Article

Hammill Institute on Disabilities

Topics in Early Childhood Special Education
2014, Vol. 34(2) 116–127

DOI: 10.1177/0271121414523652
tecse.sagepub.com
SAGE

Authentic Assessment as "Best Practice" for Early Childhood Intervention: National Consumer Social Validity Research

Stephen J. Bagnato, EdD[1], Deborah D. Goins, MS[2], Kristie Pretti-Frontczak, PhD[3], and John T. Neisworth, PhD[4]

Abstract

The early childhood professionals recognize the limitations of conventional testing with young children. This recognition has given rise to *Authentic Assessment*, now recognized officially as best practice by the major professional organizations. However, no national studies have been conducted to document the comparative qualities of either authentic or conventional approaches—*according to meaningful external standards*. Based on a national Internet survey of professionals, we report details of a social validity study of common measurement methods on eight operationally defined standards for developmentally appropriate assessment linked to professionally sanctioned practice standards and indicators. Approximately 1,500 professional responses reveal higher quality appraisals for authentic assessment over conventional testing methods for early childhood intervention purposes. Based on these results, we offer recommendations to advance valid, sensible, and contextually appropriate assessment for early childhood intervention.

Keywords

authentic assessment, early care and education, early childhood intervention, developmentally appropriate, assessment for developmental disabilities, early childhood special education

The developmentally appropriate alternative to conventional, psychometric testing in early childhood intervention is *Authentic Assessment* (Bagnato, Neisworth, & Pretti-Frontczak, 2010). We cite three major foundations that support the growing development and use of *Authentic Assessment* in our field: professional standards, practice-based research, and social validity survey research.

Professional Standards

The use of Authentic Assessment is championed and institutionalized by the major national professional organizations in their standards and white papers as "best practice" for use in early childhood intervention (Bredekamp & Copple, 2009; Division for Early Childhood [DEC], 2007; National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE), 2003; Sandall, Hammetter, Smith, & McLean, 2005). Furthermore, our professional values and beliefs are refined in books by national task forces (National Research Council, 2008; Schultz & Kagan, 2007; Zaslow, Calkins, Halle, Zaff, & Margie, 2000).

> Authentic assessment refers to the systematic recording of developmental observations overtime about the naturally occurring behaviors and functional competencies of young children in daily routines by familiar and knowledgeable caregivers in the child's life. (Bagnato & Ho, 2006, p. 7)

Practice-Based Evidence and Expert Consensus

Authentic assessment is grounded in practice-based evidence in diverse and community-based early childhood programs and settings by parents and professionals working together to accomplish the major purposes for measurement within early childhood intervention (Bagnato, 2005). The U.S. Department of Education, Office of Special Education

[1]University of Pittsburgh/Office of Child Development, PA, USA
[2]The Pennsylvania State University, State College, USA
[3]B2K Solutions, Hinckley, OH, USA
[4]Behavior Technics Associates, Julian, PA, USA

Corresponding Author:
Deborah D. Goins, The Pennsylvania State University, Educational Psychology, 125 CEDAR Building, University Park, PA 16802-3108, USA.
Email: dzl147@psu.edu

Programs (OSEP), funded the *TRACE Project* (*Tracking, Referral, and Assessment Center for Excellence*) to explore the evidence-base for both conventional testing and alternative assessment strategies to accomplish the major purposes of measurement in early childhood intervention (Dunst, Trivette, Appl, & Bagnato, 2004). Specifically, research syntheses and practice guides were produced for conventional tests and testing (Macy, Bagnato, Lehman, & Salaway, 2007), authentic assessments (Macy, Bagnato, Salaway, & Lehman, 2007), team assessment models (Bagnato, McKeating-Esterle, & Bartolomasi, 2007), clinical judgment/informed opinion (Bagnato, Fevola, Smith-Jones, & Matesa, 2005; Bagnato, McKeating-Esterle, Fevola, Bortolamasi, & Neisworth, 2008), assessments of social and self-regulatory deficits (Ho & Bagnato, 2008), and presumptive eligibility (Bagnato, Fevola, Smith-Jones, & Matesa, 2006; see www.earlychildhoodpartnerships.org).

Contrary to popular wisdom and marketing pronouncements, the general evidence-base for measurement in early childhood intervention, especially for young children with disabilities, is weak and often nonexistent. There is a dearth of applied research on the validity of most measurement practices, particularly conventional tests, to accomplish any purpose in early childhood intervention (e.g., eligibility, individualized program planning, progress/performance monitoring, and accountability). The TRACE research syntheses detail a thorough analysis of more than 1,800 studies; the TRACE analyses identified less than 30 studies with sufficient rigor to provide even minimal support for the validity of current testing practices for use in early childhood intervention; even fewer studies were conducted in real-life circumstances. While still emerging, the evidence-base for authentic assessment encompasses the critical purposes of eligibility determination, program planning, progress/performance monitoring, and accountability in early childhood intervention.

Moreover, renowned national task forces have reached consensus on needed changes in the use of conventional approaches for eligibility determination; their publications have influenced changes in national government policies in special education methodology, specifically. The President's Commission on Special Education used the report of the National Academy of Sciences/National Research Council (2002) to affirm changes in special education eligibility determination practices to highlight greater use of tiered response-to-intervention models instead of conventional testing to gain quicker access to educational support services for high-risk children—". . . the committee regards the effort to assess students' decontextualized potential or ability as inappropriate and scientifically invalid" (pp. 8–23). More recently, other distinguished national task forces have produced texts based on the deliberations of expert panels that have proposed recommended practices for assessment in early childhood relating to various purposes; in each case, the contributed recommendations of the panels needed input specific to early childhood special education, and external national expert consultants (including the first author) were enjoined to provide needed changes and/or to compose rejoinder statements (National Research Council, 2008; Schultz & Kagan, 2007).

