Journal of School Psychology 51 (2013) 97–115

Contents lists available at SciVerse ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Assessing changes in socioemotional adjustment across early school transitions—New national scales for children at risk ☆

Paul A. McDermott [a,*], Marley W. Watkins [b], Michael J. Rovine [c], Samuel H. Rikoon [a]

[a] University of Pennsylvania, USA
[b] Baylor University, USA
[c] Pennsylvania State University, USA

ARTICLE INFO

Article history:
Received 27 February 2012
Received in revised form 9 October 2012
Accepted 15 October 2012

Keywords:
Early childhood
Preschool education
Head Start
Behavioral adjustment
Item response theory
Transition

ABSTRACT

This article reports the development and evidence for validity and application of the Adjustment Scales for Early Transition in Schooling (ASETS). Based on primary analyses of data from the Head Start Impact Study, a nationally representative sample ($N = 3077$) of randomly selected children from low-income households is configured to inform developmental–transitional stability and change in socioemotional adjustment. Longitudinal exploratory and confirmatory factor analysis of the ASETS revealed behavioral dimensions of Aggression, Attention Seeking, Reticence/Withdrawal, Low Energy, and higher-order dimensions of Overactivity and Underactivity. Each dimension was vertically equated through IRT, with Bayesian scoring across 2 years of prekindergarten, kindergarten, and 1st grade. Multilevel modeling provides evidence for concurrent validity, assessment of future risk, and detection of differential growth trajectories across the 4 years of early school transition.

1. Introduction

We are experiencing a period in American history where child mental health has been elevated to its highest priority (President's New Freedom Commission on Mental Health, 2003). It is now estimated that nearly 20% of the nation's children encounter notable behavioral and emotional problems (Egger & Angold, 2006). Many of the problems are manifest during early childhood as detected by parents and child care personnel (Kataoka, Zhang, & Wells, 2002), but fewer than one in five affected children ever receive early intervention services, with such services usually not fully engaged until later in elementary school (Feeney-Kettler, Kratochwill, & Kettler, 2011). Absent and delayed early interventions are known to increase substantially the duration and intensity of subsequent childhood disturbances and the attendant costs to society (Campbell, 2001; Campbell & James, 2007; Rescorla et al., 2011).

Recent research has tended to establish two prominent trends connecting early childhood socioemotional status and subsequent social and educational development: (1) many early childhood problems, especially those evident during preschool, are not transient and will persist through formal schooling and beyond, and (2) many preschool manifestations of socioemotional distress, although transient, portend other sorts of distinctly different problems in later years. Buss (2011) has found that toddlers' failure to develop coping mechanisms for their withdrawal tendencies is a precursor to later withdrawal and associated anxiety, whereas Briggs-Gowan, Carter, Bosson-Heenan, Guyer, and Horowitz (2006), as well as Pihlakoski et al. (2004) have demonstrated the persistence of early undercontrolled versus overcontrolled behavior. Similarly, Petitclere, Boivin, Dionne,

☆ This research was conducted with the cooperation of the U.S. Department of Health and Human Services, Administration for Children and Families, and supported in part by the U.S. Department of Education, Institute of Education Sciences (grant no. R305C050041-05).

* Corresponding author at: Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104-6216, USA.
E-mail address: drpaul4@verizon.net (P.A. McDermott)

ACTION EDITOR: Patricia Manz.

0022-4405/$ – see front matter © 2012 Published by Elsevier Ltd. on behalf of Society for the Study of School Psychology.
http://dx.doi.org/10.1016/j.jsp.2012.10.002

Zoccolillo, and Tremblay (2008) have shown the long-term persistence of early-detected rule-breaking propensities, and Mathiesen, Sanson, Stoolmiller, and Karevold (2009) have discovered that early childhood temperament accounts for both longitudinal stability and change in behavioral growth trajectories. It is also well understood that preschool socioemotional status can serve as a precursor to later success or failure in other, phenomenologically different, areas of functioning, such as cognitive and scholastic proficiency and social resiliency (Dobbs, Doctoroff, Fisher, & Arnold, 2006; Fantuzzo, Bulotsky-Shearer, Fusco, & McWayne, 2005; Fantuzzo et al., 2007; Hair, Halle, Terry-Humen, Lavelle, & Calkins, 2006; McWayne & Chang, 2009; von Suchodoletz, Trommsdorff, Heikamp, Wieber, & Gollwitzer, 2009).

Given the new priority on child mental health, early assessment of preschool socioemotional adjustment has increased precipitously (Campbell & James, 2007; Feeney-Kettler et al., 2011; Rescorla et al., 2011), even culminating in special guidelines for the development, validation, and application of instruments designed for early childhood populations (National Research Council [NRC], 2008; Waterman, McDermott, Fantuzzo, & Gadsden, 2012). Much concern emanates from the fact that many popular assessment devices used with preschoolers are somewhat poorly-conceived step-down versions of tests originally designed for elementary school populations, or that fail to yield the same psychometric integrity found for instruments intended for older children (Campbell & James, 2007; Fantuzzo, Manz, & McDermott, 1998; Fantuzzo, McDermott, Manz, Hampton, & Burdick, 1996; Manz, Fantuzzo, & McDermott, 1999; McDermott et al., 2009; Oades-Sese, Kaliski, & Weiss, 2010). Thus, for example, item content is often insensitive to developmental staging or is so invariant or limited in relevant scope that it impedes accurate assessment or detection of growth.

Concern with instrument quality is especially pertinent in the socioemotional domain where assessments are dependent on adult observations and ratings. In such circumstances, there is a common tendency to base content on features of clinical psychiatric symptomatology (Drotar, Stein, & Perrin, 1995; LeBoeuf, Fantuzzo, & Lopez, 2010), where teacher and parent observers are asked to respond to items that assume some knowledge of children's internal, mediating, psychological processes, such as thinking and feeling (Furr, 2011; McDermott, 1993; Strickland et al., 2011). Moreover, instruments often are composed exclusively of negative or presumably pathognomonic item content with no opportunity for observers to select alternative (and nomothetically more probable) normal or healthy indicators of child behavior. This composition can easily induce response bias because some observers think that they are expected to find something negative or problematic notwithstanding actual observations, while other observers may reactively decline to admit to any child deviations as a means to protect the child's image (McDermott, 1993, 1994).

Nearly all socioemotional instruments compile information on the frequency or intensity of given indicators but fail to contextualize the data by explaining under what circumstances and with whom the problems emerge, thus availing no clues as to motivation for specific problems or their pervasiveness across contexts (McDermott, Steinberg, & Angelo, 2006). Moreover, Waterman et al. (2012) have found that when observation-based ratings are provided by classroom teachers, substantial portions of the variance conveyed by resulting scores have nothing to do with children's actual individual differences (the primary focus of most assessment) but rather reflect idiosyncrasies of the teachers themselves and contextual features of the classroom (also see Hoyt & Kems, 1999, on related measurement bias in psychological assessment).

Perhaps the most compelling aspect of preschool socioemotional development is its transitional nature, particularly as it characterizes the child's journey from the home to prekindergarten or daycare environs to kindergarten and then on to formal schooling (Entwisle & Alexander, 1993; Entwisle, Alexander, & Olson, 2005; Pianta, Cox, & Snow, 2007). In prekindergarten, the child encounters for the first time the challenges of adaptation to organized programs with all sorts of other children and new rules and expectations for engagement and cooperation. The new involvements consume many hours of most days. Meals and naptimes pertain to all, and there can even be new practices around matters of cleanliness and toileting. As the child proceeds into kindergarten and then first grade, the emphasis on discovery learning is somewhat supplanted by more deliberate and structured activities and one moves from activities centered around individual and companion play to larger group-centered common curricula. Self-reliance and competition are encouraged, diffidence is discouraged, desks replace play circles and work tables, vocality becomes imperative, and the written word becomes central to comprehending what is going on in the classroom. One eventually even faces performance grading and the prospect of promotion or retention. Emanating from the Piagetian and Eriksonian tradition, transition theory argues that the child's longstanding socioemotional adjustment framework is heavily influenced by these early interpersonal and environmental changes and by the nature and course of the child's immediate and latent responses to those circumstances (Benner & Crosnoe, 2011; Buss, 2011; Goldsmith & Davidson, 2004; Gurin, Day, Hurtado, & Gurin, 2002; Heckman, 2006; Hemmeter & Ostrosky, 2006; Pianta et al., 2007; von Suchodoletz et al., 2009). These concepts are consistent with the aforementioned research that links preschool socioemotional phenomena to important later outcomes. Indeed, the import of early school transition to our understanding of broader development has essentially risen to a metatheoretical level that regards early transitions in schooling as a major developmental milestone (Eivers, Brendgen, & Borge, 2010).

A rather more troubling picture emerges when considering the preschool socioemotional transitions experienced by children who come from economically disadvantaged families (Benner & Crosnoe, 2011; Hair et al., 2006). Already underresourced, these children typically exhibit profound deficits in academic readiness skills that presage wide achievement gaps persisting through elementary and secondary school (Aud et al., 2011). Compared to peers nationwide, these children tend to function in the 15th–20th percentile for school readiness skills (U.S. Department of Health and Human Services [USDHHS], 2003). Recognizing the gaps, recent federal policy has prioritized methods to improve socioemotional as well as cognitive functioning of these at-risk children and to enhance the transition from preschool to formal schooling (Race to the Top Fund, 2009). Closing the performance gaps and easing the transitions for such at-risk populations have long been goals of Head Start and varied more recent initiatives,

and some of these initiatives do tend to improve the odds of success for many children (Kolker, Osborne, & Schnurer, 2004; Lynch, 2004; Magnuson, Ruhm, & Waldfogel, 2007; Sameroff & Fiese, 2000).

In service of this imperative, the Adjustment Scales for Preschool Intervention (ASPI; Noone-Lutz, Fantuzzo, & McDermott, 2002) was introduced for work with Head Start populations. It was carefully adapted from its predecessor K-12 instrument, the Adjustment Scales for Children and Adolescents (ASCA; McDermott, 1993) by revision of some item content and consultation with preschool teachers in the creation of new items. Like its predecessor, ASPI applied exclusively descriptive behavioral item content rather than items requiring surmise about unobservable psychological processes and incorporated both positive and negative item content to offset teacher response bias. It further embedded each item within a specific classroom situational frame that could inform motivation and context. Because of its unique features and popularity, ASPI content was selected by the federal government for the 2002–2006 longitudinal Head Start Impact Study (HSIS; USDHHS, 2010a), with all 144 items and their special contextual formats preserved intact. Whereas the original ASPI was developed and standardized for Head Start in one major northeastern city in the United States, it was applied in the HSIS for both Head Start and non-Head Start enrollees across a large nationally representative sample, and then readministered annually in a longitudinal, multiple cohort fashion that went beyond 2 years of prekindergarten, through kindergarten, and finally first grade. Because of its broad population use beyond preschool and its new focus on socioemotional stability and change over the critical transition years, the instrument is renamed the Adjustment Scales for Early Transition in Schooling (ASETS).

In this article, we report the evidence for the validity and application of the ASETS. Specifically, the HSIS multiple-cohort sample was reconfigured to inform developmental–transitional stability and change in socioemotional adjustment for the nation's children at greatest risk, whether Head Start prekindergarten enrollees or not. Longitudinally-based exploratory and confirmatory factor analyses were applied, with special provisions for the treatment of discrete binary item data. Dimensional structure was investigated at both first-order surface syndrome and higher-order levels, and resultant dimensions were calibrated via item response theory (IRT) models, with scaling based on vertical equating using linking items across the longitudinal progression from the first year in prekindergarten through first grade. Bayesian scoring was applied with IRT reliability estimates (du Toit, 2003). Validity analyses targeted relations of resulting ASETS scores to concurrent external performance criteria and future outcomes, detection of socioemotional change over 4 years, and observation of differential change trajectories associated with positive and negative outcomes at the close of first grade. All validity studies assessed, using multilevel modeling, the proportions of ASETS score variation that are indicative of children's individual differences rather than assessor or classroom phenomena.

2. Method

2.1. Sample and participants

The original HSIS sampling frame (USDHHS, 2010b) was constructed in the summer and early months of academic year 2002–2003 (AY0203). The sampling frame was designed to be representative of all children eligible for initial entry into Head Start nationwide, excepting only those children who would have enrolled in agencies exclusively serving migrant or farmwork families, Alaskan–Native American tribal populations, and children enrolled in Early Head Start. Based on proportional

Table 1
National longitudinal sample of the Adjustment Scales for Early Transition in Schooling.

Sample characteristic	Developmental level			
	Pre-kindergarten 1	Pre-kindergarten 2	Kindergarten	First Grade
N children	1377	2764	2873	3077
N teachers (classrooms)	867	1815	2280	2576
N preschool centers	540	1032		
% center affiliation				
Nonschool organization	80.94	66.75		
School	13.66	30.90		
Home-based care	5.40	2.35		
% center activity				
Full day	40.63	40.51		
Part day	59.37	59.49		
N schools			1469	1617
% school type				
Public			94.84	95.12
Private			3.77	3.57
Charter			1.24	1.21
Home school			0.15	0.10

Note. Maximum total $N = 3077$.

probability, the sample encompassed all five national regions (Northeast, North Central, South, Plains and West) and 223 grantee agencies. Two cohorts were drawn, one referred to as the 3-year-old cohort (whose constituents were modally 3 years old with a few younger than 3 and a few older) and a 4-year-old cohort (modally 4 years old with a few younger and older). The 3-year-old cohort members were essentially eligible to enter the first year of a two-year Head Start experience, and 4-year-old cohort members were eligible to enter the last year of Head Start albeit their first exposure to Head Start. Overall, 4667 children were selected randomly from the eligible population, 2733 for the 3-year-old cohort and 1884 for the 4-year-old cohort, the imbalance owing to the fact that, whereas the 4-year-old cohort would be tracked through AY0506 (a three-year span), the 3-year-old cohort would also be tracked through AY0506 (a four-year span) thus incurring greater attrition.

The primary purpose of HSIS was to ascertain the relative effectiveness of the Head Start experience as compared to a viable control condition. Therefore, at the opening of AY0203, the 4667 Head Start eligible children were randomly permitted either to enroll in Head Start or (if available) another prekindergarten program (the control condition), and thereafter, the two age cohorts were separately followed through AY0506.

ASETS must be completed by a teacher following at least 1 month of observation of a child in a classroom-type setting. Not all HSIS participants actually enrolled in preschool programs or in programs that involved a teacher and a classroom for the required time period. Thus, the ASETS national sample included a maximum of 3077 children between AY0203 and AY0506. Because the purpose of the present investigation was to explore psychometric properties of the items from a transitional perspective, the HSIS sample was reconfigured to enhance its developmental–longitudinal aspects. Consequently, the two cohorts were pooled to form a 4-year longitudinal sample (Prekindergarten 1 [PreK 1], Prekindergarten 2 [PreK 2], Kindergarten [K], First Grade [1st Grade]). PreK 1 level included all 3-year-old cohort children in their first year of Head Start or another prekindergarten program; PreK 2 level included all 3-year-old cohort children who had moved on to their second and last prekindergarten year and all 4-year-old cohort child entering their first and last year of prekindergarten; K included all who had moved into kindergarten from the preceding level; and 1st Grade included all progressing from the K level. Table 1 presents the structural breakdown of the entire ASETS sample.

Note that the sample increases in size at each level from 1377 in PreK 1 to 3077 in 1st Grade. This pattern reflects the fact that many control children did not necessarily find alternative placements until compulsory schooling was available (1st Grade in most cases) and the fact that not all Head Start or non-Head Start placements necessarily entailed the requisite teacher and classroom setting for the 1-month observation period. It also should be pointed out that neither Table 1 nor subsequent presentations in this article report sample divisions by randomized Head Start versus non-Head Start conditions. This reporting trend emerges from (a) the fact that the current investigation was primarily a psychometric development project and not a randomized control experiment project (by formal agreement between USDHHS's Administration for Children and Families and the ASETS/ASPI publisher) and (b) many of the meaningful distinctions between experimental and control conditions were effectively blurred because, following initial random assignment, numerous children drifted from one experimental condition to another and many non-Head Start programs featured important aspects that were indistinguishable from Head Start (USDHHS, 2010b).

Focusing on the total ASETS sample ($N = 3077$), the mean age at entry to the study was 4.0 years ($SD = 0.5$), with 50.4% of children being boys; 37.8% Hispanic, 29.5% African American, 32.7% White or other race/ethnicity; 25.7% speaking primarily Spanish; 12.8% identified with special needs; and 82.7% residing in urban areas. As per Head Start eligibility criteria, most children came from households at or below the federal poverty level, with up to 35% of children coming from households with incomes up to 1.3 times the poverty level. Additionally, 50.3% of children lived with both biological parents, with 38.9% of mothers never married, 45.3% currently married, and 15.8% currently separated or divorced. Almost 20% of mothers were recent immigrants, and 16.3% were teenagers at the time of their child's birth. Educationally, 38.0% of mothers had not graduated high school, and 71.1% had no schooling beyond high school. By all accounts, the participant sample qualified as at relative risk for serious academic, social, and emotional problems (Federal Interagency Forum on Child and Family Statistics, 2008; Huston & Bentley, 2010).

2.2. Instrumentation: external criterion measures

ASETS results were validated against various types of teacher ratings, parent ratings, and direct assessments. Such measures are reported here for two developmental levels (PreK 2 and 1st Grade), but they were available for all levels. Because the volume of analyses would permit only a representative presentation of results, it was decided to use results from the culminating point of the PreK levels (PreK 2) and post-PreK levels (1st Grade). This decision allowed the maximum amount of participant data to be presented where psychometric properties for the external measures were acceptable. A number of measures administered for HSIS were eliminated from the current study because, at a given developmental level, they failed to yield sufficient data to ensure reasonable statistical power. Measures were also eliminated because the original instruments were altered for HSIS without report of requisite psychometric support (per Smith, McCarthy, & Anderson, 2000), or because they failed to produce minimally acceptable reliability (viz., $\geq .70$ as recommended by Fabrigar, Wegener, MacCallum, & Strahan, 1999; Guilford, 1956; and Nunnally, 1978) for the at-risk population assessed by HSIS (USDHHS, 2010b, pp. 3.32–3.43).

2.2.1. Teacher ratings

The Student–Teacher Relationships Scale (STRS; Pianta, 1996) features 15 items, such as, "This child easily becomes angry at me," rated on a 5-point scale ranging from 1 = *Definitely does not apply* to 5 = *Definitely applies*. Three scales are available:

Closeness (7 items), Conflict (8 items), and Total Positive Relationship (15 items). Ample concurrent and predictive validity evidence is provided (Pianta, 2001; Pianta & Stuhlman, 2004), and internal consistency reliability coefficients for the relevant HSIS developmental levels range from .73 to .82 for Closeness, .76 to .89 for Conflict, and .88 to .89 for Total Positive Relationship (USDHHS, 2010b).

Teacher report of Academic Ability is rated at the end of 1st Grade for Language and Literacy, Mathematics, and Social Science, as based on observed attainment of multiple skills compared to the attainment of peers (USDHHS, 2010b). Initially rated on a 5-point scale from 1 = *Far Below Average* to 5 = *Proficient*, HSIS staff subsequently reduced teacher response data to a simple binary scale (1–2 versus 3–5) to enhance parsimony and reliability. Since each measure is essentially a single index, internal consistency estimates are infeasible. Thus, the appropriate standard error of the mean (*SEM*) is reported here, where (assuming binary scores = 0 versus 1) *SEM* for Language and Literacy = .008, Mathematics = .008, and Social Science = .007. Moreover, given the discrete scaling, subsequent statistical analyses apply logit link functions and the Bernoulli response distribution.

2.2.2. Parent ratings

Parents were asked to rate children's aggressive or defiant, hyperactive, and withdrawn or depressed behavior using the Total Behavior Problems scale. The scale contained 14 dichotomous items, such as "Is disobedient at home," "Can't concentrate, can't pay attention for long," and "Feels worthless or inferior." This scale was originally developed for the Head Start Family and Child Experiences Survey national study building upon prior work by Rutter and Achenbach (USDHHS, 2001). Development and validity evidence for this measure are provided for the FACES national study (USDHHS, 2001, p. 2.27) and for HSIS in USDHHS (2010b). Additional validity evidence has been reported by other researchers (e.g., Vaden-Kiernan et al., 2010; Ziv, Alva, & Zill, 2010). For the 1st Grade sample as reported for HSIS, internal consistency reliability coefficients ranged .78 to .79. For the Head Start and kindergarten samples as reported for the FACES national study, internal consistency ranged from .76 to .80.

2.2.3. Direct assessments

The Peabody Picture Vocabulary Test, Third Edition (PPVT-III; Dunn, Dunn, & Dunn, 1997) assessed listening comprehension for the spoken word (a.k.a., receptive vocabulary). It was adapted for HSIS use by shortening and equating to the full-length version and applying three-parameter IRT calibration and Bayesian scoring (USDHHS, 2010b). Criterion-related validity evidence is abundant (e.g., Dumont & Willis, 2006; Salvia, Ysseldyke, & Bolt, 2007), and internal consistency reliability coefficients ranged .70 to .78 for the HSIS population. Also, various tests of the Woodcock–Johnson III Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2002) were administered, having been abbreviated through IRT calibration and revised stopping rules (USDHHS, 2010b). The current study included scores from the following WJ III tests: Letter–Word Identification (measuring reading identification skills; HSIS reliability coefficients ranging .90–.91), Applied Problems (measuring practical math problem solving by recognizing process and counting or calculating; reliability coefficients ranging .89–.90), Spelling (measuring letter and word writing skills; reliability coefficients ranging .78–.81), Word Attack (measuring phonic and structural analysis skills; reliability coefficients ranging .93–.94), and Quantitative Concepts (measuring number concepts and recognizing patterns and missing aspects; reliability coefficients ranging .86–.87). Some of these tests contributed to composites that were included: the Pre-Academic Skills cluster (composed of scores from Letter–Word Identification, Applied Problems, and Spelling; reliability coefficients ranging .76–.78), the Basic Reading Skills cluster (composed of scores from Letter–Word Identification and Word Attack; reliability coefficients ranging .90–.91), and the Mathematics Reasoning cluster (composed of scores from Applied Problems and Quantitative Concepts; reliability coefficients ranging .71–.78). Considerable validity evidence has been reported for these tests and composite scores (e.g., Dumont & Willis, 2006; Salvia et al., 2007).

2.3. Instrumentation: socioemotional adjustment

ASETS contains 144 items embedded in 24 situational contexts. Each item may be endorsed by the responding teacher to describe the child's behavior over the past month. The 24 contexts encompass relationships with the teacher and with other children, coping with classroom expectations, and demeanor during games and play. Thus, a typical context would inquire, "How is the child at free play (individual choice)?" Within that context the teacher may describe child behavior by endorsing one of more of the following items: "Engages in appropriate activities," "Rather loud but not disruptive," "Is too timid to join in," "Disturbs others' fun," "Wants to dominate and have his/her own way," "Starts fights and rough play," "Needs teacher assistance to get involved," and "Usually plays by him/herself." Note that all of the items are behavioral and require no speculation about unobservable thoughts or feelings, clinical language is avoided, and most contexts provide one or two item choices that are normal or healthy behavior variants. Furthermore, each of the relatively negative items is theoretically reflective of a potential surface syndrome or phenotype, such as aggression, withdrawal, and dependence. Unlike other published instruments that attempt to detect psychopathology by asking respondents to estimate the frequency or intensity of each symptom devoid of context, ASETS subscribes to the theoretical premise that pathology is not determined by the intensity of a symptom in a given context because such a symptom can easily be isolated to the single context and is thus, by implication, reactive and perhaps adaptive rather than pathognomonic. Alternatively, ASETS determines pathology as based on its pervasiveness across many contexts (consistent with the view of Horn, Wagner, & Ialongo, 1989).

McDermott (1993) recounted development of the ASCA items and Noone-Lutz et al. (2002) recounted their revision for the ASPI (herein renamed ASETS because of its transitional nature and extension beyond preschool). Based on a large Head Start sample from a northeastern city, Noone-Lutz et al. (2002) showed evidence for seven phenotype scales: Aggressive, Inattentive–

Hyperactive, Oppositional (and their higher-order composite, Overactive), Withdrawn–Low Energy, and Socially Reticent (and their composite, Underactive). These scales have been validated for convergent and divergent validity (Bulotsky-Shearer & Fantuzzo, 2004; Fantuzzo, Bulotsky, McDermott, Mosca, & Noone-Lutz, 2003; Fantuzzo et al., 2007; Noone-Lutz et al., 2002). As previously described, the instrument was applied in its entirety for HSIS but extended beyond Head Start-type PreK into K and 1st Grade.

2.4. Procedure

2.4.1. Data preparation

Teachers responded to ASETS items during the spring semesters of AY0203–AY0506. The various external criterion measures were administered during the same semesters, with PPVT-III and WJ III scales administered by trained technicians and teacher reports on child academic ability provided at the close of the AY0506 semester (USDHHS, 2010a, 2010b). The average numbers of children assessed per classroom were as follows: AY0203 = 1.59, AY0304 = 1.52, AY0405 = 1.26, and AY0506 = 1.08. A relatively minor portion of ASETS item-level data was missing, where the mean proportion of missing data was .016 (SD = .007) and maximum for any item = .110 (SD = .069). Multiple imputation was performed in SAS 9.3 using Markov-chain Monte Carlo estimation for nonmonotone data and regression imputation for monotone missingness (SAS, 2011). Imputation was assisted with information from demographic variables (child age at entry, sex, ethnicity, primary language, special needs, urban residence), none of which experienced missingness. The process was successful, with mean efficiency = .997 (SD = .0001) and minimum efficiency for any item = .970 (SD = .013).

2.4.2. Subsampling

The subsequent strategy was designed to identify several representative subsamples from the larger sample reported in Table 1. Exploratory and confirmatory factor analyses and vertical scale equating each required a mutually-exclusive cross-sectional sampling across PreK 1, PreK 2, K, and 1st Grade. To this end, a computer routine was written to consecutively draw at random one child from each developmental level (with no child being drawn twice) until a sample of 1600 children was constructed, with 400 different children representing each level. The total size of this sample was dictated by the necessity to (a) generate for statistical power purposes the largest sample possible without redundant membership and (b) that contained equal representation at each developmental level. This sample was termed the *calibration sample* because it was later used for vertical equating and derivation of scoring parameters. Thereafter, the calibration sample of 1600 was randomly split to form an exploratory factor analysis (EFA) sample of 800 and confirmatory factor analysis (CFA) sample of 800.

As described, ASETS contains 144 items. Of these, 22 were deliberately written as alternative positive behaviors to reduce response bias. None of these items was conceived as a potential member of a problem surface syndrome. These items were excluded from subsequent structural analyses and scaling for several important reasons: (a) to avoid the likelihood of consequent difficulty factors (Bernstein & Teng, 1989) associated with binary item factoring, where items would form factors not on the basis of underlying phenotype dimensions but on the basis of items with comparable prevalence (relatively common or positive versus relatively rare or negative), (b) to avert the possibility of valence factors or bipolar factors that were likely to emerge in the presence of oppositely valenced items, and (c) to preserve the intended content focus of the instrument as a measure of variation in problem behavior. The remaining 122 items were used for scale development.

2.4.3. Exploratory analyses

Many researchers have cautioned that unstable or spurious factors may result when dichotomous item data are treated as continuous (Bernstein & Teng, 1989; McDonald & Ahlawat, 1974; Mislevy, 1986; Mooijaart, 1983; Muthén, 1987; Waller, 2001). Alternatively, Waller (2001) suggested iterative factoring of a smoothed tetrachoric matrix. Specifically, we used Waller's MicroFACT software to apply two-stage maximum-likelihood estimation (Olsson, 1979), and the matrix of 122 problem behavior items was smoothed for positive semidefiniteness through least-squares approximation of the original matrix (as per Knol & Berger, 1991). Given the large number of items, it was not possible to produce a nonsingular matrix. The smoothed matrix was submitted for minimum average partialling (MAP; Velicer, 1976) to suggest the maximum number of components for retention and thereafter submitted to iterative principal components analyses (consistent with the recommendation by Snook & Gorsuch, 1989 for sets or 50 or more items) with varimax, equamax, and promax rotations. The ideal structure was that which satisfied multiple criteria; namely, the solution must (a) approximate simple structure as reflected in maximum hyperplane count (Yates, 1987) and item coverage, (b) have at least 4 salient items per factor where loadings $\geq$.40 indicate salience, (c) produce internally consistent factors (i.e., $r \geq .70$), and (d) make theoretical sense in terms of parsimonious coverage of the data and compatibility with leading research in the area (Fabrigar et al., 1999).

2.4.4. Confirmatory analyses

It was anticipated that CFA involving a very large number of items or a very large number of salient items per factor would make difficult if not infeasible the successful application of structural equations modeling (SEM) with item-level data. In the present case, parameters would have to be estimated simultaneously for a large number of highly-skewed binary items, representing a higher-order structure at multiple developmental time points. These requirements would exceed the capacity of available SEM estimation procedures, resulting in nonconvergence. Numerous researchers have pointed to the inherent difficulty that SEM incurs under such circumstances when attempting to minimize the discrepancies between the observed and predicted

covariance matrix, the low reliability of correlations, correlated errors, and the added complication when item data are binary, or worse yet, binary and markedly abnormally distributed (Hall, Snell, & Singer Foust, 1999; Hau & Marsh, 2004; Kishton & Widaman, 1994; Nasser & Wisenbaker, 2003). Consequently, factor analysts have advanced the creation of item parcels (Bandalos, 2002; Hall et al., 1999; Sass & Smith, 2006; Thompson & Melancon, 1996; Wilkinson, 2007), where several items are assigned to each parcel and the lesser number of parcels is submitted for CFA. Several parceling methods have been explored, including those intended to render distributional balance within and over parcels (Bandalos, 2002; Hau & Marsh, 2004; Nasser & Wisenbaker, 2003; Thompson & Melancon, 1996), those realizing representative domain balance (Bagozzi & Edwards, 1998; Hall et al., 1999; Kishton & Widaman, 1994; Little, Cunningham, Shahar, & Widaman, 2002), and those constructed randomly (Hall et al., 1999; Kishton & Widaman, 1994; Wilkinson, 2007). Item parcels can be misused to confound sources of error variance and lead to misspecification of CFA models (Bandalos, 2002), whereas item parcels, if properly applied, can enhance psychometric qualities, including likelihood of normal distributions and reductions of sampling error (Wilkinson, 2007), and enable CFA where it might not otherwise be possible. When item data are inherently positively skewed and leptokurtic (as nearly all problem behavior data tend to be), the method of distributional balance provides best advantage because resulting parcels are more normally distributed. This goal is accomplished by creating, as in the present instance, triplet parcels where each parcel contains two items of maximum counter skew or prevalence and a third item with moderate prevalence. Parcels based on the salient markers from the EFA solution were analyzed via maximum-likelihood estimation under the Satorra–Bentler scaled difference chi-square for non-normal data (Satorra & Bentler, 2001), seeking acceptable fit where the Root Mean Squared Error of Approximation (RMSEA) $\leq .08$ and Comparative Fit Index (CFI) $\geq .90$ (Marsh, Liem, Martin, Morin, & Nagengast, 2011).

2.4.5. Scaling

Having derived ASETS dimensions, each dimension was scaled vertically, connecting PreK 1 to PreK 2, PreK 2 to K, and K to 1st Grade, using linking items. For each dimension, linking items were identified through multiple-group IRT analysis (Zimowski, Muraki, Mislevy, & Bock, 1999) of Differential Item Functioning (DIF). DIF was assessed through χ^2 tests of the residuals (based on expected comparability of item difficulty parameters) for linking items across adjacent developmental levels (e.g., Prek 1 versus PreK 2) and comparison of IRT models hypothesizing equality of difficulty parameters across levels versus models hypothesizing different parameters per level. Items displaying statistically significant DIF were removed from candidacy as linking items. One-third of the number of items composing a dimension for a given level were selected as linking items, being chosen so as to best distribute linking items across logit values covering the dimension's distribution. Vertical equating was accomplished with the longitudinal calibration sample via multiple-group IRT (Zimowski et al., 1999) testing both the one-parameter logistic and two-parameter logistic models. Thereafter, resultant item parameters were applied for the entire ASETS sample, with scores calculated through expected a posteriori (EAP) Bayesian estimation (Thissen & Wainer, 2001; Wood et al., 2002) where the scaled score (*SS*) mean = 50 and standard deviation = 10 at PreK 1, the reference level.

2.4.6. External validity

Product–moment correlations were computed to show the direction and strength of relations between scores on each ASETS dimension and scores on each external criterion variable. Given the mass of such data across developmental levels, it was determined to use the most representative levels; these levels were PreK 2 (culminating the PreK period) and 1st Grade (culminating the post-PreK period). Because the data were nested within teachers and classrooms, and consistent with the recommendations by Waterman et al. (2012), relations also were assessed using hierarchical linear modeling (HLM), where each ASETS dimension served as the group-mean centered predictor in a two-level conditional HLM model, revealing the percentage of between-children within-teacher/classroom variance accounted for by ASETS variance.

Predictive assessment was examined for the relative risk of end-of-1st-Grade academic nonproficiency for teacher-reported Language and Literacy, Mathematics, and Social Science. Because of the binary nature of those reports and the nesting within teachers, two-level, generalized multilevel logistic models were constructed with teachers as the cluster variable, ASETS dimensions as explanatory variables, and teacher-reported binary outcomes as the response variables. The generalized multilevel linear model with the logit link function was estimated for each outcome. Starting values were extracted from preceding models using pseudo-likelihood estimates for the multilevel generalized linear model. This two-step process eschews the negative bias for parameter estimates extracted through pseudo-likelihood estimation (Bauer & Curran, 2006). The final models were used to derive odds ratios expressing the increased risk for subsequent nonproficiency afforded by each increment in ASETS *SS*s.

