

Infants and Young Children
Vol. 17, No. 3, pp. 198-212
© 2004 Lippincott Williams & Wilkins, Inc.

The MisMeasure of Young Children

The Authentic Assessment Alternative

John T. Neisworth, PhD; Stephen J. Bagnato, EdD

Measurement in early care and education, and early intervention, particularly, continues to be dominated by the use of conventional, norm-referenced testing practices to the detriment of young children. Conventional tests have been neither developed for nor field-validated on infants, toddlers, and preschoolers with developmental disabilities. Thus, contrary to professional wisdom in the fields, conventional tests have no evidence-base for use in early childhood intervention. Nevertheless, the accountability movement in education embodied in No Child Left Behind legislation continues to promote the use of conventional tests, which yield distorted results for young children with special needs. It is long overdue for our interdisciplinary fields to *abandon decontextualized testing practices* and to champion the use of measurement techniques that capture authentic portraits of the naturally occurring competencies of young exceptional children in everyday settings and routines—the natural developmental ecology for children. In this article, we present the “authentic assessment alternative” to the mismeasure of young children. We review the purposes for assessment in early childhood intervention; issues related to conventional testing; 8 standards for professional “best practices”; a rationale and examples of the process and methods for authentic assessment; and guidepoints for implementing authentic assessment in action.

Key words: *alternative assessment, authentic assessment, curriculum-based assessment, early childhood education/early intervention, functional assessment, performance assessment*

PEOPLE have been testing and measuring other people for centuries. History is replete with examples of theories and ap-

proaches to test, diagnose, label, and sort people. Measurement has been conducted more often for the benefit of the assessor; certainly the history of racial and ethnic devaluation and relegation illustrates the complexity of (mis)measurement in justifying and validating social biases. We refer primarily to the attempts to assess personal qualities, often vaguely defined, such as character, morality, criminality, talent, and, of course, intelligence. As Gould (1981) asserted, “There are . . . few injustices deeper than the denial of an opportunity to strive or ever hope by a limit imposed from without, but falsely identified as lying within” (p. 28). Misrepresenting children through mismeasuring them denies children their rights to beneficial expectations and opportunities.

Given our history of (mis)measurement, it seems crucial to look closely at what we advance as accurate and worthwhile assessment with young children, especially with

*This article is a tribute to the writings of Stephen J. Gould, *The Mismeasure of Man*, New York, W.W. Norton and Company, 1981.*

From the Pennsylvania State University, University Park, Pa (Dr Neisworth); and the Children's Hospital of Pittsburgh, The UCLID Center at the University of Pittsburgh, School of Medicine, Pittsburgh, Pa (Dr Bagnato).

This article was supported in part by grant funds for faculty activities from US Department of Health and Human Services, Maternal and Child Health Bureau (5T73MC00036-06) Leadership Education in Neurodevelopmental Disabilities; US Department of Education, Office of Special Education Programs—TRACE Center; and The Vira and Howard Heinz Endowments, Pittsburgh, Pa.

Corresponding author: John T. Neisworth, PhD, Pennsylvania State University, 227 A Cedar Bldg, University Park, PA 16802 (e-mail: jtn1@psu.edu).

the impending reauthorization of the Individuals with Disabilities Act (IDEA) and the ongoing implementation of *No Child Left Behind*. We contend that much of conventional early childhood measurement—the methods and materials we studied in graduate school—have been used to diagnose, label, and sort young children, based more on inference and theory, and have resulted in questionable placements and service delivery.

We want to make 5 points very clear so that there is no misunderstanding of our beliefs about early childhood measurement and the implications of those beliefs:

1. Measurement in early childhood education and in early intervention particularly, continues to be dominated by conventional norm-referenced testing practices to the detriment of young children.
2. Conventional norm-referenced tests of development, intelligence, and psychoeducational functioning have been neither developed for, nor field-validated on young children with developmental disabilities; thus, they have no evidence-base in early childhood.
3. Conventional, laboratory-based testing procedures are decontextualized from children's natural, everyday routines, and thus, fail to capture the "true" functional capabilities of young children of varied ability levels.
4. Conventional testing must be abandoned within the early childhood fields for every purpose including screening, eligibility determination, program planning, progress monitoring, and notably, program evaluation outcomes research.
5. Only authentic or other alternative, observational assessment forms that meet current recommended practice standards of the Division for Early Childhood (DEC; Neisworth & Bagnato, 2000) and the National Association for the Education of Young Children (NAEYC; Bredekamp & Copple, 1997), and embodied in the 8 standards (to be discussed) should be promoted in the early

childhood fields (eg, Head Start, early intervention, early care and education).

To set the stage for our description of an alternative to conventional testing, we have organized this article into 4 major sections: (a) purposes for assessment, (b) issues related to conventional testing, (c) guidelines for recommended practices, and (d) the authentic assessment alternative.

Why do we assess young children?

When children are thriving, with no evident problems, assessment is rarely attempted. Of course, routine screening programs and research projects involve assessment of children with typical development, as do accountability efforts to track child achievement. Yet, assessment becomes important or "high-stakes" when problems are suspected or predicted. Thus, the specific purposes of assessment apply to children who have suspected or evident developmental delays and/or disabilities. Within early intervention, we recognize 4 assessment purposes: (a) screening and eligibility determination, (b) individualized program planning, (c) child progress monitoring, and (d) program evaluation.