Social Validity Research

LINKing Authentic Assessment and Early Childhood Intervention: Best Measures for Best Practices (Bagnato et al., 2010) was written to summarize the professional standards and practice-based evidence to promote early childhood assessment practices which are

> authentic, developmentally-appropriate, and useful for planning and evaluating beneficial experiences for young children, especially those with special needs. This book links professional standards and the evidence-base by evaluating the quality of the current testing and assessment measures with a focus on those measures and systems that are the most authentic and which have advanced practice-based evidence for their use by professionals and parents. (p. 10)

The content of the 4th edition of *LINKing* is based on a unique national consumer social validity process and study. Social validity research and measurement was first proposed by Montrose Wolf (1978) and focused on the acceptability to consumers of applied behavior analysis interventions. Arguably, social validity measurement is extremely valuable and greatly undervalued and underused (Foster & Mash, 1999; Schwartz & Baer, 1991; Turan & Meadan, 2011). *Social validity* refers to the acceptability of and satisfaction with an intervention or assessment procedure, gained through soliciting the judgments of individual consumers, participants, and implementers of the procedures (e.g., parents, children, and professionals).

Gresham (1983) was one of the first researchers to define the importance of social validation in research on assessment, specifically, of social competence in children, and to establish standards for social competency. His research underscored the critical importance of defining the real-life or "authentic" content or outcomes of the assessment based on what features were viewed as important to parents, professionals, children, schools, and agencies in their daily lives.

Based on this logic, LINKing created a social validity methodology and conducted an initial national consumer social validity study in 2008 to 2009 (n = 1,083 consumer ratings) to identify the extent to which frequently used assessments and tests in the early childhood intervention fields met professionally sanctioned quality standards based on the overarching concept of "developmental appropriateness."

Our strong belief that it is critical to conduct comparative research like this using *meaningful external criteria* rather than the outdated and, arguably, weak criteria of, for example, concurrent validity in which certain measures ascribed with the dubious status of "gold standards" are the comparison criteria. Rather, in early childhood intervention such meaningful external criteria include alignment with professional standards and even state and federal early learning standards, capacity to fulfill the purposes for measurement in the field, and authentic criteria such as generation of intervention goals or prediction of placement or success in inclusive programmatic settings. *For this LINK research, our meaningful external criteria were extent of adherence to professionally sanctioned practice standards—developmental appropriateness of content and methods.*

The LINKing social validity methodology encompassed six distinct features: (a) operational definitions of eight quality standards for assessment in early childhood intervention as meaningful external criteria (see Figure 2); (b) a multipoint rating survey format of the eight standards (see Figure 3); (c) standardized identification of more than 200 measurement instruments used with young children in early childhood intervention programs; (d) an Internet-based, electronic survey of about 1,000 consumers and their experiences in using the instruments; (e) an expert review panel of national experts to review the results and to also rate the instruments; and (f) a consumer "icon" quality classification system to identify the "best measures for best practices" which received close-up, detailed profiles in the book as a guide for consumers.

The LINK Expert Panel Consensus

For the second stage of the LINKing (LINK) process, the following steps were used in the original study:

1. Selected a national panel of assessment experts in early childhood intervention;
2. Enabled the assessment experts to complete the LINK survey across measures as a "second-pass" to guide the determination of which measures will be included in the book;
3. Facilitated a conference-call roundtable process by which the national experts review collected data and reach general consensus on the classification of the measures on the eight standards and their likely ratings in the book;
4. Conducted a final consensus analysis by the book authors using a combination of the user surveys, expert ratings, and conference-call consensus to apply the final quality classifications on each measure;
5. Applied a "consumer reports"-type of icon and nominal classification based on a 5-point scale (e.g., exemplary, notable, acceptable, marginal, and unacceptable) both to reach agreement on the final designations for each assessment measure and to determine the final group of measures to be included in the book (see Figure 1).

Purposes of the Current Study

The current study has four purposes: (a) to *augment and expand* the original LINKing 2008–2009 research; (b) to impose rigor and reality through a published, peer-reviewed study in comparing consumer quality ratings about conventional tests and authentic assessments according to their relative adherence to the eight operationally defined professional standards for developmentally appropriate measurement; (c) to explore significant and practical hypothesized interrelationships among the LINK standards for each measure as a type of social validation of these quality indicators (i.e., measures with high utility—treatment validity—for planning interventions will have higher acceptability); and (d) to derive implications for best professional practice in assessment for early childhood intervention based on the outcomes of the research.

Method

Sample

In this augmented national study, 1,445 individual consumer social validity ratings were collected from 969 survey respondents from 22 states in the United States who completed the surveys on measures of their choice in an Internet-based, electronic survey. Among the professional survey respondents/raters, there were 329 therapists/specialists, 296 researchers/faculty, 290 administrators/supervisors, 287 lead classroom teachers, 154 itinerant teachers/consultants, 71 others, and 9 classroom assistants. Of the 665 of 969 respondents who indicated their gender, 637 were female and 28 were male. Ninety-two percent were Caucasian. Majority of respondents held current validations/licenses to teach children and children with disabilities from their states. Among those actively serving children, respondents worked most frequently with Preschool Special Education programs in the urban and suburban/small town populations. Average years of experience was 17.56 for the respondents serving all young children and slightly less for respondents serving young children with disabilities (14.69 years). Ninety-two percent of respondents identified themselves as White/non-Hispanic between 46 and 55 years of age. A majority of the respondents indicated that they worked in a Preschool Special Education programs and a minority of respondents worked in Early Head Start programs (see Table 1).