2.4.7. Change detection

A prime focus in any transition study is the sensitivity of the measurement to real change over time. Each ASETS dimension was entered into a three-level, unconditional growth model, where the first level modeled temporal variability within children over the four developmental levels, level 2 modeled variability between children, and teacher/classroom variability was modeled at the third level. Models included random coefficients for intercepts, linear slopes, and higher-order slopes, as well as linear, quadratic, and cubic trends for change.

Table 2
Dimensional structure and properties of the Adjustment Scales for Early Transition in Schooling.

Item description[a]	Pattern loadings[b]				Communality	Item-scale r[c]	% prevalence[d]
	1	2	3	4			
Scale 1: Aggression (internal consistency = .96[e])							
Physically aggressive in peer conflicts	**.77**	.00	−.04	−.01	.58	.62	11.0
Starts fights and rough play during play	**.76**	.06	.03	−.04	.59	.56	4.6
Overly rough with other children in games	**.75**	.06	−.05	.02	.60	.57	6.3
Unkind to smaller or weaker children	**.70**	.04	.11	−.01	.51	.39	2.6
Attacks other children if provoked	**.69**	−.02	−.05	.07	.50	.53	9.6
Quarrels, provokes agemates	**.67**	.21	−.09	.10	.62	.62	8.9
Is often the cause of trouble in lines	**.66**	.22	−.12	.06	.61	.60	10.8
Makes unprovoked attacks on other children	**.64**	.06	−.01	.05	.45	.43	5.9
Disturbs others' fun during free play	**.64**	.09	−.10	.25	.60	.64	6.5
Disrupts games by fooling around	**.63**	.16	−.20	.26	.67	.61	10.3
Refuses to take turns	**.62**	.28	.30	−.08	.58	.48	4.1
Snatches objects away from other children	**.61**	.15	−.09	.09	.49	.53	10.9
Answers with threats when corrected	**.57**	.37	.00	.01	.57	.54	5.4
Tries to push in front of others in lines	**.57**	.38	−.14	.10	.66	.58	22.4
Sometimes unfriendly with teacher	**.57**	.27	.22	−.01	.50	.45	9.8
Bothers others during teacher directed time	**.56**	.36	−.09	.19	.67	.64	14.5
Tries to dominate agemates	**.54**	.27	−.03	−.10	.42	.47	9.5
Doesn't hesitate to lie	**.54**	.22	.04	.21	.52	.46	4.3
Deliberately destroys others' belongings	**.53**	.13	−.08	.21	.45	.47	3.1
Takes correction badly (sulks, mutters)	**.53**	.31	.12	−.05	.45	.45	15.6
Seems too unconcerned about people to greet	**.53**	.00	.30	.20	.48	.24	2.8
Misbehaves when teacher attends to others	**.52**	.23	−.18	.18	.51	.51	25.9
Wants to dominate during free play	**.52**	.39	−.01	−.13	.50	.49	14.0
Knowingly misbehaves in front of teacher	**.52**	.33	−.11	.22	.60	.58	14.9
Without warning, throws objects	**.51**	.12	.10	.10	.35	.41	6.6
Constantly restless	**.48**	.28	−.09	.22	.51	.47	22.3
Angry look when greeting teacher	**.46**	.11	.27	.23	.43	.22	0.5
Comes out swearing, without provocation	**.46**	.07	.14	.00	.24	.21	0.3
Takes children's belongings (unprovoked)	**.46**	.29	−.05	.19	.46	.53	10.3
Has stolen objects from the classroom	**.44**	.00	.01	.20	.28	.35	3.6
Tries to cheat in games	**.41**	.28	−.01	.18	.38	.43	3.4
Lies to avoid blame or punishment	**.40**	.19	−.16	.30	.42	.45	21.5
Scale 2: Attention Seeking (internal consistency = .87[e])							
Overly friendly with teacher	−.08	**.74**	−.17	.01	.57	.41	7.4
Insists on sitting next to teacher	−.19	**.74**	.11	−.06	.51	.34	7.7
Welcomes you loudly	−.07	**.67**	−.11	−.07	.45	.38	19.8
Much too talkative with teacher	.28	**.63**	−.16	−.10	.60	.49	14.0
Uses devices to gain teacher's attention	.30	**.56**	−.11	−.05	.51	.46	25.3
Clings to teacher when greeting	−.04	**.53**	.09	−.15	.28	.27	7.3
Seeks help when not needed	.17	**.47**	−.06	.04	.31	.37	10.9
Cries when corrected	.15	**.46**	.27	−.02	.32	.26	15.8
Tells on others to gain teacher's favor	.28	**.43**	.10	.09	.35	.33	13.1
Answers questions before taking time to think	.27	**.43**	−.19	−.12	.37	.37	21.4
Answers questions except when in a bad mood	.37	**.43**	.16	−.28	.40	.23	7.9
Moves quickly between free play activities	.25	**.42**	.00	.12	.34	.37	17.4
Scale 3: Reticence/Withdrawal (internal consistency = .92[e])							
Too timid to ask for help	−.14	−.04	**.70**	.09	.54	.48	5.1
Needs encouragement to join in games	−.06	.04	**.68**	.14	.52	.47	14.1
Too withdrawn to come forward for jobs	.06	−.07	**.65**	.27	.56	.49	2.8
Never any trouble because so timid	−.21	.05	**.64**	−.01	.38	.42	4.4
Won't get involved in games	.21	.10	**.63**	.27	.39	.38	2.7
Sits so quietly you don't know if attending	−.16	−.15	**.63**	.06	.49	.42	8.6
Freezes up and doesn't answer questions	−.02	−.12	**.62**	.20	.49	.44	8.1
Afraid to budge during teacher directed time	−.20	−.05	**.62**	.21	.51	.38	2.9
Does not stand up for self during conflicts	−.28	.08	**.60**	.19	.49	.39	9.6
Seems afraid to try when working with hands	−.09	.10	**.59**	.30	.49	.36	3.6
Usually plays by self during free play	−.14	.28	**.57**	.17	.43	.43	8.1
Needs assistance to engage in free play	.18	.22	**.57**	.10	.43	.34	4.9
Shy, difficult to get to speak to teacher	−.07	−.20	**.57**	.10	.43	.39	5.1
Rarely smiles at teacher	.31	−.19	**.54**	.12	.45	.30	3.1
Is too timid to join in during free play	−.05	.04	**.54**	.24	.39	.35	1.4
Allows others to push ahead in lines	−.07	.00	**.54**	.06	.31	.30	7.8
Avoids eye contact with teacher	.32	−.04	**.53**	.23	.50	.38	4.8
Waits for teacher to greet first	.20	−.01	**.52**	−.10	.28	.33	22.6
Wants teacher interest but holds back	.00	−.02	**.51**	−.04	.26	.34	12.1

Table 2 (*continued*)

Item description[a]	Pattern loadings[b]				Communality	Item-scale *r*[c]	% prevalence[d]
	1	2	3	4			
Sometimes wanders off by self	.08	.28	**.46**	.21	.38	.37	11.3
Ignores all other children	.23	−.11	**.45**	.22	.37	.23	1.6
At times does not participate in activities	.36	.08	**.44**	.07	.35	.28	20.0
Aloof, seldom says anything to teacher	.03	−.24	**.43**	.20	.32	.29	6.0
Involved in activities only with adult help	.11	.33	**.40**	.34	.47	.33	5.9
Scale 4: Low Energy (internal consistency = .77[e])							
Cannot work up energy to face anything new	−.03	.16	.24	**.60**	.50	.41	1.6
Too lethargic to ask for help	−.04	.02	.21	**.59**	.43	.42	1.9
Lacks interest, just sits	.20	−.03	.39	**.56**	.63	.49	6.7
Lacks physical energy when working with hands	−.11	.06	.26	**.55**	.30	.31	1.6
Doesn't complete projects	.12	.35	.12	**.55**	.49	.34	11.0
Seldom gets involved in any activities	−.09	−.01	.26	**.53**	.38	.24	3.8
Sluggish, apathetic in games	.00	−.09	.39	**.50**	.47	.41	1.6
Appears to live in a dream world	.09	.08	.26	**.50**	.41	.39	6.1
Hard to get started when working with hands	.24	.08	.22	**.49**	.46	.33	10.9
Not shy but rarely answers questions	.18	−.18	.12	**.48**	.34	.25	11.1
Too lacking in energy to be troublesome	−.24	.17	.35	**.46**	.41	.33	1.5
Listless, seems unmotivated	.20	−.01	.34	**.45**	.45	.47	4.3

Note.

[a] Descriptions incorporate item content and relevant situational contexts. Item content and contexts are abbreviated for convenient presentation.

[b] Values are promax-rotated pattern loadings at $k = 2$, where hyperplane count is maximized. Salient pattern loadings (≥.40) are bolded. $N = 800$ comprising the random exploratory analysis sample.

[c] Each correlation reflects the relationship between an item and the sum of the other items composing a given scale, where item distributions were standardized to unit-normal form.

[d] Entries indicate the percentage of children for whom the item behavior is scored present. Values are based on the combined exploratory and confirmatory analyses samples ($N = 1600$).

[e] Reliability is based directly on expected a posteriori estimates following item response theory scoring for the full calibration sample ($N = 1600$) including both exploratory and confirmatory samples.

3. Results

3.1. Exploratory analyses

As hypothesized, the minimum prevalence rate for any of the 22 positive behavior items was > 50% (viz., 51.9%), and these items were omitted from subsequent analyses. MAP for the remaining 122 potential problem behavior items suggested that as many as 8 components might be extracted from the smoothed tetrachoric matrix. Therefore, 1- through 8-component models were tested against the stated criteria. Having satisfied all criteria, the 4-component, promax-rotated ($k = 2$) model was selected as ideal, whereas models extracting additional components produced under-identified and unreliable dimensions, and those extracting fewer essentially compressed the 4-component model into fewer dimensions. Per Comrey (1988), 7 items having multiple salient loadings and 3 items yielding item-total scale correlations <.20 (thus suppressing internal consistency and discrimination) were eliminated. The remaining 80 items were retained, whereas the earlier Head Start ASPI solution retained 73 items (Noone-Lutz et al., 2002).

Table 2 presents rotated pattern loadings, final communalities, item-total scale correlations, and prevalence (in the full national sample) for retained items. Also posted for each scale (see the centered headings) is its internal consistency (as computed from estimated IRT scores, detailed in text that follows). Based on item content and the patterns of descending loadings, the scales were named Aggression (32 items; *M* behavioral prevalence = 9.4%), Attention Seeking (12 items; *M* prevalence = 14.0%), Reticence/Withdrawal (24 items; *M* prevalence = 7.4%), and Low Energy (12 items; *M* prevalence = 5.2%). Table 3 displays scale intercorrelations, which suggest the possibility of a second-order hierarchical structure emulating the classic, theoretical poles of Overactivity versus Underactivity (Achenbach, 2009; McDermott, 1993; Noone-Lutz et al., 2002). To examine this prospect and to decompose the variance components for the four first-order scales, Table 3 presents the second-order dimensional structure (based on iterated common factoring with squared multiple correlations on the principal diagonal, per Snook & Gorsuch, 1989, for analyses involving <40 variables). As anticipated, the Aggression and Attention Seeking scales form a second-order factor aptly termed Overactivity, while the Reticence/Withdrawal and Low Energy scales form a second-order factor termed Underactivity. Scaled scores for Overactivity and Underactivity correlated −.08. Moreover, the variance decomposition demonstrated that each of the four first-order scales retains a significant amount of variance that is both reliable and unique (45% on the average).

3.2. Confirmatory analyses

The hierarchical dimensional structure (two correlated second-order dimension, each associated with two correlated first-order dimensions) was assessed for the confirmatory sample, as represented by 1 duplet and 26 triplet item parcels. Fit of

Table 3
Second-order dimensional structure and variance components of the Adjustment Scales for Early Transition in Schooling.

First-order scale	Correlation[a]			Rotated factor loading[b]		Proportion of variance[c]		
	1	2	3	Overactivity	Underactivity	Common	Error	Specific
1 Aggression				**.70**	.05	.50	.04	**.46**
2 Attention Seeking	.61			**.67**	−.11	.45	.13	**.42**
3 Reticence/Withdrawal	.12	−.02		−.17	**.56**	.33	.08	**.59**
4 Low Energy	.38	.16	.52	.36	**.52**	.42	.23	**.35**
Average						.43	.12	.45

Note. $N = 800$.

[a] Intercorrelations are based on precision-weighted scales as derived through exploratory analysis reported in Table 2.

[b] Values are promax-rotated pattern coefficients at $k = 2$, where hyperplane count is maximized. Salient pattern loadings (≥.40) are bolded. Correlation between second-order precision-weighted scales = −.06.

