Journal of Early Intervention
Volume 33 Number 3
September 2011 171-185

10.1177/1053815111427566
http://jei.sagepub.com
hosted at
http://online.sagepub.com

JEI Guidelines for Manuscripts Describing the Development and Testing of an Assessment Instrument or Measure

Charles R. Greenwood
University of Kansas, Kansas City
Scott R. McConnell
University of Minnesota, Minneapolis

The purpose of this article is to describe guidelines for reporting and reviewing findings from studies and to provide guidance to authors seeking to report results of research on measurement and the development or evaluation of assessment procedures or instruments in early intervention (EI) and early childhood special education (ECSE). Following on a recent series of similar articles in the *Journal of Early Intervention* and *Exceptional Children*, this article will (a) provide rationale for the specific focus of assessment and measurement research in EI/ECSE, (b) describe existing standards for this research and practice, and (c) provide guidelines for preparation, review, and publication of research in EI/ECSE.

Keywords: *assessment; measurement; research; methodology*

Assessment—the act of systematically collecting information to decide what, if anything, to do—is assuming an increasingly important and more central role in education and related services for young children with disabilities and other special needs. Early

Authors' Note: Charles R. Greenwood, Juniper Gardens Children's Project, University of Kansas; Scott R. McConnell, Department of Educational Psychology and Center for Early Education and Development, University of Minnesota. When R. A. McWilliam was editor of the *Journal of Early Intervention*, he solicited short papers to describe important issues in the reporting of manuscripts using various research methods. We are in the process of updating those papers. In the original set of papers, McWilliam did not include a paper on the reporting of manuscripts focused on the development and evaluation of measures, instruments, tests, and so forth; however, we receive a number of such manuscripts. Greenwood and McConnell graciously agreed to write such an article, and it appears in this issue. Authors will find their comments useful in preparing manuscripts focused on the development and evaluation of measures, and reviewers will find their comments helpful in judging the suitability of those manuscripts for publication. Subsequent issues will contain other articles on reporting using other research methods. This work was supported in part by Grant R324C080011, the *Center for Response to Intervention in Early Childhood*, from the Institute of Education Sciences (IES), U.S. Department of Education, to the University of Kansas; Charles Greenwood and Judith Carta, Principal Investigators. However, the opinions and recommendations presented in this article are those of the authors alone, and no official endorsement from the IES should be inferred. The authors are indebted to a wide range of colleagues who have taught us the conceptual, methodological, and statistical issues relevant to measurement and assessment over the years and to the children and teachers who have joined in that instructional effort through their participation in our research and the use of its results. Correspondence concerning this article should be addressed to Charles R. Greenwood, Juniper Gardens Children's Project, 444 Minnesota Avenue, Suite 300, Kansas City, KS 66101-2914; email: greenwood@ku.edu.

interventionists, early childhood special educators, and their colleagues conduct assessments to identify children who might be eligible for and/or benefit from special services, to plan and monitor the efficacy of intervention programs for individuals or groups, to tailor interventions based on children's responses, or to describe the ongoing or summative effect of one or several complementary intervention efforts (McLean, Wolery, & Bailey, 2003). Formal assessment activities have a long history (preceding by many decades the establishment of special education mandates in the United States and other countries), although their historical roots link closely to description and assistance to individuals with disabilities. In recent years, assessment practices have assumed new and more central roles in early intervention (EI) and early childhood education with increased attention to diagnosis of higher numbers of disabilities, disorders, and syndromes and the use of systematically collected information in treatment planning (Bagnato, Neisworth, & Pretti-Frontczak, 2010; McEvoy, Neilsen, & Reichle, 2003), progress monitoring (Carta, Greenwood, Walker, & Buzhardt, 2010; Missall, Carta, McConnell, Walker, & Greenwood, 2008), and outcome reporting and evaluation (Early Childhood Outcomes Center, 2011a).

As assessment practices in EI and early childhood special education (ECSE) have expanded, so too has the systematic measurement research that undergirds these practices. Although a precise definition in psychology and education is somewhat controversial, in general we can define *measurement* as the science of determining or estimating ratios of quantities, or the ways we quantify and create metrics for specific or complex constructs of interest (Michell, 1977). As a result of this expanding knowledge base about how to develop and evaluate research and practices, our field has developed to a point where we are articulating, and beginning to adopt, common guidelines and quality standards for conducting, reporting, and evaluating work in this area.

Like assessment, measurement research has a long history generally, with a focus on young children with special needs particularly. For the purposes of this article, we distinguish *assessment* as the set of practices, procedures, and tools that are used in practice to collect information and support decision making; *measurement* represents the core features and characteristics of assessment and thus is more typically the focus of research to improve and expand assessment resources. At the turn of the 20th century, researchers like Alfred Binét, David Wechsler, and others were interested in assessing intelligence as a first step in providing therapeutic services to children with mental retardation. Fueled by the child study movement and developments in statistical analysis, measures of young children's development and skills expanded during the first half of that century (Ludy, 2007). In the latter half, researchers both deepened and intensified work on standardized measurement for young children (McLean et al., 2003), but the field also witnessed a burgeoning interest in behavioral assessment and the measurement components related to this practice (Ollendick, Alvarez, & Greene, 2004). With recent funding from the Office of Special Education and Institute of Education Sciences (IES) within the U.S. Department of Education (e.g., Early Childhood Research Institute on Measuring Growth and Development, 1996-2001; Early Childhood Outcome Center, 2004-2009), the profession's embrace of "Recommended Practices" (Bagnato et al., 2010; Neisworth & Bagnato, 2000) and formal attention from researchers and "mainstream" EI/ECSE publishers such as Brookes Publishing and the Ages and Stages Questionnaire and the *Assessment, Evaluation, and Programming System* (Bricker, Capt, & Pretti-Frontczak, 2002; Squires, Bricker, & Twombly, 2002),

formal attention to measurement—as well as critical appraisal of measurement research and its contribution to assessment practices—is now possible.