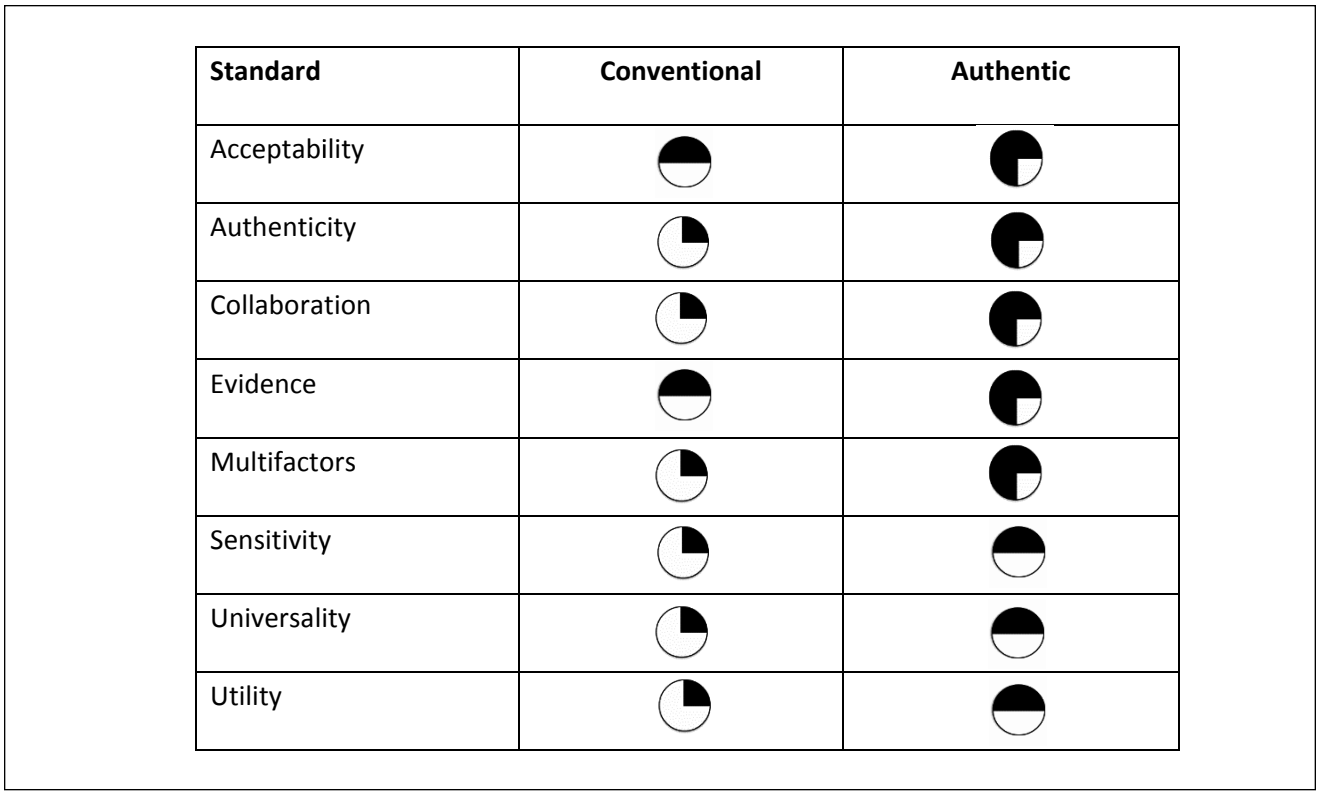

Figure 1. LINK consumer rating "icons" classifying the quality of authentic versus conventional measures.
Note. 100% black = exemplary, 75% black = notable, 50% black = acceptable, 25% black = marginal, white = unacceptable.

Table 1. Roles and Programs of Respondents.

Roles	EI-C	CC	EHS	HS	EI-H	ECSE	PS-Prv	PS-Pub
Lead classroom teachers	49	13	4	21	11	116	19	48
Itinerant teacher/consultant	25	23	6	29	36	45	16	30
Administrator/supervisor	45	36	17	37	57	60	28	54
Classroom assistant	1	0	0	1	0	1	1	1
Therapist/specialist	53	23	9	27	133	79	19	57
Researcher/faculty	32	27	12	25	62	55	31	41
Other	14	15	11	13	14	12	13	16
Total	219	137	59	153	313	368	127	247

Note. EI-C = EI-Center-based; CC = Child Care; EHS = Early Head Start; HS = Head Start; EI-H = EI-Home-based; ECSE = Early Childhood Special Education; PS-Prv = Private Preschool; PS-Pub = Public Preschool.

Procedures and Formats

Bagnato et al. (2010) selected the measures from publisher catalogs, websites, published user surveys, and discussions with early childhood personnel both at the local and state levels, literature review, and review of assessment databases. For this study, assessment measures were selected based on the following criteria: (a) intention for use in early years of child (birth to 8 years old), (b) accessible to the U.S. public, and (c) capacity of assessment measures for developing individual goals and identifying possible intervention strategies. As a result, the ratings of 80 early childhood assessment measures were included in the national consumer social validity survey for analyzing differences in ratings (see www.earlychildhoodpartnerships.org for a complete table of included tests and individual consumer ratings for each measure which is beyond the space of this article). Of the 80 measures, 61 were classified as authentic assessment measures and 19 as conventional tests. The measures were further subdivided into subtypes; these included three types of authentic assessment measures: Curriculum-Referenced With Norms ($n = 21$),

Curriculum-Referenced (n = 18), and Curriculum-Embedded (n = 23). Seven out of 19 conventional tests were IQ measures. Ratings of assessment measures on the operationally defined eight LINK standards used a multiple-choice, 5-point rating scale ranging from 1 (*Unacceptable*; for example, most items focus on competencies not considered worthwhile) to 5 (*Exemplary*; for example, most items identify competencies judged as worthwhile, appropriate, and important for young children's development; see Figure 1).

