E. and Steven P. Reise. *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

Federal Interagency Forum on Child and Family Statistics. "America's Children in Brief: Key National Indicators of Well-being, 2008." Washington, DC: U.S. Government Printing Office, 2008.

Frick, T. and M. L. Semmel. "Observer Agreement and Reliabilities of Classroom Observational Measures." *Review of Educational Research,* vol. 48, no. 1, 1978, pp. 157-184.

Garcia, Eugene and Bryant Jensen. "Early Educational Opportunities for Children of Hispanic Origins." *Social Policy Report,* vol. 23, no. 2, 2009, pp. 1-19.

Geisinger, Kurt F., Robert A. Spies, Janet F. Carlson, and Barbara S. Plake (eds.). *The Seventeenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements, 2007.

Grehan, Anna and Lana J. Smith. *The Early Literacy Observation Tool (E-LOT).* Memphis, TN: The University of Memphis, Center for Research in Educational Policy, 2004.

Gresham, Frank M. and Stephen N. Elliott. *Social Skills Rating System.* Bloomington, MN: Pearson Assessments, 1990.

Groves, Robert M. and Mick P. Couper. *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons, Inc., 1998.

Halle, Tamara and Jessica Vick. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures." Submitted to the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC: Child Trends, November 2007.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications, 1991.

Harms, Thelma, Richard M. Clifford, and Debby Cryer. *Early Childhood Environment Rating Scale, Revised Edition.* New York: Teachers College Press, 1998.

Hill, H. C., D. L. Ball, and S. G. Schilling. "Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students." *Journal for Research in Mathematics Education,* vol. 39, 2008, pp. 372-372.

Hoogendoorn, Adriann W. and Dirk Sikkel. "Response Burden and Panel Attrition." *Journal of Official Statistics,* vol. 14, no. 2, 1998, pp. 189-205.

Jackson, Russell, Ann McCoy, Carol Pistorino, Anna Wilkinson, John Burghardt, Melissa Clark, Christine Ross, Peter Schochet, and Paul Swank. "National Evaluation of Early Reading First: Final Report." Submitted to the Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office, May 2007.

James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, and Bonnie Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students." No. (NCEE 2009-4032). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Kisker, Ellen E., Kimberly Boller, Charles Nagatoshi, Christine Sciarrino, Vinitia Jethwani, Teresa Zavitsky, Melissa Ford, and John M. Love. "Resources for Measuring Services and Outcomes in Head Start Programs Serving Infants and Toddlers." Washington, DC: Mathematica Policy Research, Inc., 2003.

Kim, J. and H.K. Suen. "Predicting children's academic achievement from early assessment scores: A validity generalization study." *Early Childhood Research Quarterly*, 18, 2003, 547-566.

LaParo, K.M. and R.C. Pianta. "Predicting children's competence in the early school years. A meta-analytic review." *Review of Educational Research*, 70 (4), 2000, 443-484.

Linn, Robert L. and Norman E. Gronlund. *Measurement and Assessment in Teaching, 8th Edition.* Des Moines, IA: Prentice Hall, 1999.

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

Mather, Nancy, Donald D. Hammill, Elizabeth A. Allen, and Rhia Roberts. *Test of Silent Word Reading Fluency. Examiner's Manual.* Austin, TX: Pro-Ed, Inc., 2004.

May, H., I. Perez-Johnson, J. Haimson, S. Sattar, and P. Gleason. "Using State Tests in Educational Experiments: A Discussion of the Issues, Institute of Education Sciences, U.S. Department of Education." No. NCEE 2009-013. Washington, DC: National Center for Education, Evaluation, and Regional Assistance, 2009.

McDermott, Paul A. "National Standardization of Uniform Multisituational Measures of Child and Adolescent Behavior Pathology." *Psychological Assessment,* vol. 5, no. 4, 1993, pp. 413-424.

McKenna, M. C. and K. A. Dougherty Stahl. *Assessment for Reading Instruction. Second Edition.* New York, New York: Guilford Press, 2009.

Messick, Samuel. "Foundations of Validity: Meaning and Consequences in Psychological Assessment." *European Journal of Psychological Assessment,* vol. 10, no. 1, 1994, pp. 1-9.