Our purpose in this article is to describe guidelines for reporting and reviewing findings from studies and to provide guidance to authors seeking to report results of research on measurement and the development or evaluation of assessment procedures or instruments in EI and ECSE. Following on a recent series of similar articles in the *Journal of Early Intervention* (McWilliam, 2000; Snyder, 2000; Wolery & Dunlap, 2001) and *Exceptional Children* (Odom et al., 2005), this article will (a) provide rationale for the specific focus of assessment and measurement research in EI/ECSE, (b) describe existing standards for this research and practice, and (c) provide guidelines for preparation, review, and publication of research in EI/ECSE.

Assessment and Measurement Research in EI/ECSE

As we have noted, assessment practices and measurement research with and for children with disabilities and other special needs have a relatively long and robust history (McLean et al., 2003). In contrast with dominant perspectives in "general" early childhood education (Kagan, Moore, & Bredekamp, 1995), early interventionists and early childhood special educators have routinely adopted assessment practices and turned to measurement research as standards for "best practice" in our clinical activities (Division of Early Childhood [DEC], 2007; Neisworth & Bagnato, 2000) and as one of the central pillars of our discipline more broadly (Bagnato, 2007). With a focus on the developmental status and needs of young individual children, we have embraced procedures suited to describing these individuals in a variety of ways. This reflects one of what might be the core tenets of our practice—that better understanding (as achieved through assessment) will lead to making better decisions regarding services and supports that children need, that will in turn lead to better outcomes for individuals *and* groups of children. The case supporting accurate decision making and correctly made inferences from data produced by using a measure or set of measures is made through a series of evidentiary statements (findings) that are the product of careful development and validation research.

Historically, early interventionists and early childhood special educators have focused their assessment practices and measurement research on four primary purposes of assessment (compare Hawkins, 1979). First, we are interested in the rigor, accuracy, and efficiency of our *assessment for determining eligibility for special services*. Here, individual child characteristics are most typically compared with some a priori standard—either normative or functional—and services are allocated based on children's meeting (or not meeting) these standards. Eligibility evaluations that precede and contribute to eligibility determinations in EI/ECSE are the most common current examples, but similar practices such as universal screening and progress monitoring are emerging in early childhood Response to Intervention (Greenwood et al., 2011) and other contemporary practices.

Second, EI/ECSE professionals have a long history of conducting *assessment for program planning*. This work, often embedded in development of Individualized Family Service Plans (IFSPs) and/or Individualized Education Plans (IEPs), can be formal and linked to specific interventions (Dunlap, Kern, Clarke, & Robbins, 1991) or curricula (Bricker,

2010), or can be more general and conducted by practitioners developing highly idiosyncratic intervention programs. Assessment for program planning includes evaluation of specific skills, competencies, or areas of development that warrant intervention attention (i.e., determining intervention goals and objectives for individual children) as well as assessment of the differential effectiveness of different approaches to intervention to identify the most promising practice(s) for an individual at one point in time.

Progress monitoring is the third function of assessment common in our field, both as a standard of best professional practices and as a statutory and procedural requirement of many services that we offer. Here, assessment data are collected periodically and repeatedly to (most often) assess the efficacy of an individual's intervention or support services; this practice brings into stark relief the central role that assessment and measurement play in EI/ECSE as a way of operationalizing the idiographic, intensive, and intervention-oriented approach of our field. Fourth, EI/ECSE has a long tradition of *assessment for program evaluation*, at the level of the individual and at the level of groups of children or families. At the individual level, assessment for program evaluation is also a required element of IFSP- and IEP-driven services; as a matter of course, EI/ECSE professionals and programs assess and evaluate summative effects of services and supports (particularly at times of transition). Similarly, EI/ECSE has a long tradition (as represented by research presented in this journal) of evaluating effects of larger program efforts—intervention approaches and techniques, curricula, and/or programmatic or policy innovations—on developmental outcomes for larger groups of students. Most recently, the U.S. Department of Education has selected and begun to implement a common set of outcomes and metrics for describing early childhood outcomes at the child, program, state, and national levels to document program impact (Early Childhood Outcomes Center, 2011b; Greenwood, Walker, Hornback, Hebbeler, & Spiker, 2007). All of these are examples of assessment for program evaluation.