Figures 1 to 3 outline, describe, and illustrate the operational definitions, scaling, and quality classification for the 8 LINK standards for developmentally appropriate assessment which formed the basis for creating the multiple-choice LINK Internet-based, consumer social validation survey. The eight LINK quality standards for developmentally appropriate assessment in early childhood intervention were created as descriptive categories extracted from a subsumed analysis of the 43 individual assessment competencies in the DEC Recommended Practices manual (Neisworth & Bagnato, 2005; Sandall et al., 2005), created through a focus-group expert panel process providing the content validity of the resulting competencies. Thus, the eight LINK standards have content validity as overarching categories which encompass these individual DEC practice-competencies and serve effectively as quality indicators for appraising assessment methods and procedures for their "goodness of fit" with early childhood intervention purposes and practices.

Once the survey was constructed, it was placed on an online Internet portal at Kent State University for a 9-month period in 2008 to 2009 in which respondents received an individualized access code and log in to the site to provide demographic information and to read about and then select among the various measures that they often used to complete the rating survey.

The total number of ratings received was 1,445. However, two responses were missing data in which respondents only began the online survey session. To be considered a valid response, the survey entry had to include at least one complete set of ratings for one of the eight LINK standards. After eliminating the two invalid responses, the total number of responses included in the analysis was 1,443. Pairwise deletion was used to exclude cases with missing data in the analyses. Using an alpha value of .05 for all statistical analysis, the units of analysis were independent ratings to determine the following differences and correlations between types of assessment measures.

For the purpose of this study, "incidents of judgment" were used as the units of analysis and not the rater because the central focus of the study was in the array of ratings of specific measures (not the professionals, particularly). In such a case, all incidents were treated as independent without regard to who was the rater. As each evaluation is a unit of analysis, having the same respondents evaluating several instruments is not a relevant consideration. Our analysis evaluated the instruments based on the ratings of respondents to evaluate the relative quality of authentic and conventional measures and the interrelationships among standards.

A MANOVA was used for two reasons: The eight LINK standards were correlated ranging from −.01 to .74, and control was needed for experimental-wise error when comparing authentic and conventional measure ratings. To address unequal sample sizes (273 ratings of conventional measures and 1,170 ratings of authentic measures), Box's M test was used to test for significant differences across levels of the eight LINK standards while applying a conservative alpha of .001 for additional control for Type I error. As a follow-up analysis for overall authentic and conventional measures, independent t tests were used as all "incidents of judgment" were treated as independent without regard to raters. Furthermore, due to the nested design, the treatment of each rating as independent leads to smaller standard errors which may result in great Type I error than it may appear.

As measurement ratings have a continuous underlying scale, interrelationships among LINK standards were analyzed using Pearson's product–moment correlation. Finally, significance of differences in overall ratings among subtypes of authentic and conventional measures was tested using one-way ANOVA to control for inflation of Type I error due to repeated tests. To test for significant differences among subtypes of authentic and conventional measures across eight LINK standards, Dunnett's test was used for post hoc analysis because there was a violation of homogeneity of variance assumption.

Data Analysis and Results

Ratings of Respondents

With the primary role or title at a program or agency as the grouping variable (see Table 1), there were no notable or significant differences among the demographic variables in relation to ratings of authentic and conventional measures. Fifty-two percent of the respondents indicated that they used the rated assessment more than 6 times per year. Also, 65% of the respondents indicated that they used the assessments with the following populations: typically developing children, children at risk, and children with disabilities. Thirty-three percent of the respondents used the assessments with all three populations. The majority of the respondents reported that they used the rated assessments for the following purpose—individualized programming to monitor children's progress; 79% found the assessments to be useful, appropriate, and meaningful. Twenty-one percent of respondents indicated the primary reason for using the assessments was that it was required for eligibility determination; 13% reported that their primary reason was because assessments were valid and reliable.

STANDARDS & QUALITY INDICATORS	DEFINED PRACTICE CHARACTERISICS
ACCEPTABILITY	***Social validity; social worth or appropriateness of the scale's item content as perceived by parents and other caregivers***
Social competencies	Emphasizes socially valued and relevant content
Social detection	Yields socially noticeable changes in functioning within real-world settings
Social appropriateness	Uses assessment procedures acceptable to parents and other important caregivers
AUTHENTICITY	***Extent to which the assessment content and methods sample naturally occurring behaviors in everyday situations***
Functional content	Emphasizes competencies that are necessary for the child to participate effectively in daily life activities and routines
Observational methods	Relies only upon "in vivo" observations and reports of familiar people to document child competencies
Natural situations	Captures information on child competencies in familiar classroom, home, and community settings and routines, including play
COLLABORATION	***Parent-professional and interdisciplinary teamwork***
Interdisciplinary procedures	Uses procedures that encourage different models of teamwork (e.g., interdisciplinary, transdisciplinary) and role-sharing among parents and professionals
Family/culture-centered practices	Enables the integral engagement of parents, family members, and friends via "friendly" jargon-free materials and procedures, and practices that respect and align with cultural values; among which the family and partners can voice a preference
EVIDENCE	***Has a clear evidence-base for use in early childhood intervention; materials designed, developed, and field-validated for young children, particularly those with special needs***
Professional standards	Adheres to the unique philosophy, standards, and practices established by the various professional organizations within the early childhood intervention field (e.g., Division for Early Childhood, Head Start, National Association for the Education of Young Children)
Diversity representation	Incorporates children from diverse cultural, linguistic, socioeconomic, and disability backgrounds in the standardization group, if norm-referenced, and in field-validations
Disability specificity	Provides evidence of a pooled typical/atypical norm group or disability-specific standardization for field-validation samples
Early intervention validation	Shows field-validation studies to demonstrate efficacy to fulfill each identified or targeted early childhood intervention assessment purpose (e.g., eligibility, programming, outcomes evaluation, accountability)
MULTIFACTORS	***Collection of data across multiple methods, sources, settings, and occasions***
Multiple situations	Gathers and records information about children's competencies across diverse places (e.g., classroom, home, community), routines (e.g., group circle, playground, lunch), and situations (e.g., morning, evening)
Multiple persons	Pools data from several familiar caregivers (e.g., parents, family, friends, and professionals) who have attachments to the child and interact with the child during daily events, life activities, and across different settings
Multiple methods	Gathers information through multiple methods (i.e., interview, direct probes, permanent products, observations)
Multiple time points	Incorporates evidence of children's preintervention competencies and performances over several assessment time-points
SENSITIVITY	***Sequential arrangement and density of items in the skill hierarchy and the graduated scoring of children's performance on those items***
Functional hierarchy	Organizes assessment content in a sequence of developmental competencies (e.g., younger to older; easier to harder) and/or known instructional steps (e.g., simple to complex)
Sufficient "item-floors"	Contains a sufficient number of items in an assessment sequence to record even low functional levels and to detect the smallest increments of measurable changes in performance, both quantitative and qualitative
Graduated scoring	Uses multipoint ratings or classifications to record and document the extent and conditions under which competencies are demonstrated
UNIVERSALITY	***Design and/or accommodations, which enable all children to demonstrate their underlying and often-unrealized functional capabilities (i.e., identifies both strengths and limitations)***

(continued)

Figure 2. Operational definitions for the eight LINK standards: Foundation for the consumer survey.

Figure 2. (continued)

Equitable design	Designs assessment items so that any child can demonstrate underlying competence; emphasizes functional rather than topographical content (form) and adheres to universal design concepts (i.e., designed for all including children with disabilities without heavy reliance on adaptations or special design, promotes full integration, acknowledges differences as a part of everyday life) (e.g., gets across the room vs. walks across the room)
Alternate materials	Allows the use of alternate, and often multisensory materials to elicit an individual child's functional capabilities
Alternate responses	Allows alternate ways for individual children to show their competencies despite sensory, physical, behavioral, social-emotional, linguistic, and cultural differences or functional limitations
UTILITY	***Treatment validity; usefulness of the scale and the assessment process to accomplish specific early childhood intervention purposes, especially planning and evaluating interventions***
Curricular linkages	Encompasses assessment items whose functional content can match to curricular competencies as instructional objectives, specifically or generally
Intervention content	Identifies what to teach (i.e., assessment identifies which children need to learn which skills/concepts, and where to begin instruction/intervention)
Intervention methods	Informs how to teach (i.e., assessment provides guidance on instructional strategies that facilitate the child's optimal functioning)
Performance monitoring	Detects changes in performance across skills and concepts during/after intervention

Sensitivity: Sequential arrangement and density of items in the skill hierarchy and the graduated scoring of children's performances on those items

How is assessment content organized?

- ☐ **1. Very little assessment content is organized in developmental sequences and/or in known instructional steps**
- ☐ **2. Falls between statements 1 and 3 in meeting the standard**
- ☐ **3. Some of the assessment content is organized in developmental sequences and/or in know instructional steps**
- ☐ **4. Falls between statements 3 and 5 in meeting the standard**
- ☐ **5. Most assessment content is organized in developmental sequences or instructional steps**

Figure 3. Example of the sensitivity standard from the LINK Internet survey.

Three of the most frequently rated *authentic* assessment measures were Ages and Stages Questionnaires®, Third Edition: A Parent-Completed Child-Monitoring System (ASQ-3™); Assessment, Evaluation, and Programming System for Infants and Children (AEPS®), Second Edition; and Ages and Stages Questionnaires®: Social-Emotional (ASQ:SE): A Parent-Completed, Child-Monitoring System for Social-Emotional Behaviors. The top three of the most frequently rated *conventional* tests were the Battelle Developmental Inventory–Second Edition (BDI-2), Developmental Assessment of Young Children (DAYC), and the Bayley Scales of Infant Development®–Second Edition (BSID®-II). Overall, authentic assessment measures received higher ratings than conventional tests. With

Table 2. Differences Between Authentic and Conventional Measures Among the Eight LINK Standards.

	Authentic		Conventional			
Standards	*M*	*SD*	*M*	*SD*	*d*	*p*
Acceptability	3.02	1.11	3.12	1.02		.160
Authenticity	3.04	1.16	2.61	0.84	.425	<.001*
Collaboration	2.86	1.08	2.42	0.96	.431	<.001*
Evidence	2.96	1.16	3.11	1.13		.068
Multifactors	2.86	1.11	2.47	1.07	.358	<.001*
Sensitivity	2.80	1.03	2.73	0.93		.358
Universality	2.87	1.09	2.64	0.88	.232	<.001*
Utility	2.63	0.95	2.32	0.77	.359	<.001*

Note. d = .2 (small), d = .5 (medium), d = .8 (large).
*p < .05.

the exception of classroom assistants (0.60%), all of the roles of the respondents indicated that they used and rated both authentic and conventional measures. Researchers/faculty indicated the highest overall rating of 3.07 on authentic assessment measures. Administrators/supervisors indicated the highest overall rating of 2.89 on conventional tests.

Authentic Assessments and Conventional Tests

The average rating of authentic assessment measures was 2.87 (SD = 1.25), and the average rating of conventional tests was 2.67 (SD = 1.24). The overall mean rating and resulting effect size were significantly higher (η^2 = .158—medium effect size) for authentic assessment measures than conventional tests (see Table 2). As there were significant correlations among the LINK standards, a one-way MANOVA was conducted to test for overall differences between authentic and conventional measures on the eight LINK standards. MANOVA results revealed significant differences among authentic and conventional measures, Wilks's λ = .842, $F(8, 1278) = 29.868$, $p < .001$, $\eta^2 = .158$. As there were unequal sample sizes between authentic and conventional assessment ratings, Box's M test revealed that there are significant differences across levels of the eight LINK standards. However, with a high power of 1.000 for MANOVA and a conservative alpha of .001, significant results remained consistent. Therefore, unequal sample sizes were not problematic in determining significant overall differences between authentic and conventional measures on the eight LINK standards.