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

Pavlenko, Aneta. "New Approaches to Concepts in Bilingual Memory." *Bilingualism: Language and Cognition,* vol. 2, no. 3, 1999, pp. 209-230.

Pew Research Center. Questionnaire Design. (2010) Retrieved from: http://people-press.org/methodology/questionnaire/. Accessed on January 10, 2010.

Raundenbush, S. W. "Advancing Educational Policy by Advancing Research on Instruction." *American Educational Research Journal,* vol. 45, no. 1, 2008, pp. 206-230.

Raundenbush, S. W. and Sally Sadoff. "Statistical Inference when Classroom Quality is Measured with Error." *Journal of Research on Educational Effectiveness,* vol. 1, 2008, pp. 138-154.

Regional Education Laboratory Program. The Impact of Professional Development Strategies on Teacher Practice and Student Achievement in Math. (n.d.) Retrieved from http://ies.ed.gov/ncee/projects/evaluation/tq_mathematics.asp. Accessed on December 8, 2008(a).

Regional Education Laboratory Program. Impacts of a Problem-Based Instruction Approach to Economics on High School Students. (n.d.) Retrieved from: http://ies.ed.gov/ncee/edlabs/projects/rct_89.asp?section=ALL. Accessed on December 8, 2008(b).

Ross, Christine, Gretchen Kirby, Peter Schochet, John Hall, Susan Sprachman, Kimberly Boller, Diane Paulsell, and Sheena McConnell. "Design Options for the Assessment of Head Start Quality Enhancements." Washington, DC: Mathematica Policy Research, 2005.

Salvia, John and James E. Ysseldyke. *Assessment in Special Education and Inclusive Education – Ninth Edition*, Boston: Houghton Mifflin Company, 2004.

Schochet, Peter. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Sciences,* vol. 33, no. 1, 2008, pp. 62-87.

Shavelson, Richard J. and Noreen M. Webb. *Generalizability Theory: A Primer.* Newbury Park, CA: Sage Publications, 1991.

Shrout, P. and J. L. Fleiss. "Intraclass Correlation: Uses in Assessing Rater Reliability." *Psychological Bulletin,* vol. 86, no. 2, 1979, pp. 420-428.

Shulman, Lee S. "Those Who Understand: Knowledge Growth in Teaching." *Educational Researcher,* vol. 15, no. 2, 1986, pp. 4-14.

Smith, Miriam W., Joanne P. Brady, and Louisa Anastasopoulos. *Early Language & Literacy Classroom Observation Pre-K Tool.* Baltimore: Paul H. Brookes Publishing Co., 2008.

Strauss, E., M. S. Sherman, and O. A. Spreen. *A Compendium of Neuropsychological Tests. Administration, Norms, and Commentary. Third Edition.* New York, New York: Oxford University Press, 2006.

Traub, Ross E. *Reliability for the Social Sciences. Theory and Applications.* Thousand Oaks, CA: Sage Publications, 1994.

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs. "27th Annual (2005) Report to Congress on the Implementation of the Individuals with Disabilities Education Act." Volume 1. Washington, DC: U.S. Department of Education, 2007.

Umbel, Vivian M., Barbara Z. Pearson, Maria C. Fernandez, and D. K. Oller. "Measuring Bilingual children's Receptive Vocabularies." *Child Development,* vol. 63, 1992, pp. 1012-1020.

W.K. Kellogg Foundation. "W.K. Kellogg Foundation Logic Development Guide." Battle Creek, MI: W.K. Kellogg Foundation, 2001.

Woodcock, Richard W., Kevin S. McGrew, Fredrick A. Schrank, and Nancy Mather. *Woodcock-Johnson III Normative Update.* Rolling Meadows, IL: Riverside Publishing, 2001, 2007.

Wright, Benjamin D. and Mark H. Stone. *Best Test Design*. Chicago: The Phaneron Press, 2009.

Young, John W. "Review of the Test of Silent Word Reading Fluency." In *The Sixteenth Mental Measurements Yearbook,* edited by Robert A. Spies and Barbara S. Plake. Lincoln, NE: The Buros Institute of Mental Measurements, 2005.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA