Early Childhood Research Quarterly 27 (2012) 46–54

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

The matter of assessor variance in early childhood education—Or whose score is it anyway?☆

Clare Waterman*, Paul A. McDermott, John W. Fantuzzo, Vivian L. Gadsden

Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104-6216, United States

ARTICLE INFO

Article history:
Received 12 October 2010
Received in revised form 21 June 2011
Accepted 23 June 2011

Keywords:
Early childhood
Assessment
Score variability
Examiner
Hierarchical linear modeling

ABSTRACT

Useful assessment outcomes (as manifest through assigned scores) must show reasonable variation across children because it is that variation that presumably defines children's individual differences. Alternatively it is conceivable that some portion of the variability in assessment outcomes does not reflect child differences but rather differences in the performance of the assessors who carry out assessments. Hierarchical linear modeling is applied in this article to identify the amount of score variation attributable to assessors rather than children. Working with multiple cohorts of Head Start and kindergarten children, score variation is analyzed for measures administered outside of classrooms by extramural assessors and for teacher-administered measures within classrooms. The amount of assessor variance (vs. actual child variance) was negligible as associated with extramural assessors but substantial for teacher assessors, indicating that large portions of the variability in teacher-administered assessments have nothing to do with children's unique performances. Recommendations are provided to assist the interpretation of assessment outcomes in future research and practice.

The significance of quality early childhood assessment has increased substantially over the past decade. The increase is the result of a nationwide call for enhanced accountability for schools and government education programs dependent on taxpayer funding (National Education Goals Panel [NEGP], 1998; National Research Council [NRC], 2008; U.S. Department of Health and Human Services [USDHHS], 2003; U.S. Department of Health and Human Services, Administration for Children and Families [USDHHS/ACF], 2010). Substantial emphasis is now placed on having children *ready for school* by entry into kindergarten (Head Start Act, 2007; Shepard, 1997) — an imperative especially important for low-income, urban children who are most at risk for poor educational outcomes. Whereas the assessment of young children has traditionally focused on tracking child development milestones and providing instructional guidance for early childhood educators (NEGP, 1998), assessments are increasingly being used to document children's skill development, to hold programs accountable for child outcomes, to identify children in need of special services, and to certify children as ready for school (Head Start Act, 2007; Love, 2006; NRC, 2008; Shepard, 1997; USDHHS/ACF, 2010).

It is generally understood within the assessment community that tests are created for succinct and limited purposes. Some purposes are restricted to teacher classroom use (e.g., teacher monitoring of child growth, guiding teacher instruction), while others carry the weight of higher-stakes decisions (e.g., program accountability, identification of children with special needs) (Love, 2006; NRC, 2008; Shepard, 1997; USDHHS/ACF, 2010). It is fundamental that the application of an assessment matches its intended purpose (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 1999; NRC, 2008). Yet, mismatches between assessment applications and intended purposes are commonplace (Perie, Marion, & Gong, 2007; Wang, Waterman, Perie, & Marion, 2010) and operate to undermine critical and costly educational processes. The recognized standards for the psychometric integrity of an assessment (e.g., requisite reliability and relevant validity) will vary depending on the purpose of the assessment; a position that has been highlighted by a recent set of guidelines put out by the National Research Council (2008) for assessment in early childhood. In general, the reliability of a measure refers to the consistency or repeatability of scores obtained by an assessment. Validity is indicative of whether or not the assessment is measuring what it was intended to measure and the appropriateness of

☆ This research was supported by the U.S. Department of Health and Human Services' National Institute of Child Health and Human Development, the Administration for Children and Families, the Office of the Assistant Secretary for Planning and Evaluation, and the Head Start Bureau, by the U.S. Department of Education's Office of Special Education and Rehabilitative Services (Grant Nos. P21 HD043758-01 and R01HD46168-01), and by the U.S. Department of Education's Institute of Education Sciences (Grant No. R305C050041-05).

* Corresponding author. Tel.: +1 860 908 9128.
E-mail address: clarewaterman@gmail.com (C. Waterman).

0885-2006/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2011.06.003

the inferences made based on the results of the assessment. The higher the stakes of assessment, the more evidence there should be for reliability and validity of the measure (AERA/APA/NCME, 1999). That is because the highest level of confidence must reside in decisions having life-altering ramifications for children, teachers, and educational programs.

There is a trade-off between the method of an assessment (e.g., direct assessment, observation, work sampling) and its reliability and validity. The less standardized the assessment, the more difficult it is for resultant scores to be reliable or to maintain focus on relevant phenomena. For example, when teachers are asked to make judgments about a child's performance based on observations made in the classroom, differences in the time of day, relationship to the child, and other factors may influence ratings. In turn, acquired stability and focus place a ceiling on the potential validity of assessment scores. This interplay between the method of assessment and its psychometric properties is one major reason why users need to articulate clearly the intended purpose of an assessment when choosing the method (NEGP, 1998; NRC, 2008). In general, if one intends to make comparisons across children and classrooms, or to make placement decisions or evaluate programs, a sufficiently standardized assessment is necessary – an assessment for which explicit administration and scoring routines are articulated and followed. This ensures that the judgments made about children and programs are accurate and not biased by presentiments of the assessor (NRC, 2008).

There are some significant assessment challenges in measuring the competencies of preschool children (NEGP, 1998; NRC, 2008). Preschool children are not generally capable of completing traditional paper-and-pencil tests and so other assessment methods are applied. Such methods include direct assessment (where a trained assessor presents an individual child with tasks or questions), observations, portfolio (work sampling) assessments, and teacher rating scales (where a teacher reports on typical behavior for a child over a set period of time). There is much written about the relative propriety of each type of assessment for use with young children (NEGP, 1998; NRC, 2008). Recently, many researchers and practitioners have espoused the use of authentic assessment (or work sampling) with young children (NEGP, 1998) and expressed opposition to the use of standardized direct assessments because they question the developmental appropriateness of content or administration procedures for preschool children (Meisels, 1998; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education [NAEYC/NAECSSDE], 2003). Proponents of standardized assessments point to the subjective nature of authentic assessments and argue that high-stakes decisions should be based on more objective measures of child performance (NEGP, 1998; NRC, 2008).

Realistically, there are pros and cons to all assessment methods. Less standardized assessments, such as work sampling, allow for more flexibility and provide detailed information on child development but the lack of standardized scoring protocols makes difficult or implausible any comparisons across children. Standardized assessments (such as direct assessments) allow for easy comparisons across children, classrooms, and programs, but too often lack the capacity to provide real-time, detailed information useful for instructional planning. Furthermore, different types of assessors (e.g., teachers vs. independent or extramural assessors) may be more or less appropriate depending on both the method and stakes of the assessment. It makes little sense to have extramural assessors conduct work sampling assessments or observations intended to guide teacher instruction or track daily growth of children. Likewise, expecting teachers to remain objective while administering standardized assessments in high-stakes environments is unrealistic and may jeopardize decisions based on the assessment results.

Table 1
Overview of type of assessors, method of assessment, constructs assessed, and stakes associated with the assessments evaluated in the current study.

Construct	Type of assessor	
	Extramural	Teacher
Early language and literacy		Observation
Alphabet knowledge	Direct[a]	Direct[a]
Vocabulary	Direct[a]	
Listening comprehension	Direct[a]	
Phonemic awareness		Direct[a]
Early mathematics	Direct[a]	Observation

Note: Entries reflect the method of assessment used by each type of assessor for each construct assessed.
[a] Assessment results are used for higher-stakes purposes.

The present study focused on a potential threat to measuring children's actual skill levels – *assessor variance*. Assessor variance is the amount of variation in preschool children's assessment outcomes that is attributable to the assessor or other factors; it detracts from the amount of variance that is objectively attributable to the children themselves. Put more simply, although ideally all of the information in a child's score should reflect child performance, it must be acknowledged that when an assessment is administered by an assessor, some of the phenomena that make up the child's score will reflect things about the assessor and not the child; this is what we refer to as assessor variance. Although assessor error in early childhood measurement is recognized as an important consideration in assessment, the paucity of empirical studies indicates that this issue has not been systematically addressed to date (NRC, 2008). The present study examines the composition of assessment outcomes (as manifest through summative scores) used for various purposes in early childhood settings. Table 1 highlights the different types of assessors (extramural vs. teacher), methods of assessment (direct vs. observation), and stakes associated with each assessment (lower- vs. higher-stakes) reviewed in the current study for each construct assessed. Seven assessments are used in the study; five direct assessments administered by extramural assessors, one observational measure administered by teachers, and one direct assessment administered by teachers. The externally administered assessments include the Learning Express (McDermott et al., 2009), Test of Early Reading Ability–Third Edition (Reid, Hresko, & Hammill, 2001), Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), Oral and Written Language Scales (Carrow-Woolfolk, 1995), and the Test of Early Mathematics Ability–Third Edition (Ginsburg & Baroody, 2003). The teacher-administered observational measure is the Preschool Child Observation Record (High/Scope, 2003) and the direct assessment is the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002). We apply multilevel modeling to separate child-related and assessor-related score variance and describe the relative proportions of child- and assessor-related score variance across different types of assessors, assessment methods, and testing purposes.

1. Method

1.1. Sample

1.1.1. Participant children

Multiple cohorts of children were enlisted, each comprising the full enrollments of classrooms drawn from the 254 Head Start classrooms operated by the public school system of one of the largest cities located in the eastern United States. For three of the four cohorts, random selection was conducted at the classroom level through the use of a random number generator, where each

Head Start classroom was assigned a random number, reordered accordingly, and selected in ascending order until the target number of classrooms was reached. Cohort 1 ($N = 526$) comprised the enrollments of 40 random classrooms gathered primarily for administration of norm-referenced tests (NRTs) at the close of academic year 2004–2005 (AY0405). They ranged 33–69 months of age ($M = 50.5$, $SD = 6.8$) with 50.9% being female, 71.1% African American, 16.1% Latino, 11.6% Caucasian, and approximately 9.5% English language learners (ELLs) and 7.1% receiving services for special needs. A follow-up subsample ($N = 321$) also was assessed during their 2005–2006 (AY0506) kindergarten year. The sample reduction reflected mainly the fact that Head Start includes 3–5-year-olds, whereas only 4- and 5-year-olds are advanced to kindergarten and 3-year-olds remain in Head Start for an additional year. The follow-up sample aged 57–80 months ($M = 66.4$, $SD = 3.8$), with 53.3% females, 71.7% African Americans, 16.8% Latinos, 10.5% Caucasians, 7.8% ELLs, and 8.1% special needs.

Cohort 2 contained 1667 children constituting the enrollments of 80 Head Start classrooms drawn randomly at the opening of academic year 2006–2007 (AY0607) and mutually exclusive of the classrooms comprising Cohort 1. Ages ranged 35–68 months ($M = 49.8$, $SD = 6.8$), with 51.2% being female, 10.2% ELLs, and 10.4% receiving services for special needs. Cohort 3 ($N = 1426$) was constructed from the same 80 random classrooms as they existed in the subsequent academic year (2007–2008; AY0708) and encompassed all of the Head Start children who had not moved on to kindergarten and those who newly enrolled for AY0708. Ages spanned 35–70 months ($M = 50.1$, $SD = 6.7$), with 50.3% females, 12.7% ELLs, and 9.3% special needs. Cohorts 2 and 3 featured very similar ethnic strata, where approximately 69.2% were African American, 19.1% Latino, 4.5% Caucasian, and the remaining children having varied other ethnic backgrounds.

Finally, Cohort 4 was a general sample ($N = 2840$) of AY0607 kindergarten children who had previously been enrolled in the school district's Head Start program. Child ages ranged 56–86 months ($M = 66.2$, $SD = 3.6$), where 50.8% were male, 67.8% African American, 20.9% Latino, and 5.1% Caucasian. Approximately 58.3% were eligible for free/reduced lunch allocation and 4.8% were eligible for special education services.

1.1.2. Participant assessors

Two types of assessors were recruited; teacher assessors and extramural assessors (extramural indicating outside of and independent from the classroom). As part of a larger study to evaluate the effectiveness of preschool curricula (Fantuzzo, Gadsden, & McDermott, 2010), three teams of *extramural assessors* were hired to conduct individually administered assessments of children. One team was hired for each given academic year, AY0405, AY0607, and AY0708, the teams comprised of 20, 45, and 38 assessors, respectively. The preponderance of hired extramural assessors were undergraduate and graduate students and generally ranged in age from 18 to approximately 60 years (median ages in the mid to late 20s). Nearly 20% of extramural assessors were males and more than 40% were ethnic minorities (primarily African American).

Teacher assessors also provided data pertaining to the participant children. The first group consisted of 85 Head Start teachers in charge of the randomly selected classrooms encompassing AY0607 Cohort 2 children. The teachers had 1–37 years ($M = 9.6$, $SD = 8.1$) teaching Head Start children and 2–44 years ($M = 15.6$, $SD = 10.2$) overall teaching experience. Two additional groups of teachers provided kindergarten follow-up assessments of prior Head Start children. The first group included 165 teachers during AY0506 and the second 506 teachers during AY0607. Per district policy, teaching experience and other demographics for kindergarten teachers were not made available.

1.2. Measures

1.2.1. Externally administered measures

The Learning Express (LE; McDermott et al., 2009) is an individually administered criterion-referenced test mapped to the national Head Start Indicators (USDHHS, 2006) early learning standards in the areas of literacy, language, and mathematics. The LE was administered in four testing waves (October, January, March, and May/June) during AY0607 and AY0708 for the purpose of evaluating the efficacy of preschool curricula as part of a multimillion dollar grant. Two equivalent forms of the LE (Form A and B) were used to reduce practice effects. The LE was calibrated via two-parameter logistic, item response theory (IRT) models and scored applying ex a posteriori Bayesian estimation where $M = 200$ and $SD = 50$. Composite reliability for each subscale (Alphabet Knowledge, Vocabulary, Listening Comprehension, Mathematics) ranges 0.93–0.98, with reliability for child subgroups according to age, sex, ethnicity, dual- versus single-language learner status, and special needs status uniformly >0.90. McDermott et al. (2009) further provide substantial evidence for LE concurrent validity with nationally standardized tests of the same content areas and predictive criterion-related validity with future assessments by classroom teachers.

Four NRTs were administered at the end of AY0405, one from each of the four areas assessed on the LE. The Test of Early Reading Ability–Third Edition (TERA-3; Reid et al., 2001) was used to examine alphabet knowledge and is appropriate for children between 3 years, 6 months, and 8 years, 6 months of age. Reliability of scores has been shown to be substantial and validity is supported through correlations with other measures of achievement and ability. The Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997) was used to assess receptive vocabulary. The PPVT-III has displayed evidence of moderate interrater reliability and high test-retest reliability (>0.90). Correlations between the PPVT-III and other tests of cognitive ability support the validity of the measure. The Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1995) was administered to determine children's listening comprehension skills and is used with children as young as 3-years-old through adolescence. Studies provide substantial evidence for the reliability and validity of OWLS test scores (Carrow-Woolfolk, 1995). Finally, the Test of Early Mathematics Ability–Third Edition (TEMA-3; Ginsburg & Baroody, 2003) was applied. It is appropriate for use with children from 3 years, 0 months, to 8 years, 11 months of age. Research evinces appreciable support for score reliability and validity, with reliability coefficients ranging 0.94–0.95. All four NRTs used in this study are direct assessments that have been normed on nationally representative samples of preschool children.

1.2.2. Teacher-administered measures

Archival data were obtained from the school district pertaining to test information routinely gathered by classroom teachers and used by the district for monitoring student responsiveness to educational programs. These data included test scores from the Preschool Child Observation Record (COR; High/Scope, 2003), as administered by teachers to Cohort 2 Head Start children over AY0607, and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), as administered to Cohort 1's kindergarten follow-up sample over AY0506 and to Cohort 4's general kindergarten sample over AY0607. The COR is a teacher-observation measure that occurs at three times over the course of the school year (fall, winter, and spring) intended to provide a systematic assessment of child development. The COR was used by the district to monitor child progress and individualize instruction in the classroom. Of particular interest for this study were the scores that teachers assigned to children in the areas of Language and Literacy as well as Mathematics. Internal consistency for the

Language and Literacy subtest ranges 0.80–0.85 and Mathematics ranges 0.75–0.88.

The DIBELS is a standardized direct assessment of reading ability for kindergarten through grade six used to make placement decisions about children and monitor child growth. Kindergarteners are administered the DIBELS at three time points during the fall semester. At each administration, children are expected to reach particular benchmarks. Those not reaching benchmarks are provided immediate supplemental enrichment services. In addition, the district used DIBELS scores as a measure of school accountability by measuring the percentage of children in each school that met or did not meet the specified benchmarks. The subareas presented here are Initial Sounds Fluency, Letter Naming Fluency, Phonemic Segmentation Fluency, and Nonsense Word Fluency. As reported on the DIBELS website, test-retest reliability coefficients range 0.72–0.88 and concurrent validity coefficients 0.36–0.70 against pertinent criteria in kindergarten (University of Oregon Center on Teaching and Learning, n.d.).

1.3. Procedure

1.3.1. Assessor training

Per national standards (AERA/APA/NCME, 1999) and leading assessment texts (viz., Aiken, 2003; Sattler, 2001), extramural assessors were trained extensively in the administration of the LE and NRTs. Each assessor underwent approximately 35 h of training which included (1) information on child development and working with preschool children, (2) types of assessments and the role of the assessor, (3) basic psychometrics, (4) essential aspects of the LE or pertinent NRTs (starting and stopping rules, prompts, stimulus delivery and response recording), and (5) practice with peers and with preschool children not otherwise involved in the project. The bulk of the training was focused on becoming proficient at using the assessment tools, during which time the supervisors underscored the importance of rapport building, standard test administration, and accurate scoring.

Whereas the NRTs were administered within a brief time period at the close of AY0405, LE administration continued in four waves across AY0607 and AY0708 each, thus requiring ongoing assessor monitoring and refresher training. Extramural assessors were required to work between three and five days per week during respective testing waves. All testing was undertaken in isolated areas (apart from classrooms) of children's actual Head Start centers, with onsite testing supervisors. Assessors displaying the highest attention to detail were recruited for training in the cross-verification of all answer keys, thus ensuring the accuracy of scoring for each test administered. If it was determined that scoring procedures were not adhered to during a test session, the test protocol was voided. Booster training sessions preceded each successive assessment wave in order to review new item starting-point rules and to reinforce practices that would reduce drift of assessor reliability.

Select teacher assessors were recruited to participate in two- and three-day trainings provided by High Scope and DIBELS trainers in the administration of the COR and DIBELS. Teachers who participated in the trainings were subsequently expected to train additional teacher assessors at their respective schools. As policies for training and retraining were established at the school level, we are unable to report a universal model in the current study. It is our understanding, however, that in addition to initial training on the measures, teacher assessors participated in several professional development days that were dedicated to training and discussions.

1.3.2. Assessor assignment

As noted, children who were administered the LE were tested by extramural assessors over four waves during AY0607 and AY0708, respectively. Assignment of children to assessors was quasi-random. Practical logistics precluded complete randomness in assignments inasmuch as Head Start centers had to be informed in advance of arrivals of assessment teams, isolated testing locations had to be secured, and participant children were often absent when their turns occurred, thus completing testing on alternate days. Nevertheless, multiple assessors were assigned to given centers each day and each child was selected according to the classroom's alphabetical student listing. A given child was tested by one assessor per wave applying a given LE form (A or B), where the form was alternated at each successive wave and an effort was made to assign a different assessor to a given child on subsequent waves. All four LE subscales were administered to a child at each wave during a single testing session that lasted approximately 20 min, but no more than 30 min. This was feasible because the LE is an adaptive test using IRT scaling. The alternating forms served to reduce practice effects.

With respect to administration of NRTs, this was accomplished in May AY0405, where extramural assessors had been trained to apply the Alphabet Knowledge subtest of the TERA-3, Form A of the PPVT-III, the TEMA-3 (for Mathematics), and Listening Comprehension subtest of the OWLS. Matrix sampling assigned assessors to children in a quasi-random manner such that approximately four children in any given classroom received any given test. In total, 168 children were administered TERA-3, 154 PPVT-III, 171 OWLS, and 157 TEMA-3. The matrix sampling arranged that no child was assessed on more than two NRTs, ensuring that testing time never exceeded 15–20 min.

No randomization was asserted for assignment of teacher assessors. Rather, teachers had been assigned to Head Start or kindergarten classrooms by those ordinary mechanisms applied by the school district. They were directed as part of their contractual obligation to evaluate every class member using the COR or DIBELS during specified successive periods over an academic year. This business-as-usual process was ideal for present purposes because it availed a natural perspective on the variability of student test scores as it typically evolves in school practice. No specific details can be given for the amount of time teachers spent on COR and DIBELS assessments. That is because teachers reported a wide variety of approaches to completing the task, including focused evaluations on a given child, opportunistic evaluations of one or several children, performance recording that was essentially in vivo, and recording that was retrospective. Our understanding of these processes is based on a series of 2-h focus groups with 15–20 teachers and our examination of the testing materials, response criteria, and response distributions.

1.3.3. Data analysis

The critical question for this investigation centers on the ability of different types of assessors with different types of assessment tools to capture the relevant score variance that could possibly demonstrate individual differences among preschoolers. Because observed test scores reported for children can reflect distinctly different sources of origin, some of which actually inform something about individual child differences and some of which convey error or are more about who did the assessment or where it was done, it is necessary to untangle as much as possible the relevant information from the irrelevant. Considering the array of test and assessor types presented here, the partitioning of variance is possible to a substantial extent. Specifically, given the circumstance that we have observed multiple assessments of a particular type (e.g., NRTs) provided by a particular assessor, it is feasible through hierarchical linear modeling (HLM) to identify proportions of score variation that are attributable to children themselves versus those who assessed them. We call this latter source *assessor variance* to distinguish it from child variance.

To this end, series of two-level HLMs were constructed, one for each one-time NRT outcome and one for each repeated LE, COR and DIBELS outcome. Each model was an unconditional one-way, random-effects ANOVA model to estimate the amount of variability between children, nested within assessors, and the amount of variability between assessors, per se. Since the purpose here was limited to partitioning score variance rather than explaining that variance, covariates were not employed. Thus, to the extent that one might be inclined to assume that observed test scores are unique indicators of individual child differences (the common assumption in any practice where test scores are used to evaluate the performance of individual children or to render programmatic decisions regarding them), such multilevel models should reveal that all or nearly all of reported test performance is child-level (Level 1) and not assessor-level (Level 2) variation.

It was conceivable, of course, that some of what we term assessor variance is really due less to a given assessor and more to the selectivity that might lead to formation of class enrollments or to the homogenous context of the classrooms themselves. This matter would be exceedingly difficult to resolve entirely, given the intractable distinctions between what is, for example, variability spawned by simply being in a classroom versus classroom variability generated by the teacher who organizes and directs day-to-day classroom activities. Nonetheless, our study was in a position to discover the amount of score variation attributable to extramural assessors versus the classrooms from which children were drawn. This effort follows from the fact that LE score variation was jointly nested within assessors and within classrooms, and at each point in time over a school year. Here, unconditional, multilevel, cross-classification models were estimated. The models estimated the independent proportions of score variability of children nested within assessors and children nested within classrooms, as well as the conjoint between-assessor/classroom variation. Inasmuch as the LE was not administered by classroom teachers, it seems reasonable to posit that any discovery of notable amounts of between-classroom variance (especially as contrasted with between-assessors variance) would point to the relative importance of broader contextual classroom (vs. teacher) differences. Though these models do not directly untangle the amount of variation in children's COR and DIBELS scores that is *classroom* (vs. teacher-as-assessor) specific, they do uncover the amount of classroom variation one would expect for preschool children for early literacy and mathematics constructs. Without a method for directly decomposing the amount of variation in teacher-administered assessments that is due to classroom idiosyncrasies and that which is due to teacher administration, these cross-classification models are a sound way to address the amount of assessor-level variation on teacher-administered assessments as compared to assessor-level variation for assessments given by extramural assessors.

2. Results

Table 2 reports percentages of assessor variance associated with the administration of standardized measures by extramural assessors. These values equal 1 minus the model intraclass correlation (ICC) times 100, where the ICC = assessor-level random coefficient/(assessor-level random coefficient + residual). As illustration, the 1.0% posted for assessor variance with Alphabet Knowledge during Wave 1 testing of AY0607 indicates that an estimated 1.0% of the variation in scores is attributable to difference among assessors rather than children, whereas the complementary 99.0% of score variation is attributable to differences among children. Statistical significance is also indicated for some values. Significance reflects the reliable departure of a related assessor-level coefficient from 0 (0 indicating no assessor effect). Because outcomes of these inferential tests vary as a function of both the

Table 2
Percentages of assessor variance estimated for measures administered by extramural assessors.

Performance area	Learning express								Norm-referenced measures[a]
	Academic year 2006–2007				Academic year 2007–2008				
	Wave 1	Wave 2	Wave 3	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4	
Alphabet knowledge	1.0 (n = 1336)	1.1 (n = 1345)	1.3 (n = 1338)	0.2 (n = 1279)	0.2 (n = 1235)	1.7* (n = 1229)	0.0 (n = 1204)	0.9 (n = 1129)	4.1 (n = 178)
Vocabulary	0.0 (n = 1336)	0.0 (n = 1355)	1.6 (n = 1345)	2.2* (n = 1284)	0.8 (n = 1234)	1.2 (n = 1229)	0.5 (n = 1211)	0.4 (n = 1127)	0.0 (n = 171)
Listening comprehension	3.1* (n = 1336)	2.8* (n = 1349)	2.4* (n = 1339)	1.3 (n = 1279)	1.2 (n = 1236)	2.5* (n = 1230)	2.3* (n = 1208)	2.5* (n = 1130)	2.2 (n = 183)
Mathematics	1.2 (n = 1336)	0.7 (n = 1351)	2.9* (n = 1341)	1.5 (n = 1281)	1.1 (n = 1235)	0.2 (n = 1228)	0.8 (n = 1207)	2.0* (n = 1129)	5.3 (n = 163)

[a] Entries correspond to percentage assessor variance for the Test of Early Reading Ability – 3rd Addition, Peabody Picture Vocabulary Test – 3rd Addition, Oral and Written Language Scales, and Test of Early Mathematics Ability – 3rd Addition, respectively.

* $p < 0.05$.

magnitude of an effect and the differential sample sizes, we advise for present purposes that interpretations be based on the relative magnitude of the reported percentages rather than absence or degree of statistical significance.

The estimated amount of assessor variance for the LE scores is essentially negligible across content areas and time. It averaged 1.3% (*SD* = 0.9) with a range from 0.0% to 3.1%. No consistent longitudinal trends are apparent. The poorest performance was found for Listening Comprehension scores over time, peaking at 3.1% for the first wave of administration. Thus, in the worst case, 96.9% of the variability in LE scores is still conveying individual differences among children, per se.

The last column in Table 2 displays results for the four NRTs administered by extramural assessors. Here, the average assessor variance is slightly higher at 2.9% (*SD* = 2.3) and ranges 0.0–5.3%. The latter figure for TEMA-3 Mathematics performance exceeds the 5% criterion applied by Snijders and Baker (1999) for regarding the magnitude of cluster-level variance as consequential although, even here, the evidence suggests that approximately 94.7% of score variation is still driven by children's individual performances. In the present instance, extramural assessors and supervisors had reported that the TEMA-3's varying administration and stopping rules gave rise to some confusion, and this may be reflected in the detectable differences among assessors.

Tables 3 and 4 recount percentages of teacher assessor variance associated with the COR in Head Start and DIBELS in kindergarten. For both tests, the elevations in assessor variance are dramatic. For Head Start teachers applying the COR (Table 2), average assessor variance is 27.6% (*SD* = 5.8) with a low estimate of 19.7% and high of 34.5%. There appears some evidence of improvement (decrement in assessor variance) by the final assessment, although these values nevertheless indicate that only 75–80% of score variation at year's end is child centered, whereas merely 70% of score variation is child centered throughout the rest of the school year.

Table 4 average assessor variance for kindergarten teachers administering DIBELS is 30.8%, a value comparable to assessor variance among Head Start teachers (27.6%), although the variability among kindergarten teachers is quite distended (*SD* = 14.2). The variability marks a very large performance range from a low of 8.3% for Letter Naming Fluency to a high of 69.4% for Initial Sound Fluency. Also disconcerting is the trend for teacher assessor variance to increase with successive assessments. Focusing on the final DIBELS assessment each year (the assessments one would presumably hope to be the most informative about child differences), results reveal that almost 40% of score variability is unrelated to children's individual differences. Moreover, it would appear that children's scores for sound and phonemic fluency are most threatened by confounds with assessor variance.

We endeavored to separate the score variation for the LE measures that was uniquely associated with classroom context (not teachers, since teachers did not administer those measures) from that associated with the extramural assessors who administered the LE. Table 5 presents those results. It will be noted that the values reflecting between-classrooms variance generally exceed those for between-assessors variance, but the disparities are not large and the average level of between-classrooms variance is only 4.8% (*SD* = 2.0), not rising to the 5% level associated with consequentiality. The combination of assessor and classroom variance averaged 5.9% (*SD* = 1.6%), a value dwarfed by the non-child-oriented variance evident in scores rendered by teacher assessors (i.e., 27.6% and 30.8%).[1]

[1] Variance components models were also conducted and the results compared with those from the unconditional, multilevel, cross-classification models reported here. The results did not differ.

3. Discussion

Multiple direct assessments by persons external to the preschool classrooms, using the LE (a standardized direct assessment), revealed that on average 1.3% of score variation was attributable to assessors rather than to children. Similarly, norm-referenced measures produced only 2.9% assessor variance. In sharp contrast, an average of 27.6% of score variation from Head Start teachers' COR assessments and 30.8% from kindergarten teachers' DIBELS assessments were unrelated to actual child differences. Indeed, the amounts of such error variance increased with successive kindergarten teacher assessments with the most extreme values occurring for measures of sound and phonemic fluency skills.

As we suggested, some portion of what is characterized as teacher assessor variance is no doubt a reflection of overall classroom differences. This follows from the fact that students were not randomly assigned to teachers (classrooms) and classrooms certainly should be expected to manifest some differences for general academic performance. Given the availability of data for the LE measures both across assessors and across classrooms (children not having been randomly assigned to their classrooms), the level of classroom variance was less than 5.0% and the average combination of assessor and classroom variance was less than 6.0%. Thus, to the degree that one can extend the general trend to the case of teacher assessors, there is little room to dismiss the observation that alarmingly large proportions of score variation for teacher assessments have nothing to do with children's individual differences and cannot be explained away as ordinary differences between classrooms. Put simply, why should teachers' assessments of basic academic skills vis-a-vis popular, commercial, evaluation tools be saturated with five times more the amount of error variance characterizing independent direct assessments for the same types of basic skills?

Furthermore, it should be noted that, of the two teacher-administered assessments, the DIBELS, a direct assessment, displayed the highest amount of assessor-related variability. This is counter to what one would expect when comparing the results of a presumably standardized assessment such as the DIBELS with those of a teacher observation measure like the COR. These results point to the importance of taking into account the stakes of the assessment when choosing not only the method of assessment, but the type of assessor as well. The results of the current study seriously call into question the use of teacher-assessors for high-stakes assessments, even when using standardized measures. It seems reasonable to suggest that whenever high-stakes decisions (e.g., child placement, school accountability, funded intervention research) are based on assessment outcomes, extramural assessors should be used. Though this may seem costly, the ramifications of making inaccurate decisions are arguably more costly in the long run, in terms of money spent on unneeded services, misguided research, and potentially negative outcomes related to inaccurate placement decisions or school sanctions.

Although the findings were extracted through multilevel modeling, they parallel closely many studies grounded in generalization theory. Hoyt and Kems (1999) synthesized through meta-analysis the results of 79 investigations of rater bias in psychological research. They discovered an average of 37% of score variation was attributable to rater bias; specifically, raters' differential interpretations of the rules for application of the measuring devices and raters' disparate opinions about the same target phenomena being assessed. Because such variance constitutes a source of measurement error, they termed it bias variance. They further distinguished between assessment tasks where raters were required to rate explicit attributes (e.g., counting the exact

Table 3
Percentages of assessor variance estimated for the Preschool Child Observation Record in Head Start.

Performance area	Academic year 2006–2007		
	Fall (n = 1502)	Winter (n = 1477)	Spring (n = 1514)
Language and literacy	25.8	25.3	19.7
Mathematics	34.5	34.3	26.2

Note: All values are statistically significant at $p < 0.05$.

Table 4
Percentages of assessor variance estimated for the Dynamic Indicators of Basic Early Literacy Skills during kindergarten.

Performance area	Academic year 2005–2006 (fall)			Academic year 2006–2007 (fall)		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Initial Sounds Fluency	22.4 (n = 302)	69.4 (n = 309)		25.6 (n = 2545)	47.2 (n = 2610)	
Letter Naming Fluency	17.5 (n = 301)	17.0 (n = 311)	27.0 (n = 311)	8.5 (n = 2540)	17.6 (n = 2620)	21.3 (n = 2581)
Phonemic Segmentation Fluency		32.3 (n = 309)	40.6 (n = 311)		38.6 (n = 2598)	44.9 (n = 2580)
Nonsense Word Fluency		32.3 (n = 306)	36.8 (n = 309)		28.3 (n = 2553)	27.2 (n = 2570)

Note: All values are statistically significant at $p < 0.05$.

frequency of a given behavior) versus rate inferential attributes (e.g., select descriptive terms to represent frequency, estimate severity, etc.). Bias variance increased precipitously for inferential attributes, with rater main effects averaging 14% and interactions between raters and the phenomena they were observing 47%. The explicit versus inferential distinction may extend to many of the tasks that characterize standardized direct assessments (with predetermined correct answers and prompts) versus teacher assessments (allowing interpretation and irregular protocols).

Table 5
Percentages of assessor variance, classroom variance, and conjoint variance estimated for the Learning Express in Head Start.