In addition to these longstanding and relatively common functions of assessment in EI/ECSE, several other approaches or functions of assessment have emerged in recent years and are likely to continue drawing attention in research and practice in the years ahead. Briefly, these emerging areas include the following:

- *Curriculum-based or curriculum-embedded assessment.* Here, several of the more typical assessment purposes are integrated into a larger "system" of curriculum, instruction, and monitoring. At the most general level, this approach is represented by selection or development of an overarching skill, competency, or developmental task-based curriculum with procedures to monitor individual children's status across time, and to use relations between child status, curriculum hierarchies or elements, and developmental expectations or goals as the basis for designing, allocating, evaluating, and improving services over time (e.g., Boulware, Schwartz, Sandall, & McBride, 2006; Bricker, 2010).
- *General outcome measurement.* Researchers and practitioners in EI/ECSE have been adapting and expanding approaches to assessment initially developed for elementary and secondary students (e.g., Deno, 1997) for assessing the developmental status and growth of individual children and groups (Greenwood et al., 2008; McConnell & Missall, 2008). These measures are characterized as brief, repeatable, easy-to-use measures of children's development that are designed specifically to show growth over time toward a meaningful, but relatively long-term, desired outcome.
- *Status measures for differentiated intervention models, including Response to Intervention.*

Most recently, researchers, practitioners, and policy makers have increased attention to the special requirements of assessment practices that can be used to identify children who might benefit from different forms or tiers of intervention support. Most commonly associated with comprehensive interventions like the Pyramid Teaching Model (Fox, Carta, Strain, Dunlap, & Hemmeter, 2010) and emerging Response to Intervention systems (Buysse & Peisner-Feinberg, 2009; Buzhardt et al., 2011), assessment and measurement take a new and central role in program innovation for young children with special needs and their families and professionals who serve them.

- *Progress monitoring and quality rating vis-à-vis policy or program standards.* Loosely referenced as a growing "accountability movement" in EI/ECSE, as in many other channels of education and human services, assessment practices are being developed and implemented to determine the extent to which teachers, classrooms, and programs meet or conform with external standards of "quality," as well as the extent to which children served by these teachers, classrooms, or programs are achieving outcomes that meet expectations, including the expectation that children are "ready for kindergarten" (Rous, LoBianco, Moffet, & Lund, 2005).

Ongoing and continued work on the most common purposes of assessment (i.e., eligibility, program planning, progress monitoring, and program evaluation), as well as expansions and innovations in assessment and measurement practices generated by the field or the systems that influence and govern it, present ongoing challenges to practitioners, researchers, and policy makers. For example, "What assessment practices are needed to drive and support effective services?" "What characteristics should these practices possess?" "How should research and development of these practices be designed, conducted, and reported to maximize efficiency and effectiveness?" These questions, and others, lead directly to the need for guidelines that help define quality in assessment and measurement research.

Existing Standards for Assessment and Measurement Research

At the core, measures used in this field can be described in terms of five basic features: (a) the purposes they are intended to serve; (b) the content measured; (c) the methods of administration, collection, and interpretation of data; (d) the respondents whose behavior is measured; and (e) the informants who make the records and thus provide the data. For example, the *purpose* may be any of those previously stated (e.g., progress monitoring), and all measures are developed to fulfill some single or multiple purposes in mind. *Content* is the definition including the breadth and depth of what is to be measured; for example, the domains, outcomes (e.g., social or cognitive), taxonomy, and skills, including specific items. *Methods* are the particular measurement procedures used to guide collection of the data and render scores (the data). Well-known methods, for example, are rankings, ratings, interviews, tests, and direct observations. *Respondents* are the individuals whose performance is being measured (e.g., a child, a teacher). Children, for example, may be responding to objects or pictures that are part of the measure. *Informants* are those individuals who are using the method to produce the raw data. Parents, for example, might be raters of children, observers may collect data representing the performance of teachers in classrooms, or teachers may engage in collecting a 1-min probe of the language and early literacy skills using items in a deck of cards. In most cases, informants are trained, taught, or at least

instructed how to use the method (administration) to some standard to collect the data of interest.

While the purposes that measures serve will vary, the need for the resulting information to be trustworthy is central to the evaluation of any measure and the data and decisions resulting from the data collected using that measure. Thus, research developing and evaluating a new measure or improving and expanding an existing measure must be guided by standards of quality. At least three major measurement standards apply to measurement in our field: (a) the *Standards for Educational and Psychological Testing* produced by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education; (b) the *DEC Recommended Practices* of the Division of Early Childhood; and (c) the *Student Progress and Mastery Monitoring (SPMM)* standards as described by the National Center for Response to Intervention (formerly the National Center for Student Progress Monitoring).

At the core of measure development and validation broadly are the widely accepted AERA standards (see Table 1) covering key methods and associated evidence. The DEC Recommended Practices and the Student Progress Monitoring standards add specific directions and refinement given the early childhood focus of the field (i.e., DEC) and the particular purpose and types of information desired and needed (i.e., SPMM standards). One of the challenges faced by authors organizing papers and deciding what to include and by reviewers asked to evaluate manuscripts reporting assessment/measurement research for *JEI* is knowledge of these contemporary standards and the relationship between them.

Table 1 provides our attempt to integrate the three standards such that their shared and unique aspects can be seen clearly. The AERA standards in the table are listed in the left most column, DEC Recommended Practices in the center column, and SPMM standards in the right most column. We examine each briefly.

Standards for educational and psychological testing. We argue (AERA, 1999) that these standards apply generally to the development and validation of all psychological, educational, and behavioral performance measures intended to produce data that are rigorous enough to support trust of researchers, parents, practitioners, and policy makers in the inferences and conclusions made from the data (interested readers should consult the standards). Conclusions based on collected data have potentially high stakes consequences, for example, a child is eligible and may begin receiving services or not, a child is making sufficient progress to receive a change in intervention or not, or a program receiving state funding support will continue to receive support or it will not. Making these decisions must be based on high standards and rules of evidence.

These AERA standards are accepted by numerous scientific and professional stakeholders and policy makers. For example, they are prominent in most contemporary texts on tests and measurement (Salvia & Ysseldyke, 2000) and in the manuscript preparation guidelines of the APA as regards preparation of research reporting empirical findings of measurement development or improvement investigations. They are also expected elements in measurement research, for example, the Goal 5 Measurement projects funded by the IES (e.g., FY2012 Special Education Research, p. 64, Posted February 28, 2011). And, they are criteria considered in scientific merit reviews of grant applications at the IES and National Institutes of Health.

Table 1
Intersection of Three Existing Standards for Assessment and Measurement Research Matched on Similar Content or Stated Criteria

Standards for educational and psychological testing (AERA/APA/NCME)	DEC recommended assessment practices	Student progress and mastery monitoring standards	
		Progress monitoring	Mastery monitoring
1. Domain (or outcome) specification	Acceptability[a]	General outcome social validation	Skill domain social validation
2. Review of relevant literature to specify behaviors of interest	Acceptability[a]	Key skills indicators	Skill sequence specified
3. Item development and refinement	Acceptability[a], Authentic[b], Congruent[c], Collaborative[d]	Key skill indicators	Skill sequence specified
4. Content validation and steps addressing the development of measure content	Acceptability[a], Authentic[b], Congruent[c], Collaborative[d]	Social validation and evidence that content measures the outcome	Social validation and evidence that content measures the outcome
5. Practicality and utility review of proposed administration formats and scores	Acceptability[a], Authentic[b], Collaborative[d]	Usable by practitioners in intervention decision making	Usable by practitioners in intervention decision making
6. Initial field/pilot testing	Authentic[b], Collaborative[d]	Initial field/pilot testing	Initial field/pilot testing
7. Item analysis and refinement/selection	Equitable[e]	Item analysis and refinement/selection	Item analysis and refinement/selection
8. Reliability evaluation including temporal and, where appropriate, interrater/interobserver evaluations	Convergent[f] (pooling different informant info to form more comprehensive picture, rather than treating interrater disagreement as error)	Reliability of the performance-level score, reliability of the slope of improvement, alternate forms, reliability data disaggregated for diverse populations	Reliability, reliability data disaggregated for diverse populations
9. Classification reliability analysis and determination of empirical benchmarks	Equitable[e]	End-of-year benchmarks, rates of improvement specified	Pass–fail decision criterion
10. Criterion validity evaluation	Equitable[e]	Validity data, disaggregated for diverse populations	Validity data, disaggregated for diverse populations
11. Predictive utility evaluation	Equitable[e]	Validity data, evidence of classification accuracy	Pass–fail decision criterion evidence
12. Internal validity analysis of item structures	Equitable[e]	Evidence supporting skill selection from the universe of skills in the domain	Evidence supporting skill hierarchy
13. Construct validity, including evaluation of sensitivity to growth over time	Sensitive[g], Equitable[e]	Sensitivity to student improvement; validity of the performance-level score; predictive validity of the slope of improvement	Sensitivity to student improvement

(continued)

Table 1 (continued)

Standards for educational and psychological testing (AERA/APA/NCME)	DEC recommended assessment practices	Student progress and mastery monitoring standards	
		Progress monitoring	Mastery monitoring
14. Bias analysis for demographic and regional variables	Equitable[e]	Reliability and norms data disaggregated for diverse populations	Reliability and norms data disaggregated for diverse populations
15. Norming and estimations of base rates at specific classification cut points	Equitable[e]	Norms disaggregated for diverse populations	Pass–fail decision criterion

Note: AERA = American Educational Research Association; APA = American Psychological Association; NCME = National Council on Measurement in Education; DEC = Division for Early Childhood.
[a]*Acceptability* = Measurement materials and approaches have consensus among families and professionals.
[b]*Authentic* = Contrived tasks and persons unknown to the child are avoided, thus preventing reactivity.
[c]*Congruent* = Materials and approaches are designed for and field validated with the children intended to be assessed.
[d]*Collaborative* = The measure's methods engage teamwork between families and professionals in the collection and use of the information.
[e]*Equitable* = The measure accommodates individual differences.
[f]*Convergent* = Data collected on everyday behavior in natural settings by multiple informants can be pooled to provide a more comprehensive set of information about the child.
[g]*Sensitive* = The measure reflects short-term, small increments of progress so that all children, including those with severe disabilities, can be included and accurately represented.

The list of AERA standards for systematic development and technical adequacy evidence of measures include (a) domain specification; (b) review of relevant literature to specify behaviors of interest; (c) item development and refinement; (d) content validation and steps addressing the development of measure content; (e) practicality and utility review of proposed administration formats and scores; (f) initial field/pilot testing; (g) item analysis and refinement/selection; (h) reliability evaluation including temporal and, where appropriate, interrater/interobserver evaluations; (i) classification reliability analysis and determination of empirical benchmarks; (j) criterion validity evaluation; (k) predictive utility evaluation; (l) internal validity analysis of item structures; (m) construct validity; (n) bias analysis for demographic and regional variables; and (o) norming and estimations of base rates at specific classification cut points as related to both measure methods and respondents.

