ELSEVIER

Journal of School Psychology 47 (2009) 337–366

Journal of School Psychology

Measuring preschool cognitive growth while it's still happening: The Learning Express☆

Paul A. McDermott*, John W. Fantuzzo, Clare Waterman, Lauren E. Angelo, Heather P. Warley, Vivian L. Gadsden, Xiuyuan Zhang

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

Received 22 December 2008; received in revised form 30 April 2009; accepted 8 July 2009

Abstract

Educators need accurate assessments of preschool cognitive growth to guide curriculum design, evaluation, and timely modification of their instructional programs. But available tests do not provide content breadth or growth sensitivity over brief intervals. This article details evidence for a multiform, multiscale test criterion-referenced to national standards for alphabet knowledge, vocabulary, listening comprehension and mathematics, developed in field trials with 3433 3–5$^1/_2$-year-old Head Start children. The test enables repeated assessments (20–30 min per time point) over a school year. Each subscale is calibrated to yield scaled scores based on item response theory and Bayesian estimation of ability. Multilevel modeling shows that nearly all score variation is associated with child performance rather than examiner performance and individual growth-curve modeling demonstrates the high sensitivity of scores to child growth, controlled for age, sex, prior schooling, and language and special needs status.

Keywords: Project Head Start; Preschool education; Item response theory; Cognitive assessment; Growth; Longitudinal studies

☆ This research was supported by the U.S. Department of Health and Human Services's 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, and by the U.S. Department of Education's Office of Special Education and Rehabilitative Services (Grant Nos. P21 HD043758-01 and R01HD46168-01).

* Corresponding author. Tel.: +1 215 822 7906; fax: +1 215 996 1222.
E-mail address: drpaul4@verizon.net (P.A. McDermott).

0022-4405/$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2009.07.002

Introduction

The arrival and discontinuation of Head Start's National Reporting System (NRS; U.S. Department of Health and Human Services [USDHHS], 2003) has rekindled debate on how to best assess the progress of the nation's near million participant children. Should assessment focus mainly on broad accountability and cost benefits? Should it be designed to measure children's performance against uniform standards or is it intended to guide curriculum development and refinement? Or does assessment fulfill its true mission when informing teachers about which children are faring well and, if not, why not? Indeed, there are compelling reasons to expect that assessment must do all of these things for Head Start and do them all well (Gullo, 2005; National Association for the Education of Young Children, & National Association of Early Childhood Specialists in State Departments of Education, 2003; Ziegler & Styfee, 2004). The intensity of the debate motivated a congressional charge to the National Research Council (NRC) to objectively explore the issues and this has culminated in guidelines to help reconcile the disconnects between purposes and practices in early childhood assessment (Snow & Van Hemel, 2008).

Within this context, we report on a federally-funded initiative which, although undertaken primarily to design and field-test new curricula through randomized trials, found it necessary to develop and refine assessment tools that would satisfy many of the varied roles expected by Head Start, while simultaneously addressing a number of the more endemic technical problems faced by Head Start assessment, specifically, and by all preschool assessment, more generally. Given the particular nature of our larger research agenda to develop new curricula in areas of basic literacy, language and mathematics, the assessment innovations reported here are focused exclusively on children's cognitive achievement and do not as yet extend to other important areas. We emphasize that, as applied here, the terms *cognitive achievement* and *cognitive growth* refer to the cognitive constellation of the broader cognitive/physical/social–emotional preschool readiness framework ascribed to by the National Education Goals Panel (1995) and the National Association for the Education of Young Children (Bredekamp & Copple, 1997). The terms encompass early cognition and general knowledge (e.g., in literacy, language, numeracy) and do not embrace the distinction that preschool cognitive learning and development are either exclusively intellective (as in general intelligence) or academic (as in formal academic achievement). Notwithstanding the potential for broader future application, the rapid pace of policy and practice reform argue at this time for a thorough presentation of the new assessment technology and of the evidence thus far supporting it.

Practitioners and policy makers are generally concerned about the progress of children as they move through preschool programs. But assessment of progress cannot be adequately met through static measurements applied with a child once or even twice over the course of a school year. Even the important evidence gathered from true experiments with pre- and posttests has limited value. That is, by the time that posttest assessments are taken, it is too late to do anything about the curriculum or other circumstances that all year have affected the involved children (also see Gullo, 2005, 2006). Good curricula must be more dynamic than typical assessments allow. Progress implies growth and meaningful measurement of growth requires repeated measurements over relatively brief time intervals (e.g., 2–3 months) that would give opportunity to alter the relative course and speed of curricular programs. The demand for more useful assessment is further complicated by the fact that the NRS was, and available commercial tests of preschool cognitive achievement are, at best, intended for the limited before-and-after perspective associated with assessment in fall and spring. This before-and-after perspective presents constraints when working with Head Start children. We have found that for large populations of Head Start children, the highest average increment in correctly-answered literacy and language items over 6–8 months is merely 4 items—a number that would preclude any effective assessment of gains over briefer intervals and that may raise questions about the accuracy of the 6- to 8-month gains (McDermott, Angelo, Waterman, & Gross, 2006). It is unlikely that an average change of 4 items could signal meaningful growth unless the content area represented by the items was extraordinarily narrow and thus not generalizable. And to the extent that such change is typified by the average child, what kind of accuracy could be afforded for all of those evincing less than average performance?

Available commercial tests are constructed as norm-referenced tests (NRTs) based on the performance of large and representative samples of children. NRTs have many valuable applications but as Meisels (2004) has pointed out, they are not intended as measures of broader dynamic processes. This is a critical matter because, beyond a child's first two years of life, there is arguably no period wherein a child's cognitive development is more rapid and expansive than the 3- to 5-year period covered by preschool (Shonkoff & Phillips, 2000). Thus, not only must the points of measurement be closer together but the richness and sampling of content within any given domain also must be thorough enough to adequately represent the domain and to detect meaningful growth. The dilemma is exacerbated further because commercial NRTs necessarily center item content around the 50th percentile of difficulty, whereas it is common to discover that the nation's most disadvantaged preschoolers (as Head Start is commissioned to serve) perform on average well below that range (e.g., 15th–20th percentile; U.S. Department of Education, 2007). This inevitably translates into a markedly small sampling of easier NRT items that will prove relevant to average Head Start populations and fewer, if any, relevant items for Head Start's youngest or most disadvantaged children.

There is reasonable evidence to support the value of teacher-conducted assessments of preschool functioning (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). Such methods allow teachers to acquire a first-hand picture of how each child is performing and, to the extent that teacher-provided assessments systematically are aligned with the intended curriculum, they should also encourage teachers to remain on point in terms of what is relevant and what is not to curricular implementation. Yet for assessment of cognitive growth, there remains no more accurate or objective method than direct assessment of children's skills by persons who are trained technically and who are personally uninvolved in the teaching role (Shepard, Kagan, & Wurtz, 1998, p. 7). If for no other reasons, this is required because high-stakes decisions such as identification of special needs children, reliable measurement of growth, and reprogramming of curricula demand high precision measurement—measurement where variation is almost exclusively driven by child performance rather than by differences in teachers' perceptions or assessment skills.

The necessity for broad rather than narrow representation of cognitive domains and substantial breadth in item complexity from the simplest (to accommodate the less proficient three-year-olds) through the most difficult (to challenge unusually advanced 4- and 5-year-olds), presents an additional burden. A useful mathematics test, for instance, must incorporate many distinct subskills (e.g., serration, counting, cardinality, sorting, and formal operations) and numerous progressively difficult items to detect real growth, while not being so long that it would tax preschoolers or require multiple sessions to finish one testing. And, because test content would require repeated assessment over the year, there would be need for multiple equivalent test forms to offset practice effects. All of this would suggest a careful weighing of the basic demands of responsible assessment and the practical constraints of preschool programs.

The National Institute of Child Health and Human Development in consort with the U.S. Department of Health and Human Services's Administration for Planning and Evaluation and Head Start Bureau and the U.S. Department of Education's Office of Special Education and Rehabilitative Services supported (2002–2008) programs to develop and test curricula that were integrated across multiple domains. One program was located in the nation's fifth largest school system and concentrated on Head Start. This program, known as the Evidence-based Program for the Integration of Curricula (EPIC; Fantuzzo, Gadsden, McDermott, & Culhane, 2003), built curriculum modules in literacy, language, mathematics and learning behaviors; integrated the modules through pilot experiments; and conducted large randomized field trials over multiple years. Initially it had been planned to use commercial NRTs to assess student growth multiple times each year in order to shape and refine the modules. Thus, in academic year 2003–2004 (AY0304) the researchers applied the Test of Early Reading Ability–Third Edition (TERA-3; Reid, Hresko, & Hammill, 2001) to assess alphabet knowledge, the Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997) for vocabulary, Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1995) for listening skills, and Test of Early Mathematics Ability–Third Edition (TEMA-3; Ginsburg, & Baroody, 2003) for mathematics. This effort made clear the inadequacy of the measurements, including narrowness of content sampling within any given domain, paucity of items available for accurate growth measurement and use with younger and more challenged Head Start children, necessity for 1–1$^1/_2$ h testing per child (thus, requisite two sessions at each time point), and only one (listening comprehension) test form for the OWLS. This motivated the development of a test featuring (a) many diverse subskills within each subject domain as aligned primarily with the Head Start national standards, (b) content complexity that would center around Head Start and prekindergarten-aged children and at once enable precision measurement of growth over brief time intervals, (c) provide two equivalent forms for alternate administration, (d) require no more than 20–30 min to assess all content domains, and (e) take full advantage of contemporary item response theory (IRT). The test is called the Learning Express. What follows is the report of its design and validation with large and independent cohorts of Head Start children in a single municipality.

Method

Participant children

Three child cohorts were applied, each resulting from a random selection of classrooms among the 250 Head Start classes operated by the School District of Philadelphia, Pennsylvania. Cohort 1 (AY0405) consisted of 748 children comprising the enrollments of

46 classrooms, there being 48.9% males and 51.1% females ranging in age from 33 to 69 months (*M*=50.5, *SD*=6.8). Approximately 8.2% were regarded as Dual Language Learners (DLL) and 6.3% required attention for special needs. Nearly 71.9% of the children were African American, 17.0% Latino, 10.4% Caucasian, and the remaining having varied other ethnic backgrounds.