Authentic Assessments and Conventional Tests Across the Eight LINK Standards

Following significant MANOVA results for authentic and conventional measures, independent-samples t tests were conducted to compare the ratings for each of the eight LINK standards across both authentic and conventional measures (see Table 2). Results indicate that there are significant differences between authentic assessment and conventional test ratings for the following standards: authenticity, $t(522.861) = 6.902$; collaboration, $t(428.476) = 6.527$; multifactors, $t(377.345) = 5.231$; universality, $t(422.227) = 3.584$; and utility, $t(417.389) = 5.280$. The effect size results for the significant differences ranged from small (.232) to medium (.431).

Among significantly different standards, mean ratings were higher for authentic assessment measures than conventional tests. The Authenticity standard had the highest mean rating of 3.04 which reflects the purposeful development of authentic assessments to capture what children may encounter in their everyday lives. The overall everyday appropriateness of authentic assessments allows professionals and parents to use as many sources of information as possible to plan and help children in their development. Figure 1 applies the LINK icons to visually profile the similarities and differences among the authentic assessment measures and conventional tests across the eight LINK standards.

Interrelationships Among LINK Standards

Mean ratings of the eight LINK standards ranged from 2.57 for the utility standard to 3.04, for the acceptability standard. Among the LINK standards, there were significant correlations ranging from r = .04 to .74 (see Table 3). Acceptability and evidence standards had the highest correlation of r = .74. Effect sizes for significant differences across the eight LINK standards ranged from small (.03) to large (.74).

National consumer ratings, buttressed by the expert panel, resulted in a consensus that the most notable hypothesized interrelationships among LINK standards involved the following: Acceptability and Authenticity, Acceptability and Evidence, Acceptability and Sensitivity, Authenticity and Sensitivity, Evidence and Sensitivity, and University and Utility.

Overall, the consumer results indicate that the highest quality instruments are those which have functional content

Table 3. Significant Correlations and Effect Sizes Among the Eight LINK Standards.

Standards	AC	AU	C	E	MF	S	UN	UT
Acceptability	—	.68*	—	.74*	.03*	.58*	—	—
Authenticity				.65*	—	.55*	—	—
Collaboration					.11*	—	—	.06*
Evidence						.60*	.04*	—
Multifactors								—
Sensitivity								.07*
Universality								.51*
Utility								

Note. r = .10 (small effect size), r = .30 (moderate effect size), r = .50 (large effect size). AC = acceptability; AU = authenticity; C = collaboration; E = evidence; MF = multifactors; S = sensitivity; UN = universality; UT = utility.
*p < .05.

Table 4. Significant Differences Among Subtypes of Assessment Measures.

Test comparisons	AC	AU	C	E	MF	S	UN	UT
CRwN and CE	−0.34*	−0.37**		−0.32*	−0.47**	−0.32*	−0.46**	−0.31**
CRwN and CR	−0.30*	−0.37*	−0.36*	−0.37*	−0.33*		−0.28*	
CE and CR			−0.23*					0.32**
CRwN and CONV		0.05*	0.24	0.45**				
CE and CONV		0.53**	0.37**		0.51**		0.67**	0.47**
CR and CONV		0.43**	0.60**		0.37**			
CRwN and IQ		0.53*	0.70**		0.74**			0.48**
CE and IQ		1.00**	0.83**		1.21**	0.58**	0.77**	0.79**
CR and IQ		0.90**	1.06**		1.07**	0.47*	0.59**	0.47**
AGG and IQ		0.89**	0.89**		1.09**	0.49**	0.64**	0.63**

Note. AC = acceptability; AU = authenticity; C = collaboration; E = evidence; MF = multifactors; S = sensitivity; UN = universality; UT = utility; CRwN = curriculum-referenced with norms; CE = curriculum-embedded; CR = curriculum-referenced; CONV = conventional assessment measures; AGG = aggregate of curriculum-referenced with norms and curriculum-referenced.
*p < .05. **p < .001.

and methods which are acceptable (understandable and doable) to parents and professionals; a style of assessment which captures real-life information from everyday settings and routines through natural observations; measures whose evidence-base involves field-validation/norming with diverse children across diverse home, school, and community settings and for various purposes; and utility for intervention involving accommodations for children's functional limitations; and sufficient item density for individualized goal-planning.

Subtypes of Authentic and Conventional Measures

Three major subtypes of measures were analyzed: (a) curriculum-referenced with and without norms (e.g., measures which have generic content that is goal-oriented and teachable but not part of any particular curriculum); (b) curriculum-embedded (e.g., measures in which the content for assessment, goal-setting, and teaching are identical); and (c) conventional tests (e.g., nonauthentic, nonfunctional content, and tabletop testing procedures such as IQ tests with norms).

Results from a one-way ANOVA showed that the ratings of the LINK standards were significantly different for all of the test comparisons (see Table 4). Due to the violation of homogeneity of variance assumption to compare the subtypes across authentic and conventional measures, Dunnett's test was used for post hoc analysis. Dunnett's test revealed significant differences for all test comparisons among types of assessment measures on the eight LINK standards. All types of authentic assessment measures have higher means than conventional measures/IQ tests.

The first set of comparisons revealed significant differences among curriculum-referenced with norms, curriculum-embedded, curriculum-referenced-only types of assessment measures. All of the eight LINK standards were represented in these test comparisons.