Although this represents a comprehensive list of development activities, it should be noted that some may be conditional on the specific purposes and inferences that are intended in the use of a particular measure. For example, sensitivity to growth over time would be an expected indicator of a progress monitoring measure, whereas predictive utility evaluation would be an expected indicator of a measure intended for screening, identification, and eligibility decision making. Alternately, many of these development activities appear to apply universally to all measures regardless of whether they are behavioral, academic, cognitive, progress monitoring, or developmental. A universal example would be gathering evidence of practicality and utility of the administration format, field/pilot testing, and validity/reliability.

DEC recommended assessment practice. The DEC (Sandall, McLean, Smith, & McLean, 2005) recommended criteria for evaluating the appropriateness of assessment practices seek to contextualize measurement specifically for uses with very young children, their families, and their caregivers/teachers. These criteria are intended to assure that measures and practices (a) point to behavioral objectives for change that are judged important and acceptable; (b) guide change in treatment activities; (c) incorporate several instruments, informants, and scales, including observation and interviews; (d) incorporate input from parents; and (e) are used on multiple occasions (Bagnato, 2007; Sandall et al., 2005).

The DEC standards include acceptability, authenticity, collaboration, convergence, equity, sensitivity, and congruence. To be *acceptable*, measurement materials and approaches should have consensus among families and professionals. To be *authentic*, contrived tasks and persons unknown to the child should be avoided in the assessment process to prevent reactivity (Bagnato, 2007). Reactivity, for example, the "stranger effect" typical of young children, can be reduced by familiarity: assessing the child in familiar surroundings, using familiar objects/materials, having the parent present at the assessment, and, if administered by a professional assessor (e.g., occupational therapist, physical therapist, or school psychologist), spending enough time with the child beforehand. To be *collaborative*, the measure's methods should engage teamwork between families and professionals in the collection and use of the information. To be *convergent*, reliable data collected on everyday behavior in natural settings by multiple informants can be pooled; that is, where differences or unique information is obtained, it is used to provide a more comprehensive set of information about the child. To be *equitable*, the measure is able to accommodate individual differences. To be *sensitive*, the measure is capable of reflecting short-term, small increments of progress so that all children, including those with severe disabilities, can be included and represented accurately. To be *congruent*, materials and approaches are designed for and field validated with the children to be assessed.

Contrasting the two sets of standards, it is clear that the AERA standards guide research and development needed to produce evidence that data from any measure are trustworthy in performing its purpose. The DEC standards more closely emphasize the unique needs of measurement applied to very young children (particularly those with disabilities) and the necessary characteristics of these tools when used with parents or others to design, monitor, or evaluate special education and related services.

SPMM standards. The National Center on Response to Intervention (2011a, 2011b) defines *progress monitoring measures* as those that allow repeated assessment of performance in ways that inform teachers about the need to change, or sustain, an individual student's intervention. Progress monitoring is conducted often (at least monthly, if not weekly or biweekly) to estimate individuals' level of improvement (in growth or slope), to create a data set for identifying students who are not making desired progress, or to compare the efficacy of different forms of intervention for the same child.

Use of progress monitoring measures also is expected to result in better teacher planning and more frequent changes in intervention based on child Response to Intervention (National Center on Response to Intervention, 2011b). Meeting these standards involves research evidence addressing each of these features.

The attributes of high quality progress monitoring measures intersect with both the AERA and DEC standards but also contain requirements particular to the task of measuring change over time. These include sensitivity to (a) improvement over time, alternate forms; (b) end-of-year benchmarks; (c) specified rates of improvement; (d) norms, reliability, and validity disaggregated by diverse populations; and (e) improved student performance or teacher planning.

Sensitivity to improvement is demonstrated in SPMM by evidence that children grow over time and age. *Alternate forms* are important in progress monitoring because they are administered frequently and they rule out memorization ("learning the test") as a measurement confound. *End-of-year benchmarks* are expectations for yearly progress by that point in time. *Rates of progress are specified* either through benchmarks or normative data (including local normative data) illustrating progress within a year (e.g., monthly or quarterly). *Norms, reliability, and validity data that are disaggregated for diverse populations* help eliminate biased decision making with subpopulations of children. *Improved student learning or teacher planning* is demonstrated by evidence that the measure is improved by changes in intervention practices, or by use of the measure itself as feedback to teacher decision making and change in intervention.

The unique standards for mastery monitoring not shared with progress monitoring include skill sequence and pass/fail benchmarks. Mastery monitoring measurement requires a *skill sequence for learning* to be clearly specified and measured. And, *pass/fail benchmarks* are required for making decisions about level of proficiency achieved at a theoretically important point in time (i.e., end of preschool or beginning of kindergarten). Summing up, the SPMM standards, when considered jointly with both the AERA and DEC standards, seek to provide the evidence needed to demonstrate that these progress and mastery measures actually perform as intended when used with young children.

Guidelines for Reporting and Evaluating Assessment and Measurement Research in EI/ECSE

Intersection of the three standards for assuring quality in measurement design and validation. We now turn to some of the commonalities shared by the intersection of the standards reflected in Table 1 and provide a list of measurement research guidelines (Table 2). Taken together, this information helps inform authors preparing manuscripts for *JEI*, and reviewers of such manuscripts may evaluate what evidence or findings need to be included in manuscripts, given their intended purpose(s). For example, as shown in Table 1, AERA Number 1, *domain (or outcome) specification*, intersects with *acceptability* (DEC), *general outcome social validation* (progress monitoring), and *skill domain social validation* (mastery monitoring) SPMM standards. The nature of the evidence argument used to specify an ability/performance domain or outcome is the issue addressed by all three. The DEC and SPMM measurement development often uses social validation evidence processes as well as theory, and extant research evidence, as a means of selecting the domains and outcomes of importance to key stakeholders in the field. With respect to AERA Number 2, *review of relevant literature to specify behaviors of interest* intersects with the DEC *acceptability*

Table 2
List of Recommended Measurement Research Guidelines

Measurement and psychometric considerations

1-1. Is the theory (classical vs. contemporary) guiding the measurement research clearly stated?
1-2. Is the construct being assessed fully and completely defined? Is evidence of the extent that the measure samples this construct provided?
1-3. Is the process for developing items or sampling procedures and for evaluating the performance of these data gathering procedures fully described?
1-4. Are there clear description and empirical evaluation of the groups or individuals with whom this measure can be used?
1-5. What scores or results can be derived? What evidence is provided indicating that they are trustworthy (reliable, valid)?
1-6. Is the reliability clearly described as a property of the data collected and reported rather than the measure?

Application and use considerations

2-1. Are the procedures used for gathering data clear and acceptable to individuals who will participate?
2-2. Are the purposes(s) of assessment fully specified? What questions can, or cannot, be answered?
2-3. Is it clear who the expected users of the measure are and what training and certification they need to administer and interpret its results?
2-4. Are administration and scoring procedures fully specified? Are variations accounted for, and allowable variations described?
2-5. Is there logical, conceptual, and/or empirical evidence that links use of the measure with improved or more appropriate services and outcomes for children, their families, or the individuals and organizations who serve them?
2-6. Are procedures and metrics for reporting results understandable by and acceptable to parents, teachers, and other consumers of the measure?
2-7. Are the consequences of decisions made based on the resulting data evidence produced clear, appropriate, and acceptable?
2-8. Is evidence for implementation fidelity reported in research using large-scale assessment systems?

Other considerations

Has the developer(s) fully described his or her possible financial or business interest in sale of the measure.

standard and with the identification of important *key skill indicators and skill sequence specification* standards in SPMM.

With AERA Number 5, the *practical utility of the administration formats and scores* may also be described as *acceptable* to practitioners and families, *authentic* because the tasks are not contrived and the informant is known to the child, and *collaborative* because the methods engage teamwork between practitioners and families, the DEC standards. AERA Number 5 intersects with SPMM in terms of evidence that the measure is *feasible for use by practitioners* in terms of the knowledge and understanding required to administer within the time required for frequent administrations to all children. Similarly, with AERA Number 9, *classification reliability analysis and determination of empirical benchmarks* intersect with a measure described as *equitable* because it is able to accommodate individual differences (DEC), and in terms of SPMM it is often demonstrated via *end-of-year benchmarks, rates of improvement* (progress monitoring), and *pass/fail decision cut points* (mastery monitoring). Similarly, with AERA Number 13, *construct validity*, the

DEC and SPMM standards both require demonstration of *sensitivity* to growth over time as an essential source of evidence.

One area in particular in which a DEC standard takes a unique view is with respect to its *convergent* standard evidence for reliability. This DEC standard values the pooling of different informant information to yield a more comprehensive picture, rather than treating disagreement among informants as measurement error, the more commonly held view in AERA and SPMM standards. This DEC perspective appears most related to the large-scale inclusion of parent informants' information and judgments (Suen, Logan, Neisworth, & Bagnato, 1995) and multiple team members (Suen, Lu, & Bagnato, 1993) who may report information differently than teachers or other informants on the same measure.

Taken together, we encourage that authors and reviewers of *JEI* manuscripts designing or reporting measurement research consider these intersections in Table 1 in framing research questions in their work and reporting supporting evidence aligned with the question. For example, in work focused on new measure development, the relevant questions, design methods, and evidence would be reported and evaluated. In work seeking to improve existing measures, these new questions and evidence would be reported and evaluated. Similarly, for work on large-scale measurement systems supporting use of single measures as, for example, the Infant/Toddler Individual Growth and Development Indicators (Greenwood, Walker, & Buzhardt, 2010), or multiple measures as, for example, in statewide accountability systems, related questions and supporting evidence would be reported.

The guidelines in Table 2 provide additional support for key considerations. These guidelines are expressed as questions that any published research might address and are organized into three interrelated categories: measurement and psychometric considerations, application and use considerations, and other considerations. Any particular review or study may address one or more of these questions as well as questions from one, two, or all three of the categories. As research accrues on any particular approach or instrument, however, answers to most (if not all) of these questions should be available to potential users. As a result, these questions, both about the design and earlier-stage research on new approaches to assessment and later-stage investigations of the use and impact of well-established instruments or procedures, can and should guide both the design and reporting of research.

The Ethical Standard Needed to Protect Against Bias and Conflict of Interest

It should also be recognized that measures are developed, reported, and maintained by individuals and organizations with a wide range of interests and intentions as regards intellectual property and commercialization. Consequently, an ethical standard is needed for author(s) reporting findings in *JEI* wherein they disclose any commercial relations or possible conflicts of interest associated with the design, development, funding, and/or dissemination of the described assessment practices.

This need not be an onerous task, nor should the presence of a possible conflict of interest here preclude publication or careful attention from the field; the simple truth is that much of the research in EI and ECSE can and should yield "products" disseminated through commercial or fee-for-service mechanisms and that this dissemination is an impor-

tant part of expanding the capacity of practitioners to better serve children and families. Rather, these potential conflicts of interest should be fully disclosed.

In part, potential conflicts of interest can be managed by careful attention to the *questions asked* and the *methods and analyses used* in any reported research. Editors and readers should in all instances, not only in those in which a conflict might exist, attend carefully to the extent to which the questions undergirding any research report are well supported by theoretical, practical, and prior empirical analyses, and to the extent any report adds to the existing (and important) knowledge base. Methods and procedures should be described in clear, comprehensive ways that allow readers to assess and verify the integrity of each study, and analyses should be open to inspection and further analysis. Conclusions should be carefully and conservatively drawn, based on questions asked and results obtained.

Authors should also fully disclose any possible direct conflicts of interest prominently in any published report. This disclosure can take many forms; one possibility is that

> the work described here is part of the larger development of [name the product]. This intellectual property is or may be licensed or sold as a commercial product, and [name one or more of the authors of this paper] may receive financial gain for products related to the research described in this paper. [If appropriate, say "This relationship has been reviewed and managed by [one's employer] in accordance with its conflict of interest policies."]

The boundaries for when to disclose such a conflict may be blurry, with some descriptive or early-stage research conducted before the commercial potential of any research product is realized. To balance the need to disclose possible conflicts of interest with the obligation to share relevant research with our colleagues, authors submitting articles to *JEI* might begin by "over-identifying" possible conflicts of interest, and then working with the journal's editors to determine when, reasonably, such a disclosure is needed in publication. In this way, editors and authors can assure the integrity (and trust) of research published in *JEI*.

Summary/Conclusion

Our purpose in this article was to describe issues and guidelines for reporting and reviewing measurement research methods and findings for manuscripts submitted to *JEI*. These guidelines are designed to provide authors who submit manuscripts to the journal and reviewers with a uniform set of expectations regarding the designing and reporting of results from measurement development investigations. We readily acknowledge that these expectations may not be entirely comprehensive, complete, or unchanging. This information is intended to be helpful in identifying issues, improving clarity, and resolving common questions that might be addressed in measurement research. Our approach was to examine the intersection of three well-established standards for measurement development and validation as a means to this end because, in our experience, this intersection has not been explicitly addressed or recognized in published reports. Each of the standards is relevant in purpose and details to the assessment and measurement of young children. Our

hope is that this information will provide helpful structure to the author, the reviewer, and the editorial process/discussions around issues of quality in assessment/measurement. We invite your input on the appropriateness of these ideas and the issues that you think should be addressed in future revisions of these guidelines.

References

American Educational Research Association. (1999). *The standards for educational and psychological testing*. Washington, DC: Author.

Bagnato, S. J. (2007). *Authentic assessment for early childhood intervention: Best practices*. New York, NY: Guilford.

Bagnato, S. J., Neisworth, J. T., & Pretti-Frontczak, K. (2010). *LINKing authentic assessment and early childhood intervention: Best measures for best practices* (2nd ed.). Baltimore, MD: Brookes.

Boulware, G., Schwartz, I. S., Sandall, S. R., & McBride, B. J. (2006). Project DATA for toddlers: An inclusive approach to very young children with ASD. *Topics in Early Childhood Special Education, 62*, 94-105.

Bricker, D. (2010). *Assessment, Evaluation, and Programming System (AEPS)* (2nd ed.). Baltimore, MD: Brookes.

Bricker, D., Capt, B., & Pretti-Frontczak, K. (2002). *AEPS® test for birth to three years and three to six years* (Vol. 2). Baltimore, MD: Brookes.

Buysse, V., & Peisner-Feinberg, E. (2009). Recognition and response: Implementation sites in Maryland and Florida. In M. R. Coleman, F. P. Roth, & T. West (Eds.), *Roadmap to Pre-K RTI: Applying Response to Intervention in preschool settings* (pp. 9-10). New York: National Center for Learning Disabilities.

Buzhardt, J., Greenwood, C. R., Walker, D., Anderson, R., Howard, W., & Carta, J. J. (2011). Effects of web-based support on Early Head Start home visitors' use of evidence-based intervention decision making and growth in children's expressive communication. *NHSA Dialog: A Research-to-Practice Journal for the Early Childhood Field, 14*, 121-146.

Carta, J. J., Greenwood, C. R., Walker, D., & Buzhardt, J. (2010). *Using IGDIs: Monitoring progress and improving intervention for infants and young children*. Baltimore, MD: Brookes.

Division of Early Childhood, Council on Exceptional Children (DEC). (2007). *Promoting positive outcomes for children with disabilities: Recommendations for curriculum, assessment, and program evaluation*. Missoula, MT: Author.

Deno, S. L. (1997). Whether thou goest. . . . Perspectives on progress monitoring. In J. W. Lloyd, E. J. Kameenui, & D. Chard (Eds.), *Issues in educating students with disabilities* (pp. 77-99). Mahwah, NJ: Erlbaum.

Dunlap, G., Kern, L., Clarke, S., & Robbins, F. R. (1991). Functional assessment, curricular revision, and severe behavior problems. *Journal of Applied Behavior Analysis, 24*, 387-397.