Cohorts 2 (AY0607) and 3 (AY0607 and AY0708) comprised the full enrollments of 85 additional classrooms randomly drawn from the Head Start pool (where the pool did not include classes randomly drawn for Cohort 1). Cohort 2 included 1667 35–68-month-olds (*M*=49.8, *SD*=6.8), with 48.8% males, 51.2% females, 10.2% DLLs and 10.4% having special needs. Cohort 3 contained 2685 children across the two years, including 1671 children from Cohort 2 (AY0607) of whom 412 continued in AY0708 and an additional 1014 who were newly enrolled in those classes for AY0708. The total 2685 children ranged in age 35–69 months (*M*=49.2, *SD*=6.8), with 49.7% males, 53.3% females, 11.9% DLLs, and 10.2% special needs. Ethnic composition for these two cohorts was essentially the same, with approximately 69.2% of children being African American, 19.1% Latino, 4.5% Caucasian, and remaining children from varied ethnic groups. Additionally, as pertains to all children enrolled in the city's Head Start program, guidelines mandate that children come from families whose incomes are below the federal poverty level or who are eligible for public assistance (USDHHS, 2008).

Participant assessors

Inasmuch as direct assessment required individually administered testing repeated across the academic year, assessors were recruited at the beginning of each academic year and trained and supervised throughout. There were 20, 45 and 38 assessors working each respective year. The assessors were undergraduate- or graduate-level personnel associated primarily with colleges in the greater Philadelphia region. Ages ranged from 18 to approximately 60 years (median ages in the mid to late 20s), with more than 40% being ethnic minorities (primarily African American) and nearly 20% males.

External criterion measures

Early literacy

Standardized NRTs were used to evaluate the validity of the measures under development as pertains to early literacy, oral language comprehension and numeracy. Literacy was assessed through the Alphabet Knowledge subtest of the TERA-3 (Reid et al., 2001) and Form A of the PPVT-III (Dunn & Dunn, 1997). TERA-3 is appropriate for children between 3 years, 6 months, and 8 years, 6 months; the applied subtest measures basic alphabet skills. Norms are based on a stratified national sample configured to the U.S. Census. Reliability is substantial and validity is supported through correlations with established measures of academic achievement and cognitive ability. The PPVT-III measures receptive vocabulary skills having medium split-half reliability, interrater reliability, and test–retest reliability indices exceeding .90. Measures of validity show a significantly high correlation between the PPVT-III and Wechsler Intelligence Scales for Children–Third Edition (Wechsler, 1991) with *r*=.92, and correlations with the OWLS (Carrow-Woolfolk, 1995) ranging .63–.83.

Oral language

The OWLS (Carrow-Woolfolk, 1995) features a listening comprehension subtest designed to assess children's understanding of spoken language. It is appropriate for those 3 years old through young adulthood. It is standardized on a large nationally-representative sample and supported through appropriate criterion validity studies.

Early numeracy

The TEMA-3 (Ginsburg & Baroody, 2003) served as an external validity measure of mathematics-related skills. It offers assessment of both informal early mathematics (concepts of relative magnitude, counting, calculation with objects present) and formal mathematics (reading and writing numbers, number facts, calculation in symbolic form) for children from 3 years, 0 months, to 8 years, 11 months of age. The nationwide normative sample conforms to the U.S. Census and the test shows high reliability (.94–.95) and evidence for criterion validity (e.g., KeyMath; Connolly, 1998).

Other assessments

The Preschool Child Observation Record (COR; High/Scope, 2003) is a system for evaluating children's individual competencies as based on the classroom teacher's judgment. Evaluations are made several times during the school year. Areas of present interest include Language and Literacy and Mathematics and Science. Internal consistency for the former area ranged .80–.85 and .75–.88 for the latter as reported by High/Scope (2003).

Growth measures

The Learning Express (LE) is a multiple-form criterion-referenced test with very broad content referencing to national and regional Head Start standards and secondary referencing to the nation's leading NRTs. It is constructed for repeated application within a given academic year and for reapplication over consecutive years. Each form contains 195 items of varied formats (multiple choice, oral expressive, manipulation of objects) assessing alphabet knowledge, vocabulary, mathematics and listening comprehension. Item sets were developed de novo (i.e., no items identical to those used in commercial NRTs), modified and calibrated via IRT to yield maximum information and discrimination based on the minimal items actually administered.

Procedure

Development and pilot

The choice of LE content domains was inspired by three factors; theoretical and empirical literature pointing to age-appropriate critical areas for assessment, evidence that such content could be reliably assessed, and the practical constraint of repeated yet relatively brief assessments. Leading research has established that verbal abilities are consistently the best indicators of future reading success (Scarborough, 2001) and, more specifically, (a) that *alphabet knowledge* is a strong predictor of both short- and long-term reading ability (Bond & Dykstra, 1967; Chall, 1967), (b) that preschool *vocabulary* is essential for learning sound distinctions among language parts (Goswami, 2001) and for generation of abstract reasoning (Snow, 1991), and (c) that perhaps the earliest developmentally broad-span manifestation of language ability is children's understanding what is said to them (receptive language or *listening comprehension*; Snow, Burns, & Griffin, 1998). Phonological awareness also is linked to successful reading (National Reading Panel Report, 2000) and early phonemic sensitivity is deemed especially fundamental to quality language development (Snow et al., 1998). However, the practical lower age bounds for sufficient and rich content, the complexities and reliability constraints for the most disadvantaged and youngest preschoolers (e.g., rs < .50; see Burgess & Lonigan, 1998, and Lonigan, Burgess, Anthony, & Theodore, 1998), and the current state of demonstrated test technology did not support the development of phonological awareness measures for this study. Alternatively, *mathematics* development is reliably observed throughout children's first 5 years (Kilpatrick, Swafford, & Findell, 2001) and appears to convey both mediating and causal effects associated with later mastery of cultural symbol systems and general strategic approaches to learning (Miller, 2004). Hence, LE content focused on alphabet knowledge, vocabulary, listening comprehension and mathematics.

During summer 2004, senior research staff at the University of Pennsylvania's Penn CHILD Research Center, in conjunction with early literacy and numeracy experts and Head Start master teachers, built the initial LE item pools. Informed by the extant theoretical literature, the staff created a matrix (as per recommendation of Martineau, Paek, Keene, & Hirsch, 2007) of the national Head Start Indicators (USDHHS, 2006) and corresponding Pennsylvania Early Learning Standards (Pennsylvania Department of Education and Department of Public Welfare, 2005) and further mapped those skills to the item content delivered by the TERA-3 (Reid et al., 2001), PPVT-III (Dunn & Dunn, 1997), TEMA-3 (Ginsburg & Baroody, 2003), OWLS (Carrow-Woolfolk, 1995), Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R; Gardner, 1990), Head Start's fall and spring NRS (USDHHS, 2003), COR (High/Scope Educational Research Foundation, 2003), and Galileo Skills Inventory Version 2 (Assessment Technology, Inc, 2002). The alignments of standards and NRT items were used to guide estimates of relative difficulty. Because the content was criterion-referenced mainly to the national and regional standards, this process produced numerous items covering content more varied than the NRTs. Electronic clip art provided the raw material for initial item stimuli that appear on separate $8^1/_2 \times 11$ in. pages in two large flipbook binders (one for Form A and one for Form B). Most artwork was modified through *Microsoft Office Power Point 2003* (Microsoft Corp., 2003) and *Microsoft Paint* (Microsoft Corp., 2007) to enhance relevant features, eliminate distracting shadows and potentially confusing artistic accents, alter size or perspective as appropriate for preschoolers, and add or remove objects from each item page. Certain mathematics items (pertaining to counting, cardinality, sorting, etc.) entailed use of plastic poker chips and colored and laminated geometric shapes.

The development of equivalent forms was important at the earliest stages of research. This was because the primary role of the LE in the larger study was to assess curricular effectiveness from the outset and across repeated time points each year. One could not forestall a method to minimize practice effects. Moreover, the longitudinal design of assessments would require multiple forms whose equivalence was not compromised at time points more distant from the time point where equivalence was initially established. This compelled early forms equating and special measures to test equivalence as time passed. LE items were constructed in pairs whose members were intended to reflect comparable content and equal difficulty, with one member of a pair assigned to Form A and the other to Form B. For all IRT equating studies, equivalent-groups equating with linking items was applied, where forms were of equal length and multiple-groups calibration was used with children randomly assigned to forms (as per du Toit, 2003, and Zimowski, Muraki, Mislevy, & Bock, 1999). Equating accuracy was tested through comparison of uniformity of all four moments of the distributions for each form via Kolmogorov's *D* (Conover, 1999) at each time period.

Given basic reliability targets for content area and form (viz., *r*s > .90), it was planned to develop at least 40 items for each of two forms in alphabet knowledge, vocabulary and mathematics, and somewhat fewer for listening comprehension (because the items were necessarily more complex and required more testing time). To that end, the first LE edition contained 46 alphabet items per form (including 9 linking items; linking items being those appearing on both forms, thus making possible the IRT scale equating of forms), 61 vocabulary items per form (13 linking items), 48 mathematics items (13 linking), and 34 listening comprehension items (10 linking). Special attention was given to the diversity of content subskills within each subscale and inclusion of more difficult items to detect future growth and avoid ceiling effects. Item content included a wide array of artwork featuring child characters with assorted ethnic characteristics and culturally varied names. All content was reviewed by experts for domain relevance, cultural sensitivity and developmental propriety.

The LE was designed for administration by an adult assessor to an individual child during a single session ordinarily taking 20 min but no more than 30 min. The flipbook binder is placed on a table and oriented toward the child. As each successive item page is exposed to the child, the assessor asks a question (which appears in print on the reverse side of the item page facing the assessor) requiring the child to point to the correct choice, vocally express the answer, or manipulate objects. A standardized prompt is also available for non sequitur child responses or no response. Given the large number of items per subscale, adaptive testing proceeded with the items first ordered in ascending hierarchy of difficulty and the first item administered being one the vast majority of children could answer correctly. A child's basal was established as the highest level of difficulty in the hierarchy at which a certain number of successive items were answered correctly and ceiling the lowest level in the hierarchy at which a certain number of successive items were answered incorrectly. It was assumed that unadministered items below basal would have been answered correctly and that those above ceiling would have been answered incorrectly, thus enabling briefer testing time.

Assessors were recruited through email to psychology and education departments of Philadelphia area universities and through online advertisements. Each assessor was selected through interviews concentrating on formal experience with young children, personal demeanor, apparent maturity, ability to communicate clearly, and ability to commit to extensive training and 3–5 days per week during each wave of testing in the academic year. Approximately one-half of those interviewed were hired (20 hired for AY0405). Thirty-five hours of training were provided in early September, this including basic research methods and psychometrics, early childhood development, working with Head Start children and school personnel, contextual etiquette, teamwork, and 15–20 h practicing the LE, with 10 h supervised practice with Head Start children (not involved in subsequent aspects of the project) in the field. Five assessors were selected as team leaders (a team comprised of 4–5 assessors) based on prior experience teaching or working with young children. Although assignments of classrooms and children were coordinated by central staff, team leaders functioned as liaisons with teachers and administrative authorities, identified semiprivate locations for individual testing, and verified completeness of test protocols while team members were testing in a particular school. For each classroom, children were escorted to testing in the order of the class list, with no more than 5 children removed for testing simultaneously and always with the teacher's knowledge. Standardized questions inquired as to each child's status in terms of special needs, English as primary or secondary language, and health at the time of testing and the teacher's discretion as to whether testing was advisable.

Three testing waves occurred over AY0405, the first centering on October ($n=703$), the second January/February ($n=615$), and the third late April/early May ($n=553$). Whereas the LE items had been tested with Head Start children at the earliest stage of development and again during assessor training, Wave 1 (October) served as the shakedown trial with the full Cohort 1 ($N=748$). The primary goal was to determine the general propriety of the testing protocol across four subscales, to investigate any floor effects, item administration problems, and the accuracy of the assumption that items were ordered in the correct sequence of difficulty and that starting items were suitably easy but not overly so. Initial item ordering was estimated from the ordering of comparable skills (where they existed) on the NRS and NRTs.

Evidence revealed that surmised order of LE item difficulty was not appropriate for the sample and that certain items were problematic in terms of evoking unexpected responses or needing clearer prompts for assessors. It was observed, for example, that certain NRTs had apparently ordered items both according to skill difficulty and according to the convenience of associating items with visual stimuli shared by other items. The latter arrangements (known as testlets) tended to result in situations where items in some testlets were more difficult than those in subsequent testlets. We deliberately avoided use of any types of testlets in the LE in order to avert local dependency and reduction in reliability (refer to Thissen & Wainer, 2001, on problems associated with testlets). Additionally, it must be remembered that the NRTs were not standardized for Head Start populations and so the assumption of synchronized item difficulty with the LE could only have been tenuous.

Examples of unexpected responses included "Sponge-Bob's house" to name a pineapple picture, or the response "Ariel" for a mermaid picture, or a child picking out and naming only part of a picture, or a child merely pointing when a vocal response was required. Given such results, some items were eliminated or revised, prompts clarified, and item order rearranged. Thus, for instance, where proper names were given for vocabulary words, the prompt "Can you tell me another name for this picture?" was added (once only per item); if a child responded to only a piece of a larger picture, the assessor would now circle the entire picture with her/his hand, saying "What is this?" (once per item); or if a child pointed when required to speak, the assessor would prompt, "Can you tell me out loud?" or "Can I hear your big boy/girl voice?" (depending on the item context).

Because Wave 1 item difficulties were not as expected, Wave 1 item statistical properties (other than difficulty estimates to reorder items for Wave 2 application) were not used for decision making. Basal and ceiling stopping rules were suspended for Wave 2 (January/February), thereby extending Wave 2 over two testing sessions. This permitted the application of all items to most children and wrought substantial information on item behavior. Item analyses included examination of biserial correlations, the distributions of locations across ability levels, the magnitude of slope parameters, the joint location and magnitude of information functions, average information functions, item reliability indices (Zimowski et al., 1999), item characteristic curves, item information curves, and χ^2 statistics for item fit under specific logistic models (Bock, 1972). These same types of item-level statistics were used for item analyses at each subsequent calibration phase in test development.

Illustrative decisions pending Wave 2 data included removal of easier items whose information functions were low compared to other easy items (easy items were overabundant and needed to be reduced), items with low biserial *r*s (e.g., .11), items that did not fit the model ($p<.05$), and items with relatively poor average information functions (e.g., .02 in a field where all other items exceeded .20). Special attention was given to retaining items whose maximum information was relatively high, especially in higher ability ranges (in anticipation of skill levels increasing with subsequent waves and to avert any future ceiling effects).

For each subscale, 1- through 3-parameter IRT logistic models (1PL-3PL) were fit and tested via χ^2 deviance tests among -2 log likelihood statistics, model fit statistics (Bock, 1972), average slopes, empirical reliability and maximum information indices. As illustrative of results, the 2PL models were found superior to the 1PL for Alphabet Knowledge (where $\chi^2[93]$ deviance=830.69, $p<.0001$), Vocabulary ($\chi^2[92]=13344.65$, $p<.0001$), Listening Comprehension ($\chi^2[61]=616.69$, $p<.0001$), and Mathematics ($\chi^2[99]=683.20$, $p<.0001$). In contrast to the 3PL model the 2PL emerged as the more suitable because either the more complex model afforded no statistically significant improvement over the more simple model or because convergence was unattainable for the 3PL model.

Differential item functioning (DIF) was assessed through χ^2 tests of the residuals (based on expected comparability of item difficulty parameters) for linking items across forms and comparison of IRT models hypothesizing equality of difficulty parameters across forms versus models hypothesizing different parameters per form. Items displaying statistically significant DIF were removed from a given subscale and the difference between models hypothesizing identical and different difficulty parameters were retested to assure absence of further DIF. Each subscale was calibrated according to the 2PL model as this afforded best fit in all cases and children's scores were calculated through expected a posteriori (EAP) Bayesian estimation (Thissen & Wainer, 2001; Wood, Wilson, Gibbons, Schilling, Muraki, & Bock, 2002) where the scaled score $M=200$ and $SD=50$ at Wave 2. Final item sequencing comported with ascending logit values.

The final numbers of items per form (with number of linking items in parentheses) used for AY0405 Wave 2 and 3 calibration and scoring were Alphabet Knowledge=52 (10), Vocabulary=58 (23), Listening Comprehension=37 (12), and Mathematics=57 (14). With Wave 2 serving primarily to yield detailed item selection data and calibration of each subscale, Wave 3 (May/June) proceeded in the ordinary manner with starting items set at a point where approximately 65% of children passed the first item as per Wave 2 results and with one session per child. Wave 3 scores were estimated through EAP based on Wave 2 parameters, thereby allowing for scores to drift higher than for Wave 2 as a consequence of commensurate growth.

Booster training for assessors occurred prior to Waves 2 and 3 in order to maximize administration fidelity. During each wave, team leaders observed assessors and assessors observed team leaders during test administration and *guided observation sheets* were completed to systematically describe examiner conformance with testing protocol, pacing, prompts, starting points, basal and ceiling rules, and scoring. Fidelity also was checked through periodic accompaniment of senior staff with teams in the field and timely verification of correspondence between recorded responses and assessors' determination of correctness or incorrectness. Protocols that were spoiled by invalid administration (e.g., wrong prompt or inaccurate ceiling) were not processed for psychometric analyses. Toward the middle of each wave, the supervisory staff held a meeting with each individual team leader and assessor to discuss results from observation sheet data and field observations by supervisory personnel. (Also see the later section on fidelity outcomes evidence through examination of sources of score variation.) All of these procedures were replicated throughout subsequent years of the study.

During May, assessors were also trained to administer the alphabet knowledge subtest of the TERA-3, Form A of the PPVT-III, the TEMA-3, and listening comprehension subtest of OWLS. Matrix sampling was used to assign in a quasi-random fashion such that approximately 4 children in any given class received any given test. In toto, 168 children were administered TERA-3, 154 PPVT-III, 171 OWLS, and 157 TEMA-3. These data provided concurrent validity evidence for the LE inasmuch as none of the LE items were items also appearing on the various NRTs.

Curriculum alignment

Since the task for the larger project was curriculum design, the revised LE was administered during 4 waves over AY0506 to the enrollments of 13 classrooms that had been randomly drawn for Cohort 1 and 8 additional classrooms drawn randomly for AY0506. Immediately after each wave, detailed LE performance was reported on a classroom basis. Specifically, although such information was never divulged to classroom teachers, charts were prepared for the curriculum developers that showed the relative level of performance for each item within each subskill. Additionally, charts illustrated the percentage overall passing each item and highlighted the subskills mastered by 40%–60% of children at each successive point in time. This skill progression information was used by curriculum developers to empirically detect the subskills recently mastered by most children (i.e., those correctly performed by 50%–60% of children) and those subskills being newly encountered (40%–50% mastery) by most children. Thus, whereas the LE was primarily aligned to the national Head Start indicators, the curriculum was being aligned to the same standards but at a pace targeted to the maximum number of children.

Final refinement and calibration

In preparation for AY0607's randomized field trials, several LE items that had received poor feedback from assessors were removed, as were 2 linking items deleted from vocabulary because the proportion of linking items was excessive. Moreover, linking items were added to the other 3 subscales such that approximately $^1/_3$ of each subscale's items were linking. This strategy was implemented to ensure that at least $^1/_4$ of each subscale's items were linking after final calibration and DIF analyses where item culling was likely. Robust linking was deemed particularly important to maintaining the equivalence of forms as children progressed through markedly more difficult items. Thus, the numbers of items (including numbers of linking items in parentheses) for Wave 1 AY0607 were Alphabet Knowledge 56 (18), Vocabulary 54 (19), Listening Comprehension 39 (16), and Mathematics 69 (20). Starting items for each subscale were predetermined. Specifically, as estimated from Wave 2 AY0405 results where basal and ceiling rules had been suspended and all items were administered, Waves 1 and 2 AY0607 testing began at the point where approximately 60%–70% of children would pass the first item. Within that range it was additionally assured that (a) the starting item was a receptive and not an expressive Vocabulary item (such that children are eased into the testing situation) and (b) the start point would enable the same numbers of linking items above and below that point on each form. For Waves 3 and 4 AY0607, starting points were adjusted to 60%–70% passing expectancy based on Wave 2 AY0607 performance. All starting levels were intended to minimize testing time while simultaneously reducing stress for younger and less able children and to optimize reliable administration through uniform starting rules at each testing wave.

Basal and ceiling rules for number of consecutive correct and incorrect items were set at 5. The calibration and scoring protocol for AY0607 and AY0708 assumed that, without stopping rules, unadministered items below basal would have been scored as correct and those above ceiling as incorrect. That assumption is justified by the empirical observation of such passing and failing when all test items were administered during Wave 2 of the AY0405 pilot. The protocol also enabled the data essentially required to conduct the many factor-analytic studies over subscales and time as necessary to ensure unidimensionality. Specifically, the alternative missingness associated with frequently unadministered easy and difficult items effectively precluded requisite matrix smoothing of tetrachoric correlation matrices for nonsingularity and positive semidefiniteness. The scoring protocol assured that the item data used for establishing dimensionality were the same calibrated and applied for score estimation.

Forty-five assessors collected data on Cohort 2's 1667 children in 85 random classrooms at each of 4 waves. Participant child samples across the 4 waves were 1336, 1354, 1345 and 1283, respectively. Model comparisons (1PL, 2PL, 3PL), item DIF analyses for linking items, and calibration and scoring transpired as per AY0405, except that both Wave 2 (medial time-point) and Wave 4 (final time-point) were conducted for contrast of model fit, maximum test information, and reliability. Ancillary analyses were undertaken for the subscales as scored using the best model. First, unidimensionality for each subscale at each wave was tested through full-information factor analyses (Bock, Gibbons, & Muraki, 1988; Wood et al., 2002) using smoothed tetrachoric matrices as starting values and maximum-likelihood estimation of slopes and thresholds. Also, to test unidimensionality and local independence, full-information bifactor analyses (Gibbons & Hedeker, 1992; Wood et al., 2002) were applied to the exploratory solutions emergent in full-information factoring. Second, accuracy of the forms equating process was examined through form comparisons at each wave according to all 4 moments of the distributions (*M*s, *SDs*, skewness, kurtosis) for the respective data. Further, the equating solutions based on IRT equivalent-groups equating with linking items were contrasted to solutions based on equipercentile and linear equating. Third, to the extent that it is necessary to assume that the preponderance of LE score variation is driven by actual child performance and not variation in assessor performance, hierarchical linear modeling was used to determine the proportion of score variance attributable to each source for each subscale across the waves.

Evidence for concurrent and predictive validity was collected during spring of AY0607. Teacher-observed child performance via the Language and Literacy scale and Mathematics and Science scale of the COR (High/Scope, 2003) was obtained for 1520 of Cohort 2 children and correlated with concomitant LE scores at Wave 3. Spring COR scores were chosen because they reflected the cumulative record of performance across the year. Additionally, LE's predictive agency was tested by correlating Wave 1 (Fall AY0607) LE scores with those same spring AY0607 COR scores for 1520 children. To provide an empirical contrast for the examination of LE's assessor versus child score variation, multilevel modeling was repeated with the teachers' COR evaluations.

Growth assessment

Detection of change being a central goal, Cohort 3's assessment of 2685 children from the 85 classrooms randomly selected for AY0607 spanned through 4 waves of that year and 4 waves of AY0708, enabling an investigation of growth over 2 years, with intervening summer. Over the 8 waves, participant child sample sizes were 1336, 1354, 1345, 1283, 1236, 1230, 1211 and 1130, respectively. The 8 waves made feasible higher-order, multilevel individual growth-curve analyses, with assessment waves nested within children and children within classrooms. In order to sharpen the focus on cognitive growth related to instruction rather than growth affected by factors external to the theoretical network, growth trajectories were covaried for children's increasing age, sex, English language learner and special needs status, and prior exposure to prekindergarten education.

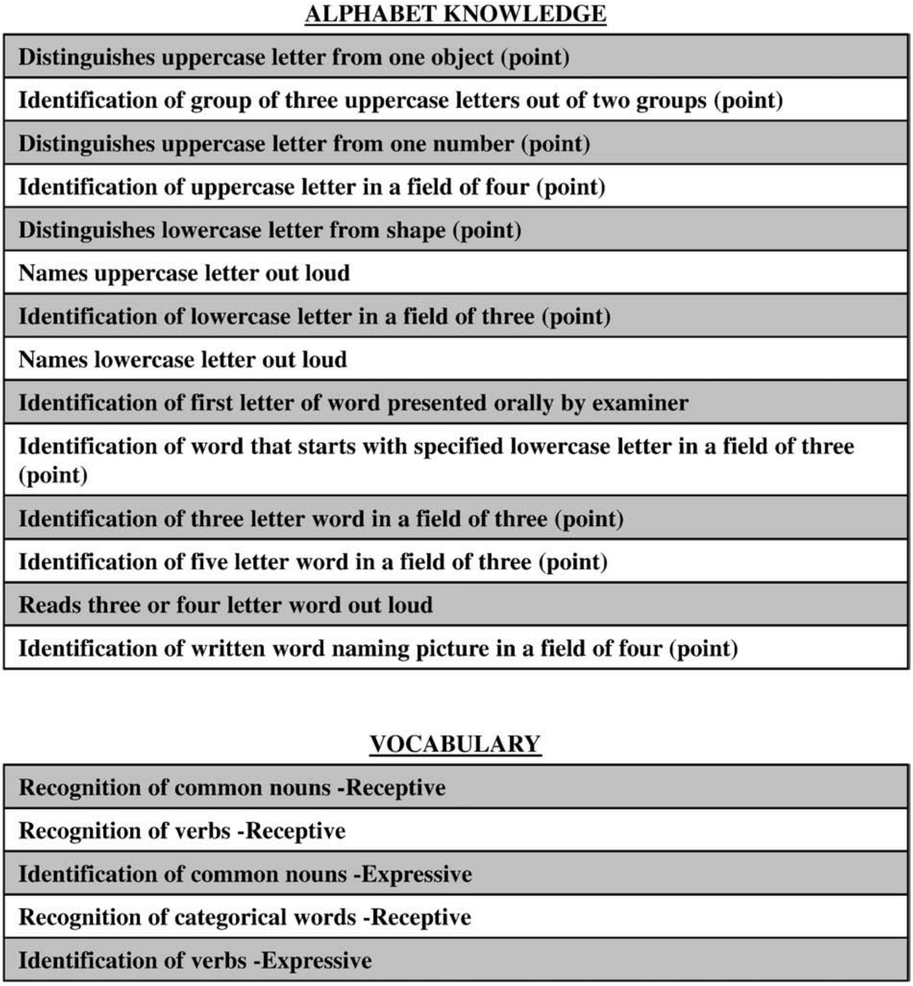
Results

Structure

The final LE edition contains 325 different items distributed over two equivalent forms (A and B) and four subscales (Alphabet Knowledge, Vocabulary, Listening Comprehension, and Mathematics). A total of 56 distinct subskills are featured, with each subscale incorporating multiple subskills (refer to Figs. 1, 2) representing content breadth within domains and fine gradients in difficulty and complexity. Table 1 summarizes the number of subskills and unique items per subscale, as well as the total number of items and linking items per form. Note that since linking items are common across forms, the total number of items per subscale is the sum of unique items and linking items.

In addition to the wide range of subskills, numerous response formats are featured. For Alphabet Knowledge, these include receptive item formats with 2, 3 or 4 response options; expressive formats with letter naming; and expressive formats with word reading. Vocabulary applies receptive formats with 4 response options and expressive formats with picture naming, while Listening Comprehension uses only receptive formats with 4 response options. Mathematics entails physical manipulation of chips or geometric shapes; receptive formats with 2 or 3 response options; and expressive formats requiring counting, numbers naming, or numerical operations.

Whereas the LE was criterion-referenced to national (USDHHS, 2006) and regional standards (PDE & PDPW, 2005), item content was also aligned with the NRS (USDHHS, 2003), TERA-3 (Reid et al., 2001), PPVT-III (Dunn & Dunn, 1997), OWLS (Carrow-



ALPHABET KNOWLEDGE

Distinguishes uppercase letter from one object (point)
Identification of group of three uppercase letters out of two groups (point)
Distinguishes uppercase letter from one number (point)
Identification of uppercase letter in a field of four (point)
Distinguishes lowercase letter from shape (point)
Names uppercase letter out loud
Identification of lowercase letter in a field of three (point)
Names lowercase letter out loud
Identification of first letter of word presented orally by examiner
Identification of word that starts with specified lowercase letter in a field of three (point)
Identification of three letter word in a field of three (point)
Identification of five letter word in a field of three (point)
Reads three or four letter word out loud
Identification of written word naming picture in a field of four (point)

VOCABULARY

Recognition of common nouns -Receptive
Recognition of verbs -Receptive
Identification of common nouns -Expressive
Recognition of categorical words -Receptive
Identification of verbs -Expressive

Fig. 1. Alphabet Knowledge and Vocabulary subskills on the Learning Express.

Woolfolk, 1995), EOWPVT-R (Gardner, 1990), TEMA-3 (Ginsburg & Baroody, 2003), COR (High/Scope, 2003), and Galileo-2 (Assessment Technology, Inc., 2002). Comprehensive mapping charts were constructed for this purpose. As an illustration, Fig. 3 presents a representative section of the charts for Mathematics items appearing on Form A. LE items are listed in order of ascending difficulty (the percentage of children passing an item) with corresponding description and specification of target concept or skill as per the standards. Listed also are NRS and commercial NRT items (in this example, TEMA-3) corresponding to the same standards. As typical across the LE mapping charts, the standards find corresponding LE items to assess the target concept or skill, while the standards often find no representative items on other tests (refer to the shaded areas of

LISTENING COMPREHENSION
Recognition of sequence in a story including concurrent, future, and past events
Recognition simple noun with positional word
Identification of object by usage
Recognition of more complex noun/adjective combinations
Identification of group (i.e. show me the one picture of the animals)
Differentiation of picture utilizing qualifiers including clauses that start with “when” or “instead”
Differentiation of picture utilizing multiple descriptors including singular versus plural subject, gender, and size, color, and number of objects
Recognition of illustration that does not match meaning of statement
Identification of 2 actions in a sequence based on picture
Recognition of picture using quote in a picture/story
Differentiation of compound subject and verb (i.e. dist plural from singular subject and verb)
Differentiation among pictures utilizing what it is not “not, no”
Recognition of more complex positional words (prepositions)
Recognition of more complex noun/verb combinations (2 or more details)
Recognition of noun and comparison of one property

MATHEMATICS
Perception of small numbers (informal)
Seriation
Counting-verbal (informal, one-to-one correspondence)
Shape identification
Counting (informal)
Subtraction with manipulatives
Number identification (formal, reading numerals)
Interrupted sequence
Concretely models addition word problem (informal with pictures)
Verbal counting by ones (informal)
Production of sets
More/less with number identification and number line
Cardinality
Addition with manipulatives
Sorting utilizing one attribute (color)
More/less with number identification without number line
Counting backwards
Concretely models subtraction word problem (informal with pictures)
Mental subtraction (word problems)
Mental addition (word problems)
Sums of small numbers (formal addition)
Subtraction of small numbers (formal subtraction)

Fig. 2. Listening Comprehension and Mathematics subskills on the Learning Express.

Table 1
Content structure of the Learning Express.

Subscale	Number of subskills	Number of items				
		Total	Unique[a]	Linking[b]	Form A	Form B
Alphabet Knowledge	14	94	76	18	56	56
Vocabulary	5	85	66	19	52	52
Listening Comprehension	15	53	38	15	34	34
Mathematics	22	93	80	13	53	53
Total	56	325	260	65	195	195

[a] Items mutually exclusive either to Form A or Form B.

[b] Items common to both forms, as required for equivalent-groups equating of forms. These items evince no differential item functioning (DIF) across forms.

Fig. 3). This reflects the broader content characterizing the LE which was, in turn, instrumental in providing sensitivity to finer gradations of growth.

Model fit, forms equivalence, and reliability

As a measure to minimize practice effects over repeated waves of assessment within a school year, the two forms were applied in a counterbalanced fashion. Children appearing as odd numbers on a class list were administered Form A at Wave 1 whereas those appearing as even numbers were administered Form B. Administration was reversed for each subsequent wave such that, for example, approximately half of the children during AY0607 received form sequence ABAB and half BABA. Each year the order of classrooms to be assessed was random for Wave 1, and from wave to wave there was an effort to maintain the same approximate order for assessing each child (e.g., a given child assessed at the start of Wave 1 was likely to be assessed at the start of other waves). This process served to minimize disparities between children in the time intervals separating their assessments (although time measures were kept to correct for any such disparities in subsequent individual growth modeling). Inasmuch as the two forms were of equal length and the groups of children were essentially equivalent by randomization, the forms were tested for equivalence following final calibration and scoring.

As stated, forms equating and item calibration and scoring were tested at two points in time in order to determine which point yielded better results. Thus, item DIF analyses, equating and calibration were conducted for Wave 2 of AY0607 (which because of its proximity to the middle of the academic year was termed *medial* calibration) and Wave 4 (*final* calibration). Results were equivocal across time points with no particular advantage found for final status calibration in terms of consequent success of equating, total test information and reliability. Thus, the original medial calibration was retained and it is these results that are reported. DIF analyses for linking items on the Alphabet Knowledge and Vocabulary subscales revealed no disparate functioning either through χ^2 fit tests examining individual items or through contrasts of models allowing different parameters for Form A and B items. One Listening Comprehension item and 7 Mathematics items

Item #	Item Difficulty	Item Description	Concept/Skill	PA Early Learning Standards for Pre-Kindergarten 2005	Head Start Child Outcomes Framework Indicators 2006	Revised COR Corresponding Items	Galileo v2 Fall/Spring Corresponding Items	NRS 2003 Category	NRS 2003 Example Corresponding Item	TEMA-3 Skill	TEMA-3 Example Item
14	70.3	Say the number that comes next "3, 4..." (child says "5")	**Interrupted sequence**	No applicable PA Early Learning Standard.	III.A.3	No parallel COR item.	No parallel Galileo item.	No applicable NRS category.	No parallel NRS item.	Number After: 1 to 9 (informal)	A13. What number comes next? (9,5,7)
15	69.2	Point to side with 2 apples (2 vs.4)	**Counting (informal)**	LM 1.5	III.A.6	No parallel COR item.	No parallel Galileo item.	Early Math Skills	E.2. Point to the picture of 3 grapes.	No applicable TEMA skill.	No parallel TEMA item.
16	68.4	Point to side with 5 cups (5 vs. 3)	**Counting (informal)**	LM 1.5	III.A.6	No parallel COR item.	No parallel Galileo item.	Early Math Skills	E.2. Point to the picture of 3 grapes.	No applicable TEMA skill.	No parallel TEMA item.
17	60.9	Count as high as you can (child counts to 10)	**Verbal counting by ones (informal)**	LM 1.1	III.A.3	No parallel COR item.	No parallel Galileo item.	No applicable NRS category.	No parallel NRS item.	Verbal Counting by Ones: 1 to 10 (informal)	A12. 1,2, 3 now count by yourself (child counts from 4 to 10)
18	58.8	What number? (says "2")	**Number identification (formal, reading numerals)**	LM 1.3	III.A.2	No parallel COR item.	No parallel Galileo item.	Early Math Skills	E.5. What is this? (4)	Reading Numerals: Single-Digit Numbers (formal)	A14. What number is this? (2,5,6)
19	58.4	What number? (says "3")	**Number identification (formal, reading numerals)**	LM 1.3	III.A.2	No parallel COR item.	No parallel Galileo item.	Early Math Skills	E.5. What is this? (4)	Reading Numerals: Single-Digit Numbers (formal)	A14. What number is this? (2,5,6)

Fig. 3. Example Learning Express item mapping chart for Mathematics.

displayed statistically significant DIF and were removed. Follow-up comparisons of Form A and B models after removal showed no differences.

Equivalent-groups equating with the remaining linking items was performed for each subscale. Each subscale was submitted for initial calibration under multiple-group (i.e., one mutually exclusive group of children per form) 1PL, 2PL and 3PL models, respectively. It was generally apparent that the 2PL model provided the best fit for every subscale, although 10 items were removed in order to maximize fit. Specifically, two Vocabulary items were removed due to relatively low information indices (1 also yielding a significant χ^2 for poor fit), and two items of comparable difficulty were removed from the opposite form (in order to maintain equality of form length). Six Listening Comprehension items were removed based on low information indices and very low logit values (indicating excessively easy items), and 1 item was removed because of dubious validity, as was a comparably difficult item removed from the opposite form. The item tallies posted in Table 1 reflect scale contents used for final calibration and scoring of all AY0607 and AY0708 data.

The 2PL models were contrasted for fit against the alternative 1PL and 3PL models. The 2PL models were found superior to the 1PL for Alphabet Knowledge (where $\chi^2[94]$ deviance = 1814.7, $p < .0001$), Vocabulary ($\chi^2[85] = 1073.9$, $p < .0001$), Listening Comprehension ($\chi^2[53] = 419.6$, $p < .0001$), and Mathematics ($\chi^2[93] = 933.5$, $p < .0001$). Similarly, in comparison to the 3PL model, the 2PL emerged as the better fit for Alphabet Knowledge ($\chi^2[94] = 786.2$, $p < .0001$), Vocabulary ($\chi^2[85] = 534.5$, $p < .0001$), Listening Comprehension ($\chi^2[53] = 427.9$, $p < .0001$), and Mathematics (where convergence for the 3PL model was unattainable).

Table 2 presents a variety of representative item-level statistics for the LE at medial calibration. Mean slopes ranged from 2.09 for Alphabet Knowledge to 1.09 for Listening Comprehension. Even the least discriminating item (a Vocabulary item with slope = .58) was suitably discriminating with a unidimensional factor loading = .50 (where loading = slope/[1 + slope^2]$^{.05}$). Mean thresholds ranged from .58 for Mathematics to .16 for Vocabulary. The lower range values for information and effectiveness are uniformly associated with very difficult items that are necessary to prevent ceiling effects during later assessment waves, as required for subsequent growth-curve modeling (see Willett, Singer, & Martin, 1998) and curriculum design.

Average total test information for Alphabet Forms A and B were 53.63 and 51.73, respectively, with Form A's approximate maximum information = 128.80 at $\theta = 0.75$ and Form B's = 132.60 at $\theta = 0.63$. Comparable values for Vocabulary (i.e., Form A and B average information and approximate maximum information) were 22.05, 23.71, 26.73 at $\theta = 1.13$, and 37.48 at $\theta = 1.00$. For Listening Comprehension the values were 14.71, 15.41, 26.44 at $\theta = 0.76$, and 26.55 at $\theta = 0.75$, and for Mathematics, 28.72, 28.36, 50.28 at $\theta = 0.75$, and 41.77 at $\theta = 0.88$.

The 2PL parameters derived from medial calibration (Wave 2) were applied to raw scores at every wave and EAP scaled scores were estimated as centered on a Wave 2 $M = 200$ and $SD = 50$. Equating was deemed adequate should the distribution of scaled scores across forms remain identical after equating (the *same distribution property* per Kolen & Brennan, 2004). Thus, Kolmogorov's D estimated the similarity of the two form distributions for each subscale at each wave, and the Kolmogorov–Smirnov goodness-of-fit index tested the probability that D was greater than the observed value under the null

Table 2
Representative item-level statistics for the Learning Express at calibration, January 2007.

Statistic	*M*	*SD*	Range
Alphabet Knowledge (n=1344)			
Threshold	0.32	0.95	−1.73/2.63
Slope	2.09	0.84	0.75/4.18
Maximum information	3.67	2.90	0.40/12.62
Maximum effectiveness [a]	0.70	0.76	0.02/3.18
Average information [b]	0.95	0.55	0.08/2.31
Reliability index [c]	0.45	0.15	0.08/0.70
Vocabulary (n=1354)			
Threshold	0.16	1.20	−1.85/2.49
Slope	1.38	0.56	0.58/2.89
Maximum information	1.60	1.34	0.25/6.01
Maximum effectiveness	0.24	0.26	0.03/1.45
Average information	0.44	0.21	0.10/1.29
Reliability index	0.29	0.09	0.09/0.56
Listening Comprehension (n=1348)			
Threshold	0.23	0.79	−1.56/1.63
Slope	1.09	0.38	0.63/2.03
Maximum information	0.96	0.69	0.28/2.98
Maximum effectiveness	0.24	0.14	0.06/0.65
Average information	0.45	0.20	0.20/0.94
Reliability index	0.30	0.09	0.16/0.48
Mathematics (n=1350)			
Threshold	0.58	1.44	−1.92/3.78
Slope	1.61	0.53	0.81/2.88
Maximum information	2.08	1.41	0.47/6.00
Maximum effectiveness	0.34	0.37	0.01/1.75
Average information	0.51	0.34	0.10/1.45
Reliability index	0.31	0.15	0.01/0.59

[a] Maximum effectiveness is the maximal product of the information function and the corresponding normal density function and it reflects the information conveyed by an item in a population with a normal ability distribution.

[b] Average information is scaled to a 0, 1 normal distribution of ability.

[c] Reliability indexes equal σ^2/σ^2+MSE and are expressions of actual rather than lower bound estimates of reliability.

hypothesis of no difference between forms (Conover, 1999). Additional contrast was provided by re-equating the forms under the equipercentile and linear methods and comparing results with those produced through equivalent-groups equating with linking items. At every wave the IRT equating outperformed the alternative methods, yielding very small *D* values (*M*=.06, *SD*=.02, range=.03–.09).

Composite internal consistency (Embretson & Reise, 2000) at medial calibration was .98 for Alphabet Knowledge, .96 for Vocabulary, .93 Listening Comprehension, and .96 Mathematics. Comparable reliability was found for the subscales across waves and for the separate forms across waves (range=.93–.98). Moreover, composite reliability was calculated at each wave for AY0607 subsamples comprising all children <48 months age, ≥48 months, males, females, those with English as primary language, English language learners, those with special needs, African Americans, Latinos, and Caucasians. No value fell below .90.

Fig. 4 displays the distribution of test information (viz., $1/\mathrm{SE}^2$) and the standard error across ability levels (θ) for each subscale at the time of calibration (Wave 2, AY0607). The useful range of LE scores is indicated by test information plot levels that remain above corresponding plot levels for error. It is evident that the useful range of LE scores is quite broad and that reliable estimates extend far into the higher ability range (the farthest extension was for Mathematics at 3.5 positive *SDs* and the least extension for Listening Comprehension at 2.2 *SDs*). This is consistent with the goal of producing measurements that will remain reliable as performance growth is assessed through subsequent test waves.

Dimensionality

The adapted IRT models assume that each subscale essentially measures one cognitive construct. The dichotomous item responses for each subscale at each wave were submitted to exploratory full-information factor analyses (Wood et al., 2002) extracting the maximum number of factors advised by the χ^2 deviance test (du Toit, 2003). For every resultant model it was found that the proportion of item variance attributable to the first factor exceeded that attributable to subsequent factors by 3- or 4-fold. Additionally, it was unambiguously evident for every model that secondary or tertiary factors were, in fact, difficulty factors (Bernstein & Teng, 1989). That is, whereas every unidimensional model featured salient loadings (viz., those ≥ .40) for every item, multiple-factor models invariably presented factors composed of easy items versus hard items versus moderately difficult items, thus evincing that the multiple factors were reflections of the substantial difficulty variation along the hypothesized construct. Full-information bifactor analysis (Wood et al., 2002) also was conducted for each subscale to test the proposition that, in addition to a superordinate general factor, one or more viable group factors also permeated the data. Hypothesized group factors included any suggested in exploratory analyses plus various item combinations based on identical subskills within subscales. No model produced any statistically significant or interpretable group factors.

External validity evidence

Concurrent validity is established by correlating scores obtained on various NRTs and the LE administered during May/June 2005. Table 3 presents results and shows that each LE subscale is substantially and significantly related to its NRT counterpart. Additionally, the order of relationships indicates that each LE subscale is more highly correlated with its NRT counterpart than with other NRT subscales (e.g., LE Alphabet Knowledge correlated highest with TERA-3's alphabet subscale, LE Vocabulary with PPVT-III vocabulary, etc.). Concurrent validity also is provided by appreciable and statistically significant Spring 2007 correlations between COR's Language and Literacy subscale and LE's Alphabet Knowledge, Vocabulary and Listening Comprehension subscales, and between COR's Mathematics and Science subscale and LE's Mathematics subscale. These values are posted

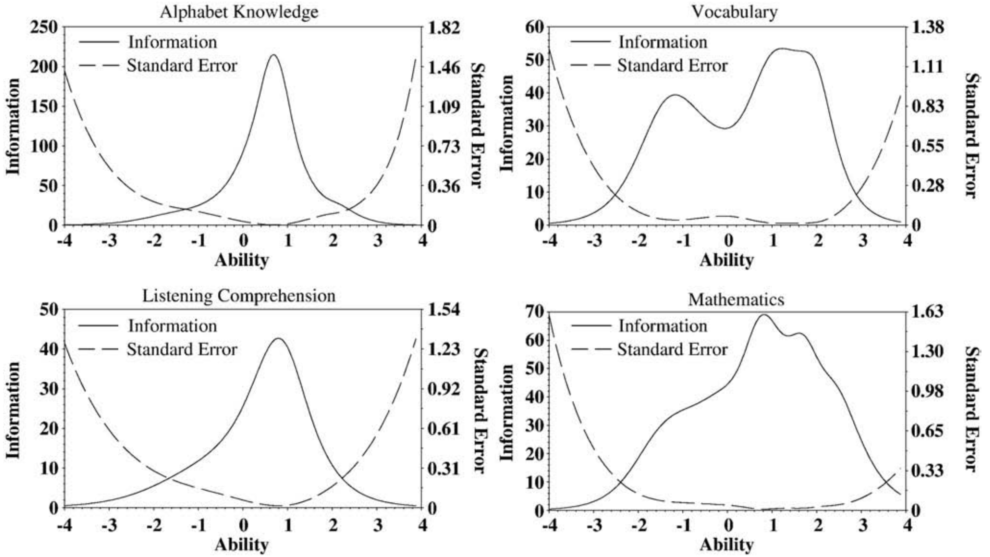


Fig. 4. Distributions of estimated information functions and standard errors for Learning Express subscales at medial calibration, January 2007 (Cohort 2, $N = 1354$).

Table 3
Correlations for concurrent relationships between Learning Express subscales and Norm-Referenced Tests (NRTs).

	Learning Express subscale			
NRT	Alphabet Knowledge	Vocabulary	Listening Comprehension	Mathematics
TERA-3 [a]	.68	.43	.30	.51
PPVT-III [b]	.41	.69	.52	.40
OWLS [c]	.38	.61	.63	.53
TEMA-3 [d]	.52	.42	.42	.59

Note. All values are statistically significant at $p<.0005$. Because of attenuation of OWLS's scores, respective correlations are corrected (Fan, 2002).

[a] TERA-3=Test of Early Reading Ability–Third Edition, $n=150$.

[b] PPVT-III=Peabody Picture Vocabulary Test-III, $n=145$.

[c] OWLS=Oral and Written Language Scales, $n=159$.

[d] TEMA-3=Test of Early Mathematics Ability–Third Edition, $n=144$.

in Table 4 as are values supporting predictive validity through correlations between Fall 2006 LE Wave 1 scores and those same Spring 2007 COR observations. The results in Table 4 show that LE Mathematics correlates higher with COR's Language and Literacy than Mathematics and Science subscale. This disordinality is a reflection of the fact that COR's Mathematics and Science subscale is highly saturated with language and literacy content ($r=.92$ between COR's Mathematics and Science subscale and Language and Literacy subscale), whereas LE's Mathematics subscale is markedly more independent from LE's three literacy- and language-type subscales ($M\ r=.66$ between LE Mathematics and Alphabet Knowledge, Vocabulary and Listening Comprehension; also see Bracken,

Table 4
Correlations for concurrent and predictive relationships between Learning Express subscales and preschool Child Observation Record (COR) subscales.

	Learning Express subscale			
COR subscale	Alphabet Knowledge	Vocabulary	Listening Comprehension	Mathematics
	Concurrent validity [a]			
Language and Literacy	.62 [b]	.56 [c]	.52 [d]	.69 [e]
Mathematics and Science	.54 [b]	.53 [c]	.50 [d]	.63 [e]
	Predictive validity [f]			
Language and Literacy	.57	.56	.52	.65
Mathematics and Science	.52	.53	.50	.58

Note. All values are statistically significant at $p<.0001$.

[a] Entries are correlations between AY0607 Wave 3 Learning Express scores and Spring 2007 COR scores.

[b] $n=1297$.

[c] $n=1303$.

[d] $n=1298$.

[e] $n=1300$.

[f] Entries are correlations between AY0607 Wave 1 Learning Express scores and Spring 2007 COR scores. $n=1215$.

1988, on numerous reasons why valid tests from the same content area often produce fluctuating correlations).

Sources of score variation

There is a fundamental assumption that test scores reflect meaningful variation in examinee performance. The assumption can be mistakenly generalized to situations wherein examinee responses are filtered by test examiners, as when different assessors orally present items to individual children, apply or withhold prompts, and in vivo determine testing floors and stopping points. We believe that the assumption warrants evidence and that its verification provides strong empirical evidence for the fidelity of the assessment process. For every LE subscale at each of the 8 waves throughout AY0607 and AY0708, scores were submitted to hierarchical linear modeling where the percentage of score variance associated with assessors was separated from score variance associated exclusively with children. Assessor-related variance ranged 0.0%–3.1% (M=1.3%, SD=0.9%). These values tended to turn statistically significant ($p<.05$) as they exceeded 1.6%, but are uniformly below the 5.0% criterion reported by Snijders and Baker (1999) as consequential cluster variance in education. However, even in the most extreme case (3.1% for Listening Comprehension during Wave 1 AY0607), a complimentary 96.9% of score variation was attributable to children's performance alone.

In order to provide a contrast for LE's proficiency, we examined through multilevel modeling the sources of score variation for the COR (High/Scope, 2003) as completed by 80 teachers for the Spring AY0607 for 1477 children. Score variability associated exclusively with the teachers was 25.4% for Language and Literacy and 34.1% for Mathematics and Science (both significant at $p<.0001$). These values are markedly disparate from the average 1.3% or upper bound 3.1% found for LE assessor variance and highlight the ability of the LE to focus on relevant phenomena.

Growth sensitivity

Experience with commercial NRTs assessing Head Start children's alphabet, vocabulary and mathematics skills had demonstrated a maximum average gain of 4–5 correctly-answered items over the school year. For the LE over a comparable period, the average gains for the same content areas were 9–15 items. Multilevel individual growth-curve analyses for LE scaled scores over the 8 waves comprising AY0607 and AY0708 further manifested significant growth rates for each subscale. To illustrate, we present in Fig. 5 the growth trajectories for mathematics as determined for children in two distinct Head Start curricula (Cohort 3, N=2685, where classrooms were randomly assigned to curricula in the larger study) and where growth is studied separately according to children's special needs status (special needs vs. nonspecial needs) within those curricular conditions. These trajectories are controlled for children's age, sex, language status (native English vs. English language learners), prior preschool experience, and the differential periods between individual children's assessment dates. Here the LE growth rate is equivalent to 0.18 scaled score points per day or 5.37 points per month. Note that the LE's cubic-curves not only sense within-school-year growth and summer plateaus or losses but also differentiate the two curriculum

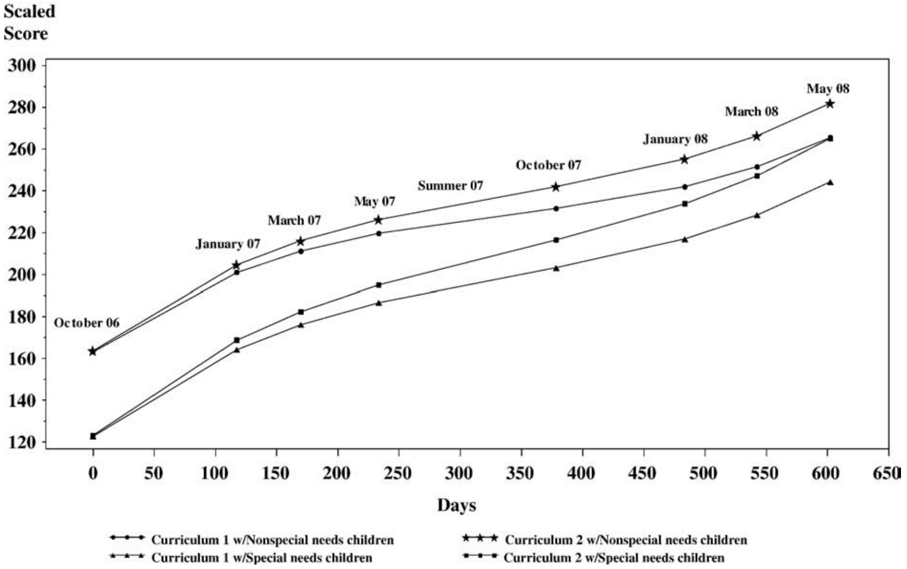


Fig. 5. Estimated average Learning Express growth trajectories for Mathematics by experimental curriculum and special needs status, 2006–2008 (Cohort 3, N=2685).

programs as well as the unique progress of special needs children within those programs. Comparable models for the controlled growth rates with other LE subscales yielded an estimated average growth of 0.14 points per day or 4.27 points per month for Alphabet Knowledge, 0.13 per day and 3.19 per month for Vocabulary, and 0.09 per day and 2.77 per month for Listening Comprehension. Moreover, when the curriculum focused intensely on Listening Comprehension during AY0708, the growth rate for that area increased to 0.15 points per day or 4.56 points per month. All growth rates are significant statistically at $p<.0001$.

Monitoring curricula

It was proposed that a more sensitive and relevant assessment device might also play an important role in curriculum design, monitoring and timely refinement. We noted the general procedure by which LE results were used in AY0506, prior to the randomized field trials of the larger study, to align curricula. It will be recalled that the LE was criterion-referenced to national Head Start standards and calibrated such that items were arranged sequentially according to their progressive difficulty. Thereafter, curriculum contents were sequenced in similar fashion but also such that the main foci of lessons comported to the empirical levels at which most children were functioning. This was accomplished in many ways but here we illustrate the general procedure wherewith curriculum contents were aligned to LE performance.

An example is drawn from LE mathematics performance as assessed in January of a given year. Fig. 6 shows the distribution of children who have successfully passed each

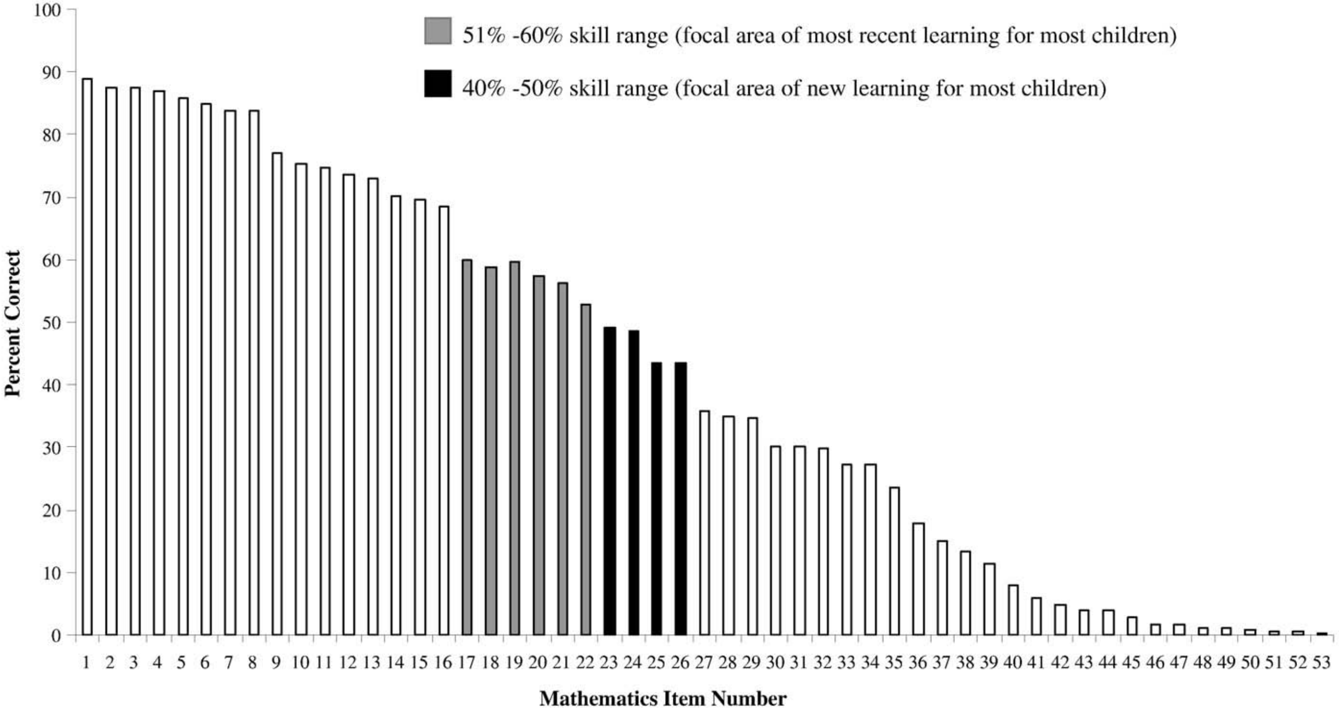


Fig. 6. Learning Express item progression chart for Mathematics performance in January 2007 (Cohort 2, $N = 1667$).

Mathematics item. Note that the items appearing toward the left represent skills mastered by most children and those to the right pertain to skills not yet mastered by most children. Now, whereas teachers will tend to appreciate a breadth of skill levels within their classrooms, it is possible with the LE to highlight more accurately the *typical* skill levels within a program or classroom, or among certain groups of children (e.g., younger vs. older, special needs, language status, etc.) at a given time and at multiple times during the school year. In the example, the shaded areas represent the center of the performance distribution in January. The lightly-shaded items have been answered correctly by 51%–60% of children in the group. So most children have mastered the underlying skills although those skills (given their position in the hierarchy of difficulty) would presumably have been mastered very recently. The premise would follow that those skills would need to be reinforced in upcoming lessons because they are new to most and as yet not learned by many. The dark-shaded area pertains to skills next in difficulty progression but not yet mastered by most children. The pertinent skills are revealed by referencing the subskills measured by those items. Referencing would inform the teacher or curriculum planner that lessons should now concentrate on tasks requiring children to identify the one-to-one correspondence among objects that are scattered about, to correctly retrieve a given number of objects from scattered sets, to identify which number is more or less where the numbers are still presented in a number-line sequence and not scattered about, to count a given number of objects and be able to conclude that the sum is the final count (cardinality). These foci ranges can be narrowed or broadened as necessity demands to maintain the changing central relevance of curricular content. It was this type of evidence-base that guided the sequencing and refinement of the EPIC Head Start curriculum. The utility of the LE curriculum-monitoring process is potentially more universal because it provides real-time assessments of progress (at any level of child grouping) at multiple times during the school year, thus permitting corrections to the direction and pace of the curriculum.

Discussion and conclusion

Development of the LE has embraced Crocker's (2003) view that content validity is the load-bearing factor that will hold or fail when educational tests are put to work. Our approach also has striven to adopt the strategic framework articulated by Lissitz and Samuelson (2007), where investigative focus will produce either *internal* evidence (content, latent process, reliability) or *external* evidence (nomological, utility, impact). From the internal perspective, we were not building a theory de novo about the structure of early childhood cognitive growth; rather, we began with the premise that national standards and multiple sources of expertise had already established the skills to be assessed. The match between the skills measured by the LE and the standards, as well as to skills measured by the popular NRTs, is almost tautologically certain at the preschool level where the standards are explicit and straightforward. Whereas there are no universal, state-level, preschool standards, the Pennsylvania standards are among those most broadly generalizable in the country (Scott-Little, Kagan, & Stebbins Frelow, 2006) because they are tied to research evidence and broader understandings in the areas of cognition and general knowledge. This argues for the useful application of the LE in other locales. Given the expansive focus on both national and widely-relevant regional standards, the LE broadens substantially the coverage of skills proposed by the standards to the point where LE items align with explicit standards not assessed elsewhere (e.g., refer to Fig. 3 and descriptive text). Our intended purposes to measure growth over short intervals and for children whose performance levels do not emulate those of the general population, further demanded finer gradations of item difficulty centered on a relevant population. The theoretical demands of IRT, which itself was adopted to exploit the precision measurements possible, demanded further that we have clear evidence of latent unidimensionality overall and for every point in time. Beyond meeting the targets for uniformly high internal consistency across areas and time, requisite internal validity also had us provide equivalent testing forms that could reduce practice effects and confirm validity of the assumption that the variation in test scores was actually a reflection of children's differences and not test-givers' differences.

The relationships between LE performance and that alternatively evaluated through NRTs and teachers' observations for the same content areas lend further validity support, although it is somewhat unclear whether this type of evidence is properly internal or external. We would suggest that evidence for external utility resides more centrally in the LE's ability to detect growth even after score variation is controlled for confounding phenomena (age, sex, prior schooling, language and special needs status, and individual variation in intervals separating assessments). We also have explained how the LE was applied in curriculum development and how it may be applied to monitor children and programs. The LE's actual impact in those areas will require more convincing evidence, such as data to demonstrate that curricula so designed are somehow superior to those uninformed by LE information.

The LE was designed to assess growth, especially as manifest in more disadvantaged populations. It was not intended to replace sophisticated normative tests. Tests such as TEMA-3 and OWLS are effective at performing their main mission yielding a general nomothesis for the distribution of certain kinds of cognitive achievement in the general population. They can measure a child's performance relative to that of others' performance at a given point in time. The point here is rather that, by fulfilling a mission more concerned with the general population, they are not ideally suited to satisfy optimally the needs of disadvantaged children within that population. Nor is the LE meant to cover every worthy area of cognitive growth that might be measured objectively. Areas like science and social studies may further lend themselves to objective assessment. Preschool growth assessment is much more broad-scoped than even cognitive areas and those additional areas (e.g., behavior, art, and hygiene) deserve close attention as well. As illustration, preschool learning behaviors have been a popular focus for some time, but it is only recently that attention has gone to growth assessment in that area such that instrumentation is able to detect growth over time and across many differentiated skills that would support curriculum design and targeted intervention (McDermott, Warley, Waterman, Angelo, & Sekino, 2009). Moreover, many facets of growth require examination through repeated work sampling (Gullo, 2005; Meisels et al., 2001) that would reveal more subtle patterns idiosyncratic to children or to aspects of the curriculum or classroom workspace or social environs.

The matter of practicality versus objectivity is a pervasive topic in the controversies about preschool assessment (Gullo, 2005; Pretti-Frontczak, Kowalski, & Douglas Brown, 2002; Shepard et al., 1998). It was emblematic of the NRC report (Snow & Van Hemel, 2008) to stress the connection between clarity of goals in preschool assessment and availability (or not) of technology to reach those goals. To be helpful to teaching staff, child assessment must be maximally informative and minimally intrusive of both teachers' and children's time. Program administrators want good information too and must balance the costs against the outcomes, whereas regulating agencies are charged to ensure through assessments that public monies and trust are wisely invested. Assessments of child cognitive growth that are provided by preschool teaching staff can be rather convenient and they may serve to focus teachers more sharply on the intended curriculum. Yet, as we have reported, a quarter to a third of the information conveyed by those assessments is not objective information about children at all; it is information about the teachers who do the assessments. We would contend that high-stakes decision making about preschool children and educational programs must be highly objective, relevant and timely. It is only that sort of information that holds any meaningful promise of practical usefulness to all concerned. It was this imperative that motivated development of the LE.

References

Assessment Technology, Inc. (2002). *Galileo Skills Inventory Version 2. Tucson.* AZ: Author.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105*, 467−477.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29−51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261−280.

Bond, G., & Dykstra, R. (1967). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly*, *2*, 5−142.

Bracken, B. (1988). The psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology*, *26*, 155−166.

Bredekamp, S., & Copple, C. (1997). *Developmentally appropriate practice—Revised.* Washington, DC: National Association for the Education of Young Children.

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology*, *70*, 117−141.

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

Chall, J. (1967). *Learning to read: The great debate.* New York: McGraw-Hill.

Connolly, A. J. (1998). *KeyMath — Revised: A diagnostic inventory of essential mathematics — Normative update.* Circle Pines, MN: American Guidance Service.

Conover, W. J. (1999). *Practical nonparametric statistics*, 3rd ed. New York: Wiley.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, *22*(3), 5−11.

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

du Toit, M. (Ed.). (2003). *IRT from SSI, BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fan, X. (2002, April). *Attenuation of correlation by measurement unreliability and two approaches for correcting the attenuation.* Paper presented at the American Educational Research Association 2002 Annual Meeting. New Orleans, LA.

Fantuzzo, J. W., Gadsden, V., McDermott, P. A., & Culhane, D. (2003). *Evidenced-based Program for the Integration of Curricula (EPIC): A comprehensive initiative for low-income preschool children* (School Readiness Grant No. R01HD46168-01). Washington, DC: National Institute of Child Health and Human Development.

Gardner, M. F. (1990). *Expressive One-Word Picture Vocabulary Test-Revised.* Novato, CA: Academic Therapy Publications.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*, 423−436.

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

Goswami, U. (2001). Early phonological development and the acquisition of literacy. In S. B. Neuman & D. Dickinson (Eds.), *Handbook of early literacy research* (pp. 111−125). New York: Guilford.

Gullo, D. F. (2005). *Understanding assessment and evaluation in early childhood education — Second Edition.* New York: Teachers College Press.

Gullo, D. F. (2006). Alternative means of assessing children's learning in early childhood classrooms. In B. Spodek & O.N. Saracho (Eds.), *Handbook of Research on the Education of Young Children 2nd Edition* (pp. 443−455). Mahwah, NJ: Lawrence Erlbaum Associates.

High/Scope Educational Research Foundation. (2003). *Preschool Child Observation Record 2nd Edition.* Ypsilanti, MI: High/Scope Press.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up.* Washington, DC: National Academy Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practice*, (2nd ed.). New York: Springer-Verlag.

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437−448.

Lonigan, C. J., Burgess, S. R., Anthony, J. L., & Theodore, A. (1998). Development of phonological sensitivity in 2- to 5-year-old children. *Journal of Educational Psychology*, *90*, 294−311.

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practice*, *26*(1), 28−35.

McDermott, P. A., Angelo, L. E., Waterman, C., & Gross, K. S. (2006, April). *Building IRT scales for maximum sensitivity to learning growth patterns over multiple short intervals in Head Start.* Paper presented at the American Educational Research Association 2006 Annual Meeting. San Francisco, CA.

McDermott, P. A., Warley, H. P., Waterman, C., Angelo, L. E., & Sekino, Y. S. (2009, March). *Multidimensionality of teachers' graded responses for preschoolers' stylistic learning behavior.* Paper presented at the American Educational Research Association 2009 Annual Meeting. San Diego, CA.

Meisels, S. L. (2004). Should we test 4-year-olds? *Pediatrics*, *113*, 1401−1402.

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.

Microsoft Corp. (2003). *Microsoft Office Power Point 2003 [Computer software].* Redmond, WA: Author.

Microsoft Corp. (2007). *Microsoft Paint (Version 3.1) [Computer software].* Redmond, WA: Author.

Miller, K. (2004, October). *Developing number names: A cross cultural analysis.* Presentation at the Early Childhood Academy, University of Michigan, Ann Arbor.

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).* Retrieved December 1, 2008, from NAECS/SDE Web site: http://naecs.crc.uiuc.edu/position/pscape.pdf

National Education Goals Panel. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary.* Washington, DC: Author.

National Reading Panel Report (2000). *Teaching children to read.* Washington, DC: National Institute of Child Health and Human Development.

Pennsylvania Department of Education and Department of Public Welfare. (2005). *Prekindergarten Pennsylvania Learning Standards for Early Childhood.* Harrisburg, PA: Author.

Pretti-Frontczak, K., Kowalski, K., & Douglas Brown, R. (2002). Preschool teachers' use of assessments and curricula: A statewide examination. *Exceptional Children*, *69*, 109−123.

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

Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. Dickinson (Eds.), *Handbook or early literacy research* (pp. 97−110). New York: Guilford.

Scott-Little, C., Kagan, S. L., & Stebbins Frelow, V. (2006). Conceptualization of readiness and the content of early learning standards: The interaction of policy and research? *Early Childhood Research Quarterly*, *21*, 153−173.

Shepard, L., Kagan, S. L., & Wurtz, E. (Eds.). (1998). *National Education Goals Panel principles and recommendations for early childhood assessments.* Washington, D.C.: National Education Goals Panel.

Shonkoff, J. P., & Phillips, D. (2000). *From neurons to neighborhoods: The science of early child development.* Washington, DC: National Academic Press.

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

Snow, C. (1991). The theoretical basis for relationships between language and literacy in development. *Journal of Research in Childhood Education*, *6*, 5−10.

Snow, C., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snow, C. E., & Van Hemel, S. B. (2008). *Early childhood assessment: Why, what, and how?* National Research Council of the National Academies Report. Washington, DC: The National Academies Press.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Magwah, NJ: Erlbaum.

U.S. Department of Education (2007). *Reading First and Early Reading First: Student Achievement, Teacher Empowerment, National Success.* Retrieved May 19, 2007, from http://www.ed.gov/nclb/methods/reading

U.S. Department of Health and Human Services. (2003). *Head Start National Reporting System–Direct Child Assessment Fall and Spring.* Washington, DC: Administration for Children and Families, Administration on Children, Youth, and Families, & Head Start Bureau.

U.S. Department of Health and 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. (2008). *Head Start family income guidelines for 2008 (ACF-IM-HS-08-05).* Washington, DC: Administration for Children and Families, Office of Head Start.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children Third Edition.* San Antonio, TX: Psychological Corp.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395−426.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2002). *TESTFACT (Ver. 4.0) [Computer program].* Lincolnway, IL: Scientific Software International.

Ziegler, E., & Styfee, S. (2004). Head Start's National Reporting System: A work in progress. *Pediatrics*, *114*, 858−859.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1999). *BILOG-MG (Ver. 3.0) [Computer program].* Lincolnway, IL: Scientific Software International.