When comparing conventional tests with the three types of authentic assessment measures, post hoc analysis indicated significant differences among consumer ratings with a majority of the p values at <.001. Authenticity and Collaboration LINK standards were significant across all of these sets of test comparisons.

Results of comparisons of curriculum-referenced with norms, curriculum-embedded, curriculum-referenced types of assessment measures with IQ tests had the highest set of mean differences ranging from 1.47 to 2.21. Ratings on authenticity, collaboration, multifactors, and utility LINK standards were significantly different.

An aggregate rating of curriculum-referenced with norms, curriculum-embedded, curriculum-referenced types of assessment measures were compared with IQ tests. LINK standards of authenticity, collaboration, multifactors, sensitivity, universality, and utility were significantly higher in quality ratings.

Discussion and Implications for Professional Practice

Synopsis

In general, this national consumer social validation survey demonstrates clearly the preference and arguable superiority of authentic assessment methods versus conventional tests to accomplish the major purposes in the early childhood intervention programs. *Simply, the results distinguish authentic assessments as the developmentally appropriate alternative to conventional tests, based on feedback from actual practice-based evidence by consumers.* Authentic assessments have been designed and developed to accomplish specific early intervention purposes; their emerging body of practice-based evidence also enhances their acceptability in the field.

We hypothesized that there would be significant correlations among the eight LINK standards which underscore the above conclusions. Results of the national consumer social validation study provide evidence for our hypothesized relationships among these LINK standards and underscore specific strengths of authentic assessment applied to early childhood intervention.

Acceptability and Evidence standards had the highest and most significant correlation ($p < .001$). Research has demonstrated that professionals and parents view measures that are understandable and sensible in form, practical in content, and valid and applicable to their own children as most desirable. Professionals and parents identified most authentic assessment instruments as more appropriate and socially valuable for use with their young children with disabilities because the measures were developed through natural field-validation studies.

Acceptability and Authenticity were also highly correlated ($p < .001$) which supports the purpose of authentic measures to capture a more accurate, representative, and easily understood picture of a child's capabilities which matches the observations and impressions of familiar and knowledgeable caregivers in the child's life. As a result, authentic assessments are rated by consumers as more applicable to early intervention purposes as their format and process are more sensible and their content more functional and translatable for individualized curricular goal-planning.

Authenticity and Evidence standards are also highly correlated ($p < .001$). Measures which are validated for specific early intervention purposes during actual community and classroom circumstances are also more authentic in their content and strategies for gathering data and linking to individualized interventions. Measures whose validation is based on representative children with both typical and atypical development are the most authentic. Simply, observations of ALL children in their natural environments lead to more appropriate and valid assessment results that are more applicable for a specific child's capabilities and needs.

Consumer ratings indicated a significant relationship between *Authenticity and Sensitivity* standards ($p < .001$). For a truer representation of a child's development and progress, multiple observations overtime are also important. Authentic assessment measures have a greater number of graduated and hierarchal items available which allow caregivers, professionals, and parents to be more effective in monitoring a child's progress because they encompass a greater number of samples of a child's natural behaviors and capabilities.

Likewise, there was a significant relationship between *Sensitivity and Evidence* standards ($p < .001$). In general, authentic assessments contain a greater density of competencies to more precisely document a child's performance and progress. Also, sensitivity to individual progress leads consumers to perceive authentic measures as more socially valid, acceptable, and practical for use in early childhood intervention ($p < .001$).

In general, Authentic Assessment measures meet the principle of universal design; this reflects the significant relationship between *Universality and Utility* standards ($p < .001$). When an assessment measure enables children to demonstrate their true or unrealized capabilities and competencies through any mode of response, the more acceptable that measure is to consumers as it is highlighting the child's individual strengths and needs; similarly, the more universal a measure, the more useful that measure is for planning and evaluating effective and individualized intervention programs for children—and for forecasting their potential for progress.

Guide Points for Practice and Research

This study revealed significantly higher ratings of authentic assessment measures than conventional tests from those interdisciplinary professionals who work directly on a daily

basis in early intervention programs with infants and young children, particularly those with disabilities and their families. *The primary goal of authentic assessments that meet the eight LINK standards is to be developmentally appropriate in the planning, implementation, and monitoring of early childhood intervention programs in natural environments and contexts for young children and families.*

The following admonitions about developmentally appropriate assessment, using the eight LINK Standards from the national research as an organizing motif, are the essential take-home points or "lessons learned" from this unique national consumer research. We believe strongly that the results of this research must influence professional practice standards, training toward professional competencies, future research, and government policies and regulations:

- Use sensible, economical, and efficient assessment procedures acceptable to both parents and professionals which sample children's skills that are socially valued and noticeable in their social participation in everyday life activities.
- Rely on knowledgeable, informed, and familiar caregivers in the child's life, especially teachers and parents, to observe, record, and capture (e.g., directly or via video and computer-assisted technology) "authentic" instances of children's functional competencies displayed in meeting the challenges of real-life routines.
- Emphasize the use of "down-to-earth" and jargon-free assessments in which both parents and professionals can collaborate as equal team members through family-friendly content and formats that facilitate collaborative decision making about goals and interventions which respect family priorities and cultural values.
- Be a knowledgeable consumer: Use only assessment measures which have a truly valid evidence-base for use in early childhood intervention—assessment content and procedures are developmentally appropriate and have been specifically designed, developed, field-validated/normed for young children, especially those with disabilities.
- Select and use assessment measures which enable the collection and synthesis of multisource information about a child's functioning across people, places, and times.
- Use assessment measures which are organized in a developmental sequence, functional hierarchy, or instructional steps; contain a sufficient number of skills (e.g., early development—later development) and density of items; and rely on a graduated, multipoint scoring system to ensure a complete and sensitive appraisal of a child's competencies and progress despite the severity of their functional limitations.
- Use assessment measures with universal design features (e.g., focus on "functional" skills; individual response modes; multisensory materials) to enable all children to demonstrate their hidden and often unrealized capabilities.
- Use authentic curriculum-based assessment measures which demonstrate a clear and precise link between skills assessed and skills taught and that also align with the program's curricular goals and state early learning standards.
- Recognize that substantial and tangible changes have occurred in the last decade in professional practice standards, professional practices, state regulations and data systems, and the availability of cost-effective and popular commercial assessment products by publishing companies which promote authentic assessment procedures; for example, note the widespread use and emergence of such measures as The Assessment, Evaluation, and Programming System (Assessment, Evaluation, and Programming System for Infants and Children [AEPSi], 2005); Work Sampling System (Meisels, Marsden, Jablon, & Dichtelmiller, 2005); and the recently released integrated system, Riverside Early Learning Assessment (Bracken, 2013).
- Advocate for and engage in future authentic assessment research and product development initiatives which infuse these standards and link their content and methods to emerging portable video and computer-assisted technologies to ensure more accurate and "real-time" appraisals of children's daily functioning.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Assessment, Evaluation, and Programming System for Infants and Children. (2005). *AEPS, AEPSi overview*. Retrieved from http://aepsinteractive.com/overview/index.htm

Bagnato, S. J. (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Intervention*, *28*, 17–22.

Bagnato, S. J., Fevola, A., Smith-Jones, J., & Matesa, M. (2005). *Research foundations for clinical judgment (informed opinion) in early intervention*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Bagnato, S. J., Fevola, A., Smith-Jones, J., & Matesa, M. (2006). *Research support for presumptive eligibility in early intervention: A research synthesis*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Bagnato, S. J., & Ho, H. Y. (2006). High-stakes testing with preschool children: Violation of professional standards for evidence-based practice in early childhood intervention. *KEDI International Journal of Educational Policy*, *3*(1), 22–43.

Bagnato, S. J., McKeating-Esterle, E., & Bartolomasi, P. (2007). *Evidence-base for team assessment practices in early intervention*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Bagnato, S. J., McKeating-Esterle, E., Bartolamasi, P., & Neisworth, J. T. (2008). Valid use of clinical judgment (informed opinion) for early intervention eligibility evidence base and practice characteristics. *Infants & Young Children*, *21*, 334–349.

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: Paul H. Brookes.

Bracken, B. (2013). *Riverside Early Assessment of Learning (REAL)*. Rolling Meadows, IL: Houghton Mifflin Harcourt.

Bredekamp, S., & Copple, C. (2009). *Developmentally appropriate practice in early childhood programs* (3rd ed.). Washington, DC: National Association for the Education of Young Children.

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

Dunst, C. J., Trivette, C. M., Appl, D. J., & Bagnato, S. J. (2004). Framework for investigating child find, referral, early identification, and eligibility determination practices. *TRACElines*, *1*(1), 1–11.

Foster, S. L., & Mash, E. J. (1999). Assessing social validity in clinical treatment research issues and procedures. *Journal of Consulting and Clinical Psychology*, *67*, 308–319.

Gresham, F. M. (1983). Social validity in the assessment of children's social skills: Establishing standards for social competency. *Journal of Psychoeducational Assessment*, *1*, 299–307. doi:10.1177/073428298300100309

Ho, H., & Bagnato, S. J. (2008). *Research foundations for early intervention eligibility based upon deficits in social and self-regulatory behavior*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Macy, M., Bagnato, S. J., Lehman, C., & Salaway, J. (2007). *Research foundations of conventional tests and testing to ensure accurate and representative early intervention eligibility*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Macy, M., Bagnato, S. J., Salaway, J., & Lehman, C. (2007). *Research foundations of authentic assessments ensure accurate and representative early intervention eligibility*. Washington, DC: U.S. Department of Education, Office of Special Education Programs.

Meisels, S. J., Marsden, D. B., Jablon, J. R., & Dichtelmiller, M. (2005). *The work sampling system* (5th ed). Retrieved from http://www.pearsonclinical.com/childhood/products/100000755/the-work-sampling-system-5th-edition.html

National Academy of Sciences/National Research Council. (2002). *NAS/NRC report on minority students in special education*. Washington, DC: Author.

National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8*. Washington, DC: National Association for the Education of Young Children. Retrieved from http://www.naeyc.org/files/naeyc/file/positions/pscape.pdf

National Research Council. (2008). *Recognizing and responding to the developmental and learning challenges of young children early childhood assessment*. Washington, DC: The National Academies.

Neisworth, J. T., & Bagnato, S. J. (2005). DEC recommended practices: Assessment. In S. Sandall, M. L. Hemmeter, B. J. Smith, & M. E. McLean (Eds.). *DEC recommended practices: A comprehensive guide for practical application in early intervention/early childhood special education* (pp. 45–70). Longmont, CO: Sopris West.

Sandall, S., Hemmeter, M. L., Smith, B. J., & McLean, M. (2005). *DEC recommended practices: A comprehensive guide for practical application in early intervention/early childhood special education*. Missoula, MT: Division for Early Childhood.

Schultz, T., & Kagan, S. L. (2007). *Taking stock: Assessing and improving early childhood learning and program quality*. Washington, DC: National Early Childhood Accountability Task Force, PEW Foundation.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis*, *24*, 189–204. doi:10.1901/jaba.1991.24-189

Turan, Y., & Meadan, H. (2011). Social validity assessment in early childhood special education. *Young Exceptional Children*, *14*(3), 13–28. doi:10.1177/1096250611415812

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.

Zaslow, M., Calkins, J., Halle, T., Zaff, J., & Margie, N. G. (2000). *Community-level school readiness: Definitions, assessments*. Washington, DC: Child Trends, Knight Foundation.