Early Childhood Outcomes Center. (2011a, August 9). *The Child Outcomes Summary Form*. Retrieved from http://www.fpg.unc.edu/~eco/pages/outcomes.cfm

Early Childhood Outcomes Center. (2011b, August 8). *Outcomes for children served through IDEA's Early Childhood Program*. Retrieved from http://www.fpg.unc.edu/~eco/assets/pdfs/outcomesforchildrenfinal.pdf

Fox, L., Carta, J. J., Strain, P. S., Dunlap, G., & Hemmeter, M. L. (2010). Response to intervention and the Pyramid Model. *Infants & Young Children, 23*, 3-13.

Greenwood, C. R., Bradfield, T., Kaminski, R., Linas, M., Carta, J. J., & Nylander, D. (2011). The Response to Intervention (RTI) approach in early childhood. *Focus on Exceptional Children, 43*(9), 1-22.

Greenwood, C. R., Carta, J. J., Baggett, K., Buzhardt, J., Walker, D., & Terry, B. (2008). Best practices in integrating progress monitoring and response-to-intervention concepts into early childhood systems. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., pp. 535-548). Washington, DC: National Association of School Psychology.

Greenwood, C. R., Walker, D., & Buzhardt, J. (2010). The Early Communication Indicator (ECI) for infants and toddlers: Early Head Start growth norms from two states. *Journal of Early Intervention, 32*, 310-334.

Greenwood, C. R., Walker, D., Hornback, M., Hebbeler, K., & Spiker, D. (2007). Progress developing the Kansas Early Childhood Special Education Accountability System: Initial findings using the ECO Child Outcome Summary Form (COSF). *Topics in Early Childhood Special Education, 27*, 2-18.

Hawkins, R. P. (1979). The functions of assessment: Implications for selection and development of devices for assessing repertoires in clinical, educational, and other settings. *Journal of Applied Behavior Analysis, 12*, 501-505.

Kagan, S. L., Moore, E., & Bredekamp, S. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary*. Washington, DC: National Education Goals Panel.

Ludy, B. T. (2007). *A brief history of modern psychology*. Oxford, UK: Blackwell.

McConnell, S. R., & Missall, K. N. (2008). Best practices in monitoring progress for preschool children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., pp. 561-573). Washington, DC: National Association of School Psychologists.

McEvoy, M. A., Neilsen, S., & Reichle, J. (2003). Functional behavioral assessment in early education settings. In M. McLean, M. Wolery, & D. B. Bailey (Eds.), *Assessing infants and young children* (3rd ed., pp. 236-261). Columbus, OH: Pearson Merrill Prentice Hall.

McLean, M., Wolery, M., & Bailey, D. B. (2003). *Assessing infants and preschoolers with special needs* (3rd ed.). New York, NY: Prentice Hall.

McWilliam, R. (2000). Reporting qualitative studies. *Journal of Early Intervention, 23*, 77-80.

Michell, J. (1977). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

Missall, K. N., Carta, J. J., McConnell, S. R., Walker, D., & Greenwood, C. R. (2008). Using individual growth and development indicators to measure early language and literacy. *Infants & Young Children, 21*, 241-253.

National Center on Response to Intervention. (2011a, May 5). *Mastery monitoring tools standards*. Washington, DC: Author. Retrieved from http://rti4success.org/progressMonitoringMasteryTools

National Center on Response to Intervention. (2011b, May 5). *Progress monitoring tools standards*. Washington, DC: Author. Retrieved from http://rti4success.org/progressMonitoringTools

Neisworth, J. T., & Bagnato, S. (2000). Recommended practices in assessment. In S. Sandall, M. E. McLean, & B. J. Smith (Eds.), *DEC recommended practices in early intervention/early childhood special education* (pp. 17-27). Longmont, CO: Sopris West.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*, 137-148.

Ollendick, T. H., Alvarez, H. K., & Greene, R. W. (2004). Behavioral assessment: History of underlying concepts and methods. In S. N. H. M. Hersen & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioral assessment* (pp. 19-36). Hoboken, NJ: Wiley.

Rous, B., LoBianco, T., Moffet, D. L., & Lund, I. (2005). Building preschool accountability systems: Guidelines resulting from a national study. *Journal of Early Intervention, 28*, 50-64.

Salvia, J., & Ysseldyke, J. E. (2000). *Assessment* (8th ed.). Riverside, CA: Houghton Mifflin.

Sandall, S. R., McLean, M., Smith, B., & McLean, M. (2005). *DEC recommended practices: A comprehensive guide for practical application in early intervention/early childhood special education*. Longmont, CO: Sopris West.

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention, 23*, 145-150.

Squires, J., Bricker, D., & Twombly, E. (2002). *Ages and Stages Questionnaire: Social-Emotional (ASQ: SE)*. Baltimore, MD: Brookes.

Suen, H. K., Logan, C. R., Neisworth, J. T., & Bagnato, S. (1995). Parent-professional congruence: Is it necessary? *Journal of Early Intervention, 19*, 243-252.

Suen, H. K., Lu, C. H., & Bagnato, S. (1993). Measurement of team decision-making through generalizability theory. *Journal of Psychoeducational Assessment, 11*, 120-132.

Wolery, M., & Dunlap, G. (2001). Reporting on studies using single-subject experimental methods. *Journal of Early Intervention, 24*, 85-89.