Screening

Early detection of developmental problems can result from referrals by parents, teachers, physicians, or community-wide screening efforts. Screening is a relatively rapid method for selecting those children who should receive more detailed assessment. Such rapid methods are socially useful because of the economy of time, effort, and cost, but are subject to error—false negatives and false positives. Clearly, false negatives (ie, indications of "no problem") when a problem really exists are the real issue. These instances will not be afforded up-close assessment, and children's problems will go unrecognized.

Eligibility

According to the current system, children who need early intervention services and supports must be evaluated and declared eligible before special help can be delivered.

Eligibility determination refers to this assessment process. Although varying across states, criteria for eligibility are based on the extent of a child's test-identified deviation from norms for typical development. Ordinarily, cut-off scores are used to make decisions about the need for services. At best, eligibility testing documents disability rather than capability and so is a grossly incomplete look at the child. More serious, however, are the several aspects of the conventional testing approach to measurement, which contradict recommended practices (to be discussed).

Program planning and progress monitoring

Measurement plays an important role in outlining the child's individual plan of instruction and therapy. The plan should be based on a range of information about the child's strengths and needs. The child's developmental progress based on his individual plan should be recorded to guide instruction and detect change. Again, conventional measures do not sample curricular content, and thus, are insensitive to the gains that children evidence in a program. For this reason, much controversy surrounds the testing of children in Head Start and prekindergarten programs using conventional tests whose content is incompatible with curricula and whose testing approach is counter to typical early childhood behavior.

Program evaluation

The quality and impact of the child's program should be clear if program evaluation is done effectively. Periodic feedback to teachers, parents, and staff is important for program modifications. Certainly, the importance of accountability driven by state and federal mandates underscores the critical role of program evaluation. Again, linkages between assessment content and instructional content are essential to reveal program outcomes and impact.

What is "conventional testing"?

Throughout this article we refer to "conventional" testing and the authentic assess-

ment alternative. For purposes of this paper, we describe key aspects of conventional testing. These features will be referred to as we present the authentic alternative.

First, conventional tests are *standardized*. The materials provided as stimuli to the tested child are held constant across all children. Many tests come with their own "kit" of toys, blocks, and, pictures that are shown to children. Sometimes, although a kit may not be provided, detailed descriptions of test materials are offered so that the tester can purchase standard materials. In either case, the requirement is that all children will be tested with the same materials, providing "a level playing field." Clearly, use of standard materials eliminates problems related to variance that might result from differences in the materials. "Point to the picture of the dog" might be more or less difficult depending on the size, color, and quality of the pictures. How the child is to respond to the test items is also typically standardized. Pointing, saying, stacking, and picking up items are often specified in the instructions to children.

Many useful assessment and instructional materials that we use are standardized, and that need not necessarily be a problem. For example, many curriculum packages come with a set of instructions on how to use the materials, recommendations on teaching, and how to assess progress. Standardization can be "tight" or "loose." Curriculum materials are typically loosely standardized so that they can be adapted to a variety of child differences. Conventional testing, however, is almost always tightly standardized to prevent variance due to differences in materials and administration procedures. What the tester says is scripted, as well as the sequence of items. There may even be requirements regarding seating, table, and room conditions. Of course, all testers are to "establish rapport," but the criterion for "rapport" is not specified, nor typically attainable in the testing situation. Tight standardization makes sense within the psychometric approach, because control of testing conditions is required for making normative comparisons. To conclude that a 4-year-old child is one standard deviation below

the normative average for her age, we must be confident that the deviation is due to the child's responding, and not nonstandard testing conditions. Likewise, we cannot coach a child to have the "right answers," because that would obviously distort the meaning of the score and norm comparison. The problems avoided by tight standardization are acknowledged when seen from the psychometric perspective and tradition. The central issue, however, is that the standardization *is tailored to and for children of typical development*, and use of such standard materials and procedures with children of diverse special needs makes little psychometric or common sense.

A second hallmark of conventional tests is the procedure for *item selection*. Items are derived through a series of steps, where an initial pool of potential items is reduced until "acceptable" items are identified. The goal of the item selection process is to identify items that discriminate between age (or diagnostic) groups. Usually, a criterion is set where an item is accepted if, for example, 80% of 4 year olds in the norming sample pass it, but only 20% of three year olds pass it. The items distilled in this way are considered "good discriminators." If a child can show the skill or answer the question, then inferences are made concerning capabilities not on the test, per se, but represented by the item. The usual outcome of this distillation process, however, is items that are nonfunctional skills, which we would not want to teach. Standing on one foot, stringing beads, and stacking 4 blocks may separate age groups, but would not be in any functional curriculum. These special items must be "kept secret" and used only for testing. Teaching to the test (not that the items are worth teaching) is clearly taboo because that would destroy any validity or normative comparisons the test might offer. The characteristics and requirements of conventional, psychometrically based testing do indeed create dilemmas for those who must assess children with special needs, and who seek useful information for making worthwhile decisions.