Performance area	Source of variance		
	Assessor	Classroom	Conjoint
Wave 1, academic year 2006–2007			
Alphabet knowledge (n = 1336)	1.0	4.4*	5.5
Vocabulary (n = 1336)	0.0	4.9*	4.9
Listening comprehension (n = 1336)	3.0*	0.7	3.8
Mathematics (n = 1336)	1.2	4.1*	5.2
Wave 2, academic year 2006–2007			
Alphabet knowledge (n = 1343)	0.9	8.8*	9.7
Vocabulary (n = 1353)	0.0	5.2*	5.2
Listening comprehension (n = 1347)	2.8*	0.8	3.6
Mathematics (n = 1349)	0.6	5.8*	6.4
Wave 3, academic year 2006–2007			
Alphabet knowledge (n = 1336)	1.1	5.4*	6.5
Vocabulary (n = 1343)	1.2	5.5*	6.7
Listening comprehension (n = 1337)	2.4*	1.5	4.0
Mathematics (n = 1339)	2.7*	4.7*	7.4
Wave 4, academic year 2006–2007			
Alphabet knowledge (n = 1275)	0.0	4.9*	4.9
Vocabulary (n = 1280)	1.7	7.1*	8.8
Listening comprehension (n = 1275)	1.4	1.7	3.1
Mathematics (n = 1277)	1.2	4.8*	6.1
Wave 1, academic year 2007–2008			
Alphabet knowledge (n = 1235)	0.1	3.3*	3.4
Vocabulary (n = 1234)	0.7	5.6*	6.3
Listening comprehension (n = 1236)	1.2	2.0	3.1
Mathematics (n = 1235)	1.0	3.6*	4.6
Wave 2, academic year 2007–2008			
Alphabet knowledge (n = 1229)	1.9*	5.2*	7.1
Vocabulary (n = 1229)	0.6	7.3*	7.9
Listening comprehension (n = 1230)	2.4*	3.0*	5.4
Mathematics (n = 1228)	0.0	6.7*	6.7
Wave 3, academic year 2007–2008			
Alphabet knowledge (n = 1204)	0.0	7.1*	7.1
Vocabulary (n = 1211)	0.5	6.4*	6.9
Listening comprehension (n = 1208)	1.9*	2.8*	4.7
Mathematics (n = 1207)	0.5	6.7*	7.2
Wave 4, academic year 2007–2008			
Alphabet knowledge (n = 1129)	0.0	6.7*	6.7
Vocabulary (n = 1127)	0.0	6.6*	6.6
Listening comprehension (n = 1130)	2.1	4.4*	6.5
Mathematics (n = 1129)	1.4	5.8*	7.2

Note: Entries are based on parameter estimates derived from hierarchical linear, cross-classification models for Learning Express data.
* $p < 0.05$.

From still a different perspective, Cote and Buckley (1987) drew upon structural decomposition of measurement variance across 70 construct validation studies. They separated trait, method, and error variance applying multitrait-multimethod designs and determined that trait variance (the type of variance that is the focal interest in most assessments) on average accounted for less than 50% of the variation represented by test scores, the majority of variation conveying irrelevant method differences and measurement unreliability. These investigations demonstrate that much, often most, of the information conveyed by test scores has nothing to do at all with the attributes of those being assessed and the degree of measurement error will vary as a function of the assessment method.

Research also informs that much of the variability in assessments that does factually represent aspects of those being assessed really represents aspects that are the wrong aspects. Teacher assessments via instruments such as COR and DIBELS may provide apt examples because, by their nature, they require or invite many subjective judgments. Within this context they are analogous to typical grading practices. Research on grading shows that teachers commonly make decisions based on factors irrelevant to children's objective academic skill levels when such skill levels are the formally intended targets (Bennett, Gottesman, Rock, & Cerullo, 1993; Brookhart, 1993; McMillan, Myran, & Workman, 2002). Their judgments are influenced by perceived student effort, attitudes, mitigating or aggravating life circumstances, cooperativeness, compliance, behavioral features substantively unrelated to academic skill, teacher beliefs about fairness or social justice or that grades are compensation for attending class, or a desire to promote or avert the social consequences associated with certain positive or negative evaluations. Teacher assessments can also reflect responses to parental pressure and the press of professional competition and accountability (Bishop, 1992; Brookhart, 1993).

The implications of appreciable assessor variance are serious and widespread. Assessor variance is real bias variance (Hoyt & Kems, 1999) or measurement error (Raudenbush, Martinez, Bloom, Zhu, & Lin, 2008). It operates not only to confound assessor qualities with child qualities but to diminish the reliability of assessments, sometimes markedly so. In turn, validity of assessments is compromised because scores no longer represent intended phenomena. The direction of bias is also predictable because the correlational correspondence and strength of relationships between biased assessments and other important constructs will break down, which will be reflected in correlations that are smaller than they would be otherwise. This negative bias, as manifest through attenuated correlations, can profoundly affect school evaluation programs and controlled research enterprises such that real differences will be masked or blunted through reduced statistical power (refer to Raudenbush et al., 2008, and Raudenbush & Sadoff, 2008).

To illustrate the point, one might easily imagine the dilemma surrounding the discovery (as in this study) that, on average, more than one-quarter of the information conveyed by children's assessment scores has nothing to do with children's performance variation, per se. Yet, educators are looking squarely at those scores and assuming that future increments reflect growth and development of children's skills (rather than changes in assessor presentiments or accuracy) and that ostensibly pathognomonic low scores earmark child problems and the need for special services (and not shifts or lapses in assessor focus). Indeed, the marked presence of assessor variance signals the absence of assessment integrity and threatens any intended legitimate purpose.

Implications of the findings of this study relate to the issues of purpose and use of assessments, training, and measurement development. Within this framework, we thought it important to lay out a set of recommendations to help guide child assessments.

3.1. Purpose/use

High-stakes educational decisions should be based on reliable scores derived from direct assessments by extramural assessors. As Bennett et al. (1993) argue, teacher judgments should be supplemented with independent objective evidence. Teachers, in general, and perhaps particularly those working with young children or more vulnerable populations, are naturally conflicted between the role of judge to determine levels of observable competency and the role of advocate to protect and promote (Bishop, 1992; Brookhart, 1993). There are times when teachers cannot simultaneously play both roles and play them competently. Brookhart (1993) has asserted that realistic assessments of academic skills require an absolute standard and for important decisions, Bennett et al. (1993), Bishop (1992), and Brookhart (1993) maintain it is important to yield to external (what we call extramural) assessments.

Whenever multiple children are assessed by the same assessor(s), it should be recognized that some portion of the variability in outcomes is due to the performance of the assessor(s) rather than the children. It is imperative that the contribution of assessor variation to observed outcomes or test scores be minimized and that the contribution of child variation be maximized. Variation must reflect children's individual differences.

3.2. Training

Assessors must be trained and monitored to apply measuring devices using the intended standardized administration protocol and to render consistent and objective judgments. This principle assumes that assessor accuracy will drift and deteriorate in lieu of requisite refresher training and continuous supervision. There is unambiguous evidence that good training reduces measurement bias for assessors (Hoyt & Kems, 1999; Raudenbush et al., 2008). Notwithstanding the explicit protocols for the direct assessments used in this study, extramural assessor training was rigorous and repetitive and oversight continuous.