Finally, both the item selection and standardized administration procedures result in

restricted *sampling* of children's authentic behaviors. Professionals who do research certainly recognize the critical role of sampling. We make inferences based on samples; we may make generalizations beyond the sample to a population to the extent which the sample is unbiased and representative. When we assess a child's developmental or behavioral status, we wish, of course, to make inferences about that child's functioning beyond the particular assessment items and situation. We may consider assessment as an attempt to obtain a fair sample of behaviors that will permit us to make inferences concerning that child's population of performances. Both the content of the sample and the sampling plan (how assessment items are selected) come into play for enabling valid inferences. "Even if our conclusion concerns only one individual, . . . the measurements we have taken are only a sample of all that might be made, and in assigning particular values to the measured magnitudes we are making an inference, based on that sample, of what other measurements would yield." (Kaplan, 1964)

At the heart of authentic assessment is the issue of sampling behavior. In authentic assessment, we observe and/or obtain reports about the child's performances in and across natural settings and occasions. Appraisal of the child's developmental skills *as practiced in the child's real environments* cannot be done through "testing" by a stranger at a table with flashcards, blocks, and beads. Clearly, such conventional testing ignores the crucial requirement for valid sampling of behavior that enable inferences about the presence, absence, fluency, and utility of skills. Use of psychometrically selected items administered in decontextualized settings results in biased samples of the child's functioning—samples that often yield results far different from how the child really behaves.

What are the professional standards for early childhood assessment?

Many interdisciplinary professional organizations have published best practice "white papers" on assessment and intervention in

early childhood, among these are the National Association of School Psychologists; The American Speech, Hearing, and Language Association, and the American Occupational Therapy Association. Yet, the most prominent and extensive professional standards regarding assessment and intervention in early childhood are published by 3 major early childhood organizations: The Division for Early Childhood, Council for Exceptional Children (Sandall, McLean, & Smith, 2000); NAEYC (Bredekamp & Copple, 1997); and the Head Start Performance Standards (Head Start Bureau, 2001).

Over the past decade, the authors (Neisworth & Bagnato, 2000) have collaborated with numerous professionals in the field to develop and field-validate the DEC Assessment Standards. The 46 DEC standards, as part of this process, have been categorized into 8 overarching standards to guide developmentally appropriate assessment. The 8 standards that we summarize below are based on 2 fundamentals. First, assessment contexts, content, and procedures must be developmentally appropriate (ie, be harmonious with and responsive to the interests, capabilities, and realities of early childhood). The imposition of school-age practices (themselves questionable) is at odds with the characteristics of all young children, with typical or atypical development.

The second fundamental element for early childhood assessment is sincere and active cooperation with parents (or primary caregivers). When parents are genuine (not perfunctory) team members, we are able to gather information not otherwise available, eg, sleep patterns, social skill/difficulties in community settings, and toileting behavior. Additionally, information provided by parents may challenge professional data. When information discrepancies exist, additional collaborative assessment is needed, rather than a presumption of parent bias. These 2 fundamentals, developmental appropriateness and parent-professional partnerships, provide not only the values basis, but also a pragmatic basis for making authentic assessment feasi-

Table 1. Eight DAP Assessment Standards

1. Utility	Usefulness for intervention
2. Acceptability	Social worth and agreement
3. Authenticity	Natural methods and contexts
4. Equity	Adaptable for special needs
5. Sensitivity	Fine measurement gradations
6. Convergence	Synthesis of ecological data
7. Collaboration	Parent-professional teamwork
8. Congruence	Special design/field-validation/ evidence-base

ble. Below, are the 8 major standards (see Table 1).

Utility

That assessment should be useful seems self-evident, yet much professional time and effort goes into testing, which has little or no use. Often, professionals are required to use materials and procedures that just do not make sense, given the child and situation. Numbers, scores, percentiles, and standard deviations are generated that simply do not inform instruction or guide intervention.

Assessment can, however, be useful in several ways. First, assessment can help us to identify functional objectives/goals for child and family. Second, assessment can sometimes inform us as to a child's preferred ways or styles for learning and interacting. Third, assessment can be useful in tracking and summarizing progress. In brief, assessment, when appropriately done, can tell us what to teach (content/curriculum), how to teach (methods), and if objectives are being reached (monitoring/accountability).

Acceptability

Sometimes the content or methods of assessment may not be acceptable. There may be items or testing demands that run counter to ethnic or cultural preferences or practices. Item content may also sample events and things not typically encountered by a child or his family, yet are used to gauge cognitive status. Assessment is acceptable when professionals and parents agree on its content and

methods and when information generated by it portrays socially detectable and socially valued competencies.

Authenticity

When assessment is “authentic,” it yields information about functional behavior in children’s typical/natural settings, what they really know and do. The pitfalls of conventional testing, the unfamiliar adult, unrealistic test demands, and nonfunctional item content distilled through psychometric item selection are avoided. Information gathered in authentic settings, within the child’s own developmental ecology, often provides us with a very different picture of strengths and needs.

Equity

Federal law, ethics, and common sense require us to accommodate individual differences in instruction. We certainly would not use the same materials or expect the same responses from a child with typical hearing and the child who is deaf. We would, of course, provide alternate ways to communicate and alternative materials. We would try to provide equitable instruction. Children with significant sensory, motor, affective, and/or cultural differences are not to be expected to relate to the standard “one size fits all” instructional materials and procedures.

What about equity in assessment? Conventional testing, as discussed earlier, requires strict adherence to standardized procedures. All children, regardless of their differences, must be administered test items in the same way as the items were administered to children in the standardization group. And who are in the standardization groups? Recall that the standardization group in almost all cases is composed exclusively of children with typical development. These children can point, say, sit, listen, explain, and tolerate a stranger in ways that children with special needs might not. In fact, children with significant problems are eliminated/excluded from standardization samples. What is wrong with this picture? The procedures are developed for a

smooth administration with “standard children,” but we do not typically need to assess these children. When we are forced to use standard procedures with “nonstandard” children, we have a dilemma: if one does not accommodate for differences, the child is penalized and results are questionable at best; if one alters the procedures, one has “violated standardization!” The crucial issue here is that conventional testing has built in contradictions; it is not equitable, nor can it be. This issue is “high-stakes” when testing “intelligence” or progress within an intervention program.

Sensitivity

Materials that are more sensitive provide more items within an age or skill range. More items permit greater precision for estimating child status and for tracking progress. As an example, a developmental or functional sequence of items related to communication skills will be more sensitive if it includes 100 vs only 10 items. Some materials designed for assessment of more severe impairments must be sensitive in order to gauge minimal functioning and detect small increments of change. Many conventional tests are unsuitable for planning or progress monitoring, especially tests based on traits or constructs, such as intelligence tests; they do not offer adequate item density for detecting change. The need for program accountability argues for the use of tools that can detect intervention-related progress, ie, assessment that is sensitive.

Convergence

Because teamwork is essential for gathering a wide base of information, we need materials that enable us to pool or to converge information. Materials facilitate convergence when professionals from multiple disciplines, as well as parents, can use them. Many new materials provide ways to gather information from multiple sources and to pool evidence. Minimal jargon, friendly formats, and ease of use and reporting are factors that make information convergence more possible.

Discipline-specific tools are, of course, needed and are often important for close-up assessment of specific problems in early childhood education. Such discipline-specific materials are typically not easily converged with other information, and require "translation" for other professionals, and certainly for parents. Curriculum-based assessment tools, on the other hand, are often useable by various professionals, allowing easy pooling of observations.

Collaboration

Authentic assessment requires teamwork; it is simply neither feasible nor sensible for one person to observe and record child functioning across developmental domains, across several everyday settings, and over several occasions. Working in union with assessment partners allows us to assemble authentic information for worthwhile decision making. Finally, we enable parent collaboration when we offer the ways and means for significant parent input. Some assessment materials, for example, provide parent-friendly versions for child appraisal, or we can make available video cameras or other ways to facilitate parent participation.

Congruence

This final standard is actually one that, if followed, somewhat guarantees equity, technical adequacy, and a valid evidence-base. Congruence refers to the similarity of the children employed during a test's development with the children whom you wish to assess. Many new (and some older) materials include children with specific special needs. Clearly, a scale or curriculum designed and developed for children with visual impairments would be a smart choice to use with such children. Some materials can be employed with children with either typical or impaired vision when suggested and alternative administration procedures and items are provided and permitted. When selecting assessment materials, then, it is important to learn about the characteristics of the children involved

in the development and field-testing of the materials. When the child you are to assess is similar to children employed in the development of the materials, you can be somewhat assured that administration procedures will be suitable. Congruence requires that the assessment measure be developed for and field-validated on children with special needs. When a measure is congruent, it demonstrates a clear evidence-base in research.

What is the authentic assessment alternative?

Authentic Assessment refers to the systematic collection of information about the naturally occurring behaviors of young children and families in their daily routines. Authentic assessment is a deliberate plan for investigating the natural behavior of young children. Information is captured through direct observation and recording, interviews, rating scales, and observed samples of the natural or facilitated play and daily living skills of children.

There are 4 major differences between authentic assessment and conventional testing: where it's done, what is assessed, how it's done, and who does it. First, a crucial distinction is the *context* (the where) for assessment. Authentic assessment relies on information that can be obtained only in the child's natural environments. These environments are the ongoing, daily routines, and typical circumstances of the child. Examples of natural environments are children at play in their own preschools, at home during bath time, at childcare, in the supermarket, and at church. This contrasts with the decontextualized, contrived arrangements that characterize conventional, psychometric practices. Conventional testing environments typically employ a clinic or "laboratory" setting such as testing rooms of schools or hospital examination rooms. As pointed out, conventional testing focuses upon standardized item content (the what) and has little instructional use. By contrast, items for authentic assessment are real behaviors that have functional importance to the child and his progress, eg, getting

across the room, communicating wants and needs, selecting an apple rather than a pear, and figuring out how a toy works. Note that these are all competencies that are worthwhile, teachable, and socially valued. Field-validation and norming of assessment instruments for individuals with disabilities must emphasize the standardization of the *function rather than form* of the behavior under examination. Conventional testing records the child's narrow response to standardized objects and procedures and does not permit accommodations for special needs (the how). Authentic assessment relies on natural observations of the child's response to daily routines; in this context, the child can demonstrate competency in any way possible. The child who is blind can show object permanence by exploring the environment tactilely in search of a hidden toy; authentic assessment does not require the child to show only the narrow response of finding and seeing a hidden toy under a standard cup.

Authentic content invites teaching because the items are precursive to or are part of the curriculum. With a functional approach, the playing field for documenting capabilities becomes level and noninferential. Conventional psychometric items are not building blocks for future competency, and psychometric procedures prohibit "teaching to the test" and, thus, are insensitive to functional progress and outcomes.

Only specific professionals, often psychologists, are permitted to conduct traditional, psychometric testing (the who). These professionals are often not integral members of the child's program and are most probably strangers to the children. In most cases, these unfamiliar professionals administer tests as individuals rather than as true team members. On the other hand, authentic assessment depends upon the observations of familiar adults in the child's life to provide convergent data on real-life functioning. An array of family members, babysitters, teachers, and interdisciplinary professionals form a team who knows the child well and works to help the child.

Research support for the authentic alternative

Bagnato, Suen, Brickley, Smith-Jones, and Dettore (2002) published a longitudinal study of the developmental impact and outcomes of an early childhood intervention model for children in high-risk communities. An "authentic assessment and program evaluation" model was employed with 1350 children over a 3-year period of intervention to profile child progress. The Developmental Observation Checklist System (DOCS; Hresko, 1994) was modified so that over 125 early-care and education providers could use it to record natural observations of child skills in everyday classroom routines. A weekly mentoring system was employed to teach the providers the specifics of conducting child quarterly assessments that would inform classroom learning activities, communications with parents, and documentation of child progress. The results of the 3-year study demonstrated the feasibility, utility, and validity of the authentic assessment methodology and the efficacy of the program.

Bagnato and Neisworth* (1995) documented the extent to which traditional tests of intelligence and development are inappropriate and fail to accomplish early intervention purposes for eligibility determination and assessment in the actual daily activities of over 250 preschool psychologists in 33 states with over 7000 children. In summary, the national survey research demonstrated that nearly 60% of the children were or would have been declared "untestable" if the psychologists followed procedures in the test manuals. Major reasons for the child's inability to respond to the tests included behavior at odds with test requirements, lack of language, poor motor skills, poor social skills, and lack of attention and other self-control behaviors. On average, psychologists followed their state

*Bagnato and Neisworth were presented with the 1995 "Best Research Article" award by Division 16 of the American Psychological Association.

requirements to use traditional tests by devoting about 90 minutes to each testing, achieving "untestable" results. After futile effort, however, these expert psychologists were able to work with their teams to declare over 90% of the children as eligible for early intervention services by using alternative and authentic measures to guide their decisions; the appropriate measures used included parent observations and reports, curriculum-based assessments by teachers and providers, play-based assessments, and observations of behavior at home or in the preschool. Clearly, both the required standardized tests and inflexible state regulations served as barriers to appropriate evaluations.

Studies demonstrate that the reliability and validity of assessments and associated decision making for children with disabilities are increased when parents and team members are involved (Briggs, 1997; Suen, Logan, Neisworth, & Bagnato, 1995; Suen, Lu, Neisworth, & Bagnato, 1993). The message of the research is "2 heads are indeed better than 1" when assessing status and progress of young children with special needs. Using generalizability analysis techniques, the researchers demonstrated that each assessor added a unique dimension to the final assessment "portrait" for each child. Additional members on the team were essential to accurate diagnosis while single-member assessments tended to be unreliable and unrepresentative. Similarly, parent congruence with professional assessments should not be required, nor expected because parents add information from natural contexts about important, often neglected, aspects of functioning (ie, sleep routines, play with friends, interactions with grandparents, self-care skills, social and self-regulatory skills with familiar people).

Performance-based or authentic assessment has been studied also for children in kindergarten and the early primary grades. (Meisels et al., 2002; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). The results of these studies demonstrate clearly that ongoing teacher assessments of children's learning within the school curriculum that are struc-

tured by a type of curriculum-based instrument accurately identify and predict those children performing well and those who are at risk. In addition, the CBA instrument (Work Sampling System) enhanced teaching, improved learning through feedback, and boosted scores on conventional group accountability tests of achievement.

The TRACE Center for Excellence in Early Childhood Assessment is a multicenter initiative funded by the US Department of Education, Office of Special Education Programs, to conduct research on the evidence-base for early childhood assessment child find, and referral practices. The second author (S.J.B.) is a coinvestigator in TRACE and director of the Pennsylvania satellite through his Early Childhood Partnerships program (see www.uclid.org). Drs Carl J. Dunst and Carol M. Trivette serve as codirectors and principal investigators of the TRACE Center.

The new TRACE center unites researchers through the 5 satellite locations from the Puckett Institute (Asheville and Morgantown, NC) with colleagues at Children's Hospital of Pittsburgh and the UCLID Center at the University of Pittsburgh (Pittsburgh, Pa), the Institute for Family-Centered Care (Bethesda, Md), the Family, Infant and Preschool Program (Morgantown, NC), and Olson Huff Center (Asheville, NC) and American Academy of Pediatrics (Chicago, Ill).

The 5-year grant will enable the national researchers at the 5 satellite centers to study and promote the use of the most beneficial and promising practices for identifying and evaluating infants, toddlers, and preschoolers who have or are at risk for disabilities or delays. Literature reviews and targeted studies to establish the evidence-base are being conducted currently on a variety of topics including clinical judgment, team assessment models, authentic vs conventional measurement, child find, eligibility determination criteria, and validity of the "delay" concept.

TRACE researchers will collaborate with state and federal government officials to determine how regulatory policies and professional standards can be refined to enhance the

Table 2. Continuum of measurement contexts

	Decontextualized		Contextualized	
	Clinical	Simulated	Analog	Natural
Where?	Laboratory situations	Replica situations	Everyday routines	Everyday routines
What?	Standard responses to standard stimuli	On-demand behaviors	Prompted natural behaviors	Spontaneous behaviors
How?	Psychometric tests	Structured tests and/or observation schedules	Direct observation; parent report; interview	Direct observation; parent report; interview
Who?	Certified/licensed professionals	Certified/licensed professionals	Professionals and caregivers	Professionals and caregivers

use of effective early childhood assessments to identify young children with special needs and to better plan their programs and track their progress.

A continuum of measurement contexts

To organize and illustrate differences among assessment approaches, we offer a continuum of measurement contexts (see Table 2). The continuum encompasses 4 categories: (a) clinical, (b) simulated, (c) analog, and (d) natural. Each category is discussed below.

Clinical

The most decontextualized measurement practices are traditional psychometric tests such as the Stanford-Binet Intelligence Scale (Thorndike, Hagen, Sattler, 2003), the Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 2002), and similar materials, including achievement/knowledge tests and personality measures. "Clinical" refers to highly scripted examiner behavior and the recording of highly scripted or "tested" child behavior in laboratory-like settings. Clinical settings (where) are the most contrived circumstances and bear virtually no resemblance to the natural environments in which young children typically function.

It is important to emphasize that administering a measure such as the Binet in the child's home or preschool classroom does not make the testing more natural or authentic.

The act of using the Binet in the child's everyday settings merely disrupts the flow of the child's natural activity and the contextual supports for behavior. In addition, such measures do not meet the 8 standards for developmentally appropriate assessment previously discussed, especially the requirement for parent participation in the assessment. Previous studies and position statements by the authors (Bagnato & Neisworth, 1995; Neisworth & Bagnato, 1994) demonstrate clearly that traditional intelligence tests in use by even specially-trained preschool psychologists fail to accomplish the major purposes of assessment for early intervention.

Simulated

We are all familiar with the use of testing rooms and clinics where more familiar, typical arrangements are introduced in order to make the environment less threatening. In this approach, concessions are made in favor of the natural environment, but often of a token nature. There is a soft rug on the floor, the testing table and chair are child-size, perhaps there is even a sofa or comfortable easy chair for the parent, etc. In general, the attempt is to have the office/clinic vaguely resemble a child's natural or typical environment. Many pediatric wards, for example, are now decorated to resemble a playful daycare environment, with colorful walls (instead of the sterile white walls of the past), posters, and toys.

Certainly, simulated arrangements are an improvement over the cold and intimidating circumstances that so often are imposed on children and their parents. Knowing what we do about behavior, we all should be clear that the closer the simulation, the greater the chance that typical behavior will be evidenced (ie, enhanced generalization). As noted, however, simulation is often minimal, and the unfamiliarity of the setting and the testing demands trump any efforts to make child and parents "feel at home." The location of the testing situation, often at a testing room off the classroom, clinic, office, or even hospital, is certainly hard to overcome. Further, the time of day for the testing appointment is often driven by administrative convenience and contradicts the child's typical schedule and circumstances. The unfamiliarity of the person doing the assessment is supposedly diminished by "establishing rapport," but how and when that is achieved is not clear, and rarely a concern for the busy professional. A "simulated rapport" is about all that one can hope for in these circumstances.

The materials presented to the child may be modified, again in an attempt to approach the child's natural circumstances. Perhaps the child's toy is substituted for a toy in the testing kit; perhaps the child's friends' names are used in the questioning, etc. Children are prompted to show a skill, or to play with objects provided by the tester.

Increasingly, revisions of standardized instruments are including recommendations for more flexible administration and the use of more child-friendly materials (popular toys). Nevertheless, the materials seldom permit too much variance from "standardized administration," ie, procedures standardized with children of typical development. These and other adjustments are attempted in recognition of the advantages of observing child behavior in context—but such efforts fall far short of observing and recording a child's real behavior in real settings. Arena assessment procedures and the Bayley Scales of Infant Development (Bayley, 1993) are examples of simulated assessment practices.

Analog

Analog practices involve arranging circumstances in the child's natural settings in order to increase the likelihood of occurrences of typical behavior. Daily settings are the contexts (where) for analog assessments, but behavior is prompted or motivated by the choice of toys and materials in these settings to create increased occasions for behavior to be observed. Analog arrangements reduce the need for, sometimes, protracted observation. Analog assessments, again, use developmental observations as the preferred mode (how) of examining and recording the display of these prompted play skills. Assessments, associated with the analog practices, include the Communication and Symbolic Behavior Scales-Developmental Profile (Prizant & Wetherby, 2002), the DOCS, and Transdisciplinary Play-based Assessment (Linder, 1999).

Natural

Assessment practices that are natural in character are distinguished in 3 ways. The "what" (content) refers to only naturally occurring behaviors; the "where" (context) refers to natural environments, ie, the child's everyday circumstances. Finally, the "how" refers to natural observations of the play and learning behaviors of children. Examples of natural or authentic assessment methods include functional behavior assessment, and direct observation and recording of behavior. Also, studies have validated the utility of the DOCS (Hresko, 1994) for authentic assessments in early childhood program outcomes research (Bagnato, 2002; Bagnato et al., 2002). Natural or authentic assessments are the most contextualized practices on the continuum. Authentic assessments in the natural context include the Pediatric Evaluation of Disability Inventory (PEDI), the Ages and Stages Questionnaire (ASQ), and the Assessment, Evaluation and Programming System (AEPS).

The authentic assessment advantage

Early childhood specialists can use the continuum of measurement contexts as a guide

to materials and practices. We contend that *context* is the major criterion regarding the evidence-based validity for authentic assessment practices over traditional psychometric testing. Researchers and policy makers have begun to recognize the overriding importance of context. The President's Commission on Special Education advanced recommendations to dramatically alter the purpose and activities of assessment (National Academy of Sciences [NAS], 2002). Referring to significant failures in appropriately identifying and assessing the needs of minority children and individuals with learning disabilities and other serious developmental disabilities, NAS stated that "While an IQ test may provide supplemental information, no IQ test should be required, and the results should not be the primary criterion on which eligibility rests...*the committee regards the effort to assess student's decontextualized potential or ability as inappropriate and scientifically invalid* [italics ours]" (NAS, pp. 8-23). The committee recommended more authentic alternatives including structured team decision-making processes, and the observational recording of changes in the child's skill acquisition during participation in tailored interventions with individualized instructional modifications.

What are guidelines for authentic assessment in action?

It is one thing to propose a conceptual model for the authentic assessment alternative to conventional testing, but quite another to demonstrate how it can be put to action. In the following section, we discuss considerations and recommendations about how authentic assessment can be implemented in real-life circumstances.

Share assessment responsibilities with a team

Assessment in early childhood must be a team enterprise, especially for children with serious disabilities. Professionals in early intervention must be committed to the worth

and advantages of sharing assessment duties with parents, other important caregivers in the child's life (eg, grandparents, babysitters), and various interdisciplinary team members. Teachers and childcare providers are perhaps the most important sources of information about child performance along with the parents.

Conduct assessment over time

Professionals often wonder how they can do authentic assessments when they require time to observe the naturally occurring behaviors of young children with disabilities, some of whom have limited and inconsistent skill repertoires. We recommend that the team consider using a combination of natural and analog practices to capture child capabilities; moreover, the assessment can be spread over a 15- to 30-day time frame for multiple children in the case of conducting assessments for eligibility determination. In this way, multiple observers can use the same or similar observational measures such as the DOCS (Hresko, 1994) and the Pediatric Evaluation of Disability Inventory (Haley, Coster, Ludlow, Haltiwanger, & Andrellos, 1992) to record spontaneous and/or prompted behaviors across several everyday settings.

Become the "orchestrator" of authentic assessments across people, contexts, and occasions

The team leader can frame his/her primary role as that of the orchestrator of authentic assessments and a facilitator of parent-professional decision making. The assessment process can be organized so that everyone's role in collecting unique or corroborative information about child capabilities and needs is clear. Similarly, an organized assessment process can ensure that comprehensive and representative information on child performance is collected across several people, situations, and times. Parents can observe and record unique information about the child's self-care, temperament, play, and sleep habits at home and with relatives and friends.

Match the team assessment model to the child

The advantages of interdisciplinary and transdisciplinary styles of teamwork for early intervention have been explored previously (Bagnato & Neisworth, 1991; Briggs, 1997). Similarly, the incompatibility of the multidisciplinary approach for early childhood has been documented. We recommend that early intervention teams use a process of collaborative decision making to determine the style of teamwork that is preferred by both parents and the involved professionals and that match the functional needs of the child. For example, authentic assessments in an analog context can be completed for a child with cerebral palsy by having the teacher, physical therapist, and parent prompt play behaviors in the home or classroom, and by observing how the child communicates, plays with toys, and moves about in natural activities.

Rely on parent judgments and observations

Despite best practices and regulations, many professionals are still leery about relying on parent information, suspicious that it is biased and unreliable. In fact, however, parent involvement in the assessment process offers important information about the child's emerging skills, other subtle and inconsistently expressed attributes, and descriptions of critical aspects of temperament or behavioral style. As parents partner with professionals over a longer time period, trust develops and the developmental observations of each become more attuned.

Select a common instrument to unify interdisciplinary and interagency teamwork

Interdisciplinary professionals on a team can choose a common measure such as a curriculum-based instrument to unify the work of the team in the assessment process. A system such as the Assessment, Evalua-

tion, and Programming System (AEPS; Bricker, Cripe, & Slentz, 2003) enables professionals and parents to survey all aspects of development and behavior in an organized manner while focusing on a common core of functional objectives for both assessment and programming. In addition, we recommend that Part C and Part B lead agencies can facilitate transition and planning partly through the use of a common authentic assessment tool such as the AEPS or other such scales.

Employ jargon-free materials

Esoteric professional jargon prevents clear communication among parents and professionals. The use of assessment tools with clear content and objectives (ie, "finds the correct toy at the bottom of a toy box," "lets you know what he means," or "uses eyes and hands together") allows all team members, including parents, to understand and agree on objectives.

Use sensitive instruments to gauge child progress

Most often, conventional measures of intelligence and development lack the sensitivity to profile the status and progress of children with significant functional limitations. These scales lack a sufficient density of items at lower levels (eg, floors); their normative samples do not include children with disabilities and thus generate standard scores, which do not describe lower functional levels (eg, lowest attainable score is 50).

Team members can choose authentic curriculum-based measures that have been field-validated on children with disabilities and have the needed equity and sensitivity features. Various techniques can be used to score and profile child status and progress, including graduated scoring options (1-7 scale) to capture the stimulus conditions under which performance is enhanced; ratio quotients; functional ages; growth curves; goal attainment scaling, and hierarchies of

skill acquisition (Bagnato, Neisworth, & Munson, 1997).

Use technology to facilitate authentic assessments and progress/program evaluations

Increasingly, teams are using various forms of technology to conduct authentic assessments in an efficient way. Videotaping of the child and parent's interactive behaviors at home or in the community is a common way for a team representative to capture functional information for view by the entire team. Such video samples on DVDs can be recorded to compare past performance and observable progress that is indisputable.

Computer software packages enable interdisciplinary teams to conduct an itemized performance analysis of data on individual children and groups on authentic assessment instruments. These programs (eg, Bagnato, 2002) not only generate individualized goals and learning experiences to be infused into daily care and education routines, but also can profile developmental growth curves from authentic assessment data. Analyses can be performed on the archived data to document child progress, intervention outcomes, and program impact for teachers, parents, com-

munity stakeholders, and funding and oversight agencies.

CONCLUSIONS

Early childhood education and intervention have become national priorities. The humanitarian as well as cost benefits of early intervention are increasingly recognized, resulting in renewed political attention and financial support. Assessment plays a pivotal role in the design, delivery, and evaluation of early childhood intervention programs. Use of decontextualized conventional testing must be relegated in favor of in-context authentic assessment. Observation and reporting of actual achievement, rather than inferences about competence based on global trait-based testing, provide real evidence of real child progress and program impact. The time has come for measurement that brings professionals and parents together; that directly informs intervention; and that enables professionals to record functional child progress and to document the success of programs. We must promote assessment that serves the social and humanitarian interests of children, professionals, and those who fund and support the education of all young children, especially those who are most vulnerable.

REFERENCES

-
- Bagnato, S. J. (2002). *Quality early learning: Key to school success*. Pittsburgh, PA: Early Childhood Partnerships, Children's Hospital of Pittsburgh and the Heinz Endowments.
- Bagnato, S. J., & Neisworth, J. T. (1991). *Assessment for early intervention: Best practices for professionals*. New York: Guilford Press.
- Bagnato, S. J., & Neisworth, J. T. (1995). A national study of the social and treatment "invalidity" of intelligence testing in early intervention. *School Psychology Quarterly*, 9(2), 81-102.
- Bagnato, S. J., Neisworth, J. T., & Munson, S. M. (1997). *LINKing assessment and early intervention: An authentic curriculum-based approach*. Baltimore, MD: Paul Brookes Publishing.
- Bagnato, S. J., Suen, H., Brickley, D., Smith-Jones, J., & Dettore, E. (2002). Child developmental impact of Pittsburgh's Early Childhood Initiative (ECI): First-phase authentic evaluation research. *Early Childhood Research Quarterly*, 17(4), 559-580.
- Bayley, N. (1993). *Bayley scales of infant development* 2nd ed. San Antonio, TX: Psychological Corporation.
- Bredekamp, S., & Copple, C. (1997). *Developmentally appropriate practice in early childhood programs* (Rev. ed.). Washington, DC: National Association for the Education of Young Children.
- Bricker, D., Cripe, J., & Slentz, K. (2003). *Assessment, evaluation, and programming system (AEPS)*. Baltimore, MD: Paul Brookes Publishing.
- Briggs, M. H. (1997). *Building early intervention teams*. Gaithersburg, MD: Aspen.
- Gould, S. J. (1981). *The mismeasure of man*. New York: W.W. Norton & Co.
- Haley, S. M., Coster, W. J., Ludlow, L. H., Haltiwanger, J. T., & Andrellos, P. J. (1992). *Pediatric evaluation of disability inventory*. Boston, MA: PEDI Research Group.

212 INFANTS AND YOUNG CHILDREN/JULY-SEPTEMBER 2004

- Head Start Bureau. (2001). *Head start program performance standards and other regulations*. Washington, DC: U.S. Department of Health and Human Services, Administration on Children, Youth, and Families.
- Hresko, W. (1994). *Developmental observation checklist system*. Austin, TX: PRO-ED.
- Kaplan, A. *The conduct of inquiry*. San Francisco, CA: Chandler Publishing Company; 1964:239.
- Linder, T. (1999). *Transdisciplinary play-based assessment*. Baltimore, MD: Paul Brookes Publishing.
- Meisels, S. J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D., & Son, S. (2002). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *American Educational Research Journal*, 39(1), 3-25.
- Meisels, S. J., Bickel, D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teacher judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, 38(1), 73-95.
- National Academy of Sciences. (2002). *NAS/NRC Report on minority students in special education*. Washington, DC: National Research Council/National Academy of Sciences.
- Neisworth, J. T., & Bagnato, S. J. (1994). The case against intelligence testing in early intervention, *Topics in Early Childhood Special Education*, 12(11), 1-20.
- Neisworth, J. T., & Bagnato, S. J. Recommended practices in assessment. In: Sandall, S., McLean, M. E., Smith, B. J., eds. *DEC Recommended Practices in Early Intervention/Early Childhood Special Education*. Longmont, CO: Sopris West; 2000:17-27.
- Prizant, B., & Wetherby, A. (2002). *The communication and symbolic behavior scales-developmental profile*. Baltimore, MD: Paul Brookes Publishing.
- Sandall, S., McLean, M. E., & Smith, B. J. (2000). *DEC recommended practices in early intervention/early childhood special education*. Longmont, CO: Sopris West.
- Suen, H. K., Logan, C. R., Neisworth, J. T., & Bagnato, S. J. (1995). Measurement of team decision-making through generalizability theory. *Journal of Psychoeducational Assessment*, 11, 120-132.
- Suen, H. K., Lu, C. H., Neisworth, J. T., & Bagnato, S. J. (1993). Parent-professional congruence: Is it necessary? *Journal of Early Intervention*, 19(3), 243-252.
- Thorndike, R. L., Hagen, E. P., Sattler, J. M. (2003). *Stanford-Binet Intelligence Scale* (4th ed.). Professional manual. Itasca, IL: Riverside Publishing.
- Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence* (3rd ed.). Professional manual. San Antonio, TX: Psychological Corporation.