Our focus groups with Head Start teacher assessors made apparent that assessor training was much less organized and often ended after introductory exposure to the measuring devices. But even good training will not eliminate all assessor variance (Hoyt & Kems, 1999; Raudenbush et al., 2008), especially for devices that depend upon assessor impressions and inferential judgments (Hoyt & Kems, 1999). Some measuring devices may be so dependent upon observer judgments that either no amount of training will suffice or training costs will preclude use of the devices (Hoyt & Kems, 1999; Raudenbush et al., 2008).

3.3. Development

Where measuring devices are intended for applications with multiple children by the same assessor(s), device developers, researchers, and practitioners should demonstrate through multilevel modeling with representative samples that amounts of attendant assessor variance are negligible or relatively small as compared to the amounts of variance associated with children's individual differences. This requires estimation and reporting of the ICC or percentage of assessor-related variance (refer to McDermott et al., 2009, and McDermott et al., 2011, for demonstrations).

In research applications where measuring devices are presented as criterion outcomes and those outcomes are reported for multiple children based on assessments by the same assessor(s), reports should provide not only evidence of simple overall relationships between predictors and criterion measures, but also the amount of estimated assessor-related variance conveyed in such relationships. Thus, reports of predictive efficiency should include reductions in ICCs associated with (or percentages of

variance explained by) any given predictor, having controlled for the assessor-related variance influencing the research. Recent examples are provided by McDermott et al. (2011) where new instruments are given validity support through predictions of future assessments that are nested within teachers. In addition to ordinary Pearson correlations to characterize effectiveness of prediction, the authors recount the percentages of between-children (vs. between-teachers) variance that are available for explanation and the percentages of that between-children variance that are actually accounted for by the new measures under study.

Finally, we have several recommendations specific to developers of early childhood assessments. Developers should make explicit: (1) the intended use of the assessment tool (e.g., research, accountability/program evaluation, formative assessment), (2) the amount of score variation that reflects true child differences as opposed to assessor differences (in addition to traditional measures of reliability and validity evidence), and (3) the requisite training and re-training required to obtain assessor variance levels similar to those reported during standardization. During the current climate of increased accountability for early childhood programs and educators, it is especially important that assessment developers and users follow these and similar guidelines (see AERA/APA/NCME, 1999 and NRC, 2008). Only then can faith be placed in the decisions made about children, teachers, and programs on the basis of assessment outcomes.

References

Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston, MA: Pearson Education Group.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology*, *85*, 347–356.

Bishop, J. H. (1992). Why U.S. students need incentives to learn. *Educational Leadership*, *49*(6), 15–18.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*, 123–142.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales – Listening Comprehension Scale*. Circle Pines, MN: American Guidance Service.

Cote, J. A., & Buckley, R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research*, *24*, 315–318.

Dunn, L. M, & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–Third Edition Form A*. Circle Pines, MN: American Guidance Service.

Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2010). An integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal*, *48*, 763–793.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability–Third Edition Form A*. Austin, TX: PRO-ED.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Head Start Act, 42 U.S.C. §9801 (2007).

High/Scope Educational Research Foundation. (2003). *Preschool child observation record* (2nd ed.). Ypsilanti, MI: High/Scope Press.

Hoyt, W. T., & Kems, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.

Love, J. M. (2006). *Uses of data on child outcomes and program processes in early childhood accountability systems: Assumptions, challenges, and consequences*. Princeton, NJ: Mathematica Policy Research.

McDermott, P. A., Fantuzzo, J. W., Warley, H. P., Waterman, C., Angelo, L. E., Gadsden, V. L., et al. (2011). Multidimensionality of teachers' graded responses for preschoolers' stylistic learning behavior: The Learning-To-Learn Scales. *Educational & Psychological Measurement*, *71*, 148–169.

McDermott, P. A., Fantuzzo, J. F., Waterman, C., Angelo, L. A., Warley, H. W., Gadsden, V. L., et al. (2009). Measuring preschool cognitive growth while it's still happening: The learning express. *Journal of School Psychology*, *47*, 337–366.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*, 203–213.

Meisels, S. J. (1998). *Assessing readiness: How should we define readiness? National Center for Early Development and Learning Spotlight Series* (No. 3). Washington, DC: Office of Educational Research and Improvement.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, *38*, 73–95.

National Association for the Education of Young Children (NAEYC) & National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE). (2003, November). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8* (Position Statement).

National Education Goals Panel. (1998). *Principles and recommendations for early childhood assessments*. Washington, DC: National Education Goals Panel.

National Research Council. (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Dover, NH: The National Center for the Improvement of Educational Assessment.

Raudenbush, Sd. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2008). *An eight-step paradigm for studying the reliability of group-level measures*. Working paper, University of Chicago.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*, 138–154.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability–Third Edition Form A*. Austin, TX: PRO-ED.

Sattler, J. M. (2001). *Assessment of children cognitive applications* (4th ed.). San Diego, CA: Jerome M Sattler Publisher.

Snijders, T., & Baker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools*, *34*, 85–97.

University of Oregon Center on Teaching and Learning. (n.d.). *DIBELS data system*. Retrieved July 12, 2010, from https://dibels.uoregon.edu.

U.S. Department of Health and Human Services. (2003). A vision for quality and accountability for Head Start. *Head Start Leaders Guide to Positive Child Outcomes*. Washington, DC: Administration for Children and Families, Office of Head Start.

U.S. Department of Health Human Services. (2006). *Head Start child outcomes framework*. Washington, DC: Administration for Children and Families, Administration on Children, Youth, and Families, & Head Start Bureau.

U.S. Department of Health and Human Services, Administration for Children and Families. (2010, January). *Head Start impact study, final report*. Washington, DC.

Wang, H., Waterman, C., Perie, M., & Marion, S. (2010, May). Investigating administrator and teacher use of interim assessments. In E.C. Wylie (Chair), *Assessment: Evidence and action*. Symposium conducted at the American Educational Research Association 2010 Annual Meeting, Denver, CO.