[c] Whereas the total proportion of common variance conveyed by a scale is expressed by final communality estimates derived in second-order common factoring, specific variance indicates the proportion of variance which is both reliable and unique to a particular scale. Specific variance is calculated by subtracting communality for a scale from its internal consistency index (posted in Table 2). Specific variance values that exceed error variance (where error variance = 1 − internal consistency) are considered significant and are bolded. The sum of row entries for variance components = 1.0.

this model was adequate, Satorra–Bentler $\chi^2(321) = 743.82$, CFI = .917, and RMSEA = .041 (90% CL = .037/.044). To test longitudinal replication of the structure, analysis was repeated for all confirmatory sample participants in PreK 1 plus PreK 2 and K plus 1st Grade, respectively. Each analysis involved 400 children. Again, adequate fit of the model was supported, $\chi^2(321) = 570.76$ for the PreK children and $\chi^2(321) = 565.09$ for the post-PreK sample. Other fit statistics for both samples were identical at CFI = .907 and RMSEA = .044 (90% CL = .038/.050). The similarity of fit statistics across developmental levels indicates that the two levels do not differ (Chen, 2007). Further, multiple-group tests of the pattern of loadings (i.e., configural invariance) and the magnitude of loadings (i.e., metric invariance) revealed nonsignificant statistical and practical differences between chi-square

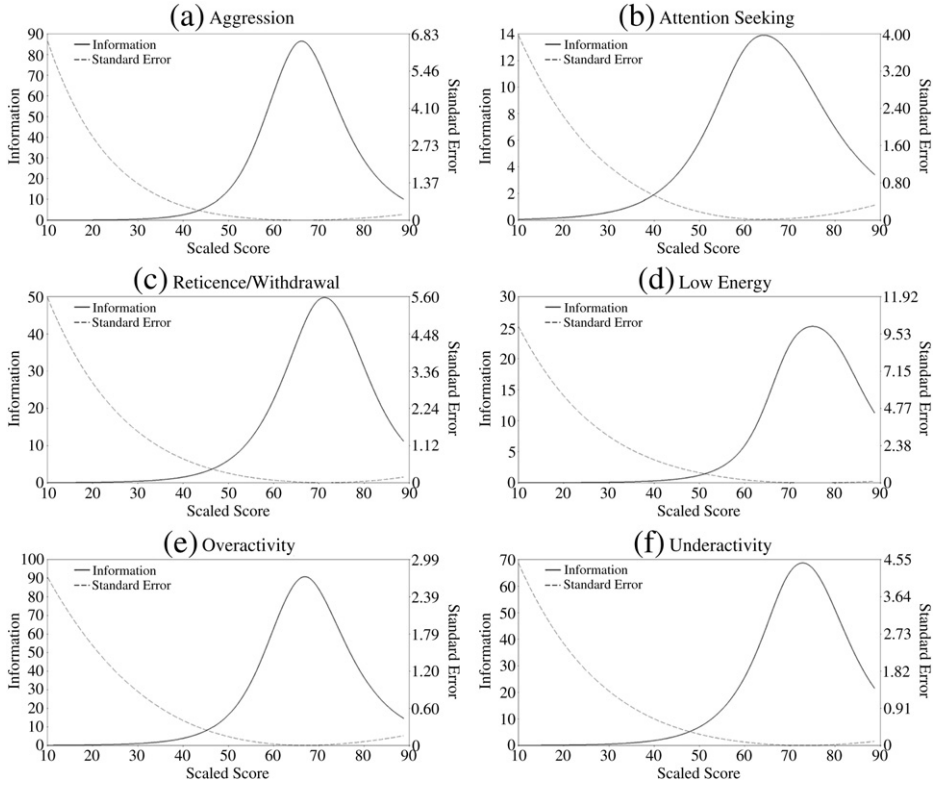

Fig. 1. Distributions of estimated information functions and standard errors for ASETS scales.

and CFI values, respectively, across the developmental levels. Consequently, the total confirmatory sample structure is deemed properly representative of the structure across levels.

3.3. Scaling and scoring

In order to equate vertically each dimension across developmental levels, multiple-group DIF analyses identified approximately one-third of the items comprising each dimension at each adjacent level to serve as linking items. Thus, for example, the 32 Aggression items at PreK 1 were linked by 11 common nonDIF items to PreK 2, another 11 nonDIF items from PreK 2 were linked in common to K, and so on, where the linking items were partly selected so as to represent (via distributed logits) the points across the dimensional continuum. For each ASETS dimension, multiple-group vertical equating was applied and contrasted the one-parameter logistic and two-parameter logistic models. Based on χ^2 likelihood tests, the two-parameter logistic model was found superior for all dimensions.

Longitudinally, the equated first-order scales yielded as follows: (a) Aggression = 95 items, with mean threshold = 1.85, slope = 2.87, information (information is the inverse of measurement error or $1/SE^2$) = 25.65, and maximum information = 36.50 at $\theta = 1.63$; (b) Attention Seeking = 36 items, mean threshold = 1.83, slope = 1.38, information = 6.66, and maximum information = 13.89 at $\theta = 1.38$; (c) Reticence/Withdrawal = 72 items, mean threshold = 2.16, slope = 1.72, information = 11.94, and maximum information = 49.79 at $\theta = 2.13$; and (d) Low Energy = 36 items, mean threshold = 2.82, slope = 1.86, information = 3.33, maximum information = 25.21 at $\theta = 2.50$. The two second-order scales yielded values as follows: (a) Overactivity = 131 items, mean threshold = 2.07, slope = 1.74, information = 27.57, maximum information = 90.47 at $\theta = 1.75$ and (b) Underactivity = 108 items, mean = threshold 2.42, slope = 1.06, information = 13.26, maximum information = 68.66 at $\theta = 2.25$. For any given item, the smallest slope = 0.47, which approximates a factor loading = .423. Each ASETS scale was scored via Bayesian EAP with mean = 50 and standard deviation = 10 for the Prek 1 reference level.

3.4. Reliability

As derived directly from EAP IRT scores (and their *SE*s), reliability coefficients for Aggression, Attention Seeking, Reticence/Withdrawal, and Low Energy were .96, .87, .92, and .77, respectively. The reliability coefficient for Overactivity was .97, and for Underactivity was .93. Figs. 1a–f illustrate the overplots of total scale information and the standard error for each scale. For each figure, the positions where the information function crosses the standard error function mark the boundaries within which test scores may be reliably interpreted. It is apparent that in every instance, except Low Energy, scores the mean (which translate to ≥50 *SS* points) are accurate enough to support distinctions between adjustment levels. For Low Energy, scores ≥ $^1/_4$ of a standard

Table 4
Relationships between Prekindergarten 2 spring ASETS scores and concurrent criterion measures.

Criterion measure	ASETS scale						
	Aggression	Attention Seeking	Reticence/ Withdrawal	Low Energy	Overactivity	Underactivity	% explainable variance[a]
Student–Teacher Relationship Scale (teacher rating)							
Closeness ($n = 2747$)	−.21 (8.1%)	.06 (3.1%)	−.41 (28.1%)	−.30 (21.3%)	−.16 (8.3%)	−.43 (33.6%)	74.9
Conflict ($n = 2743$)	.66 (56.3%)	.38 (28.4%)	.13 (3.6%)	.21 (6.8%)	.64 (54.9%)	.18 (4.7%)	83.1
Parent rating							
Total Behavior Problems ($n = 2626$)	.21 (4.7%)	.10 (3.7%)	.08 (6.2%)	.12 (5.3%)	.20 (4.7%)	.10 (5.8%)	90.9
Peabody Picture Vocabulary Test, Third Edition (direct assessment)							
Receptive Vocabulary ($n = 2699$)	−.02 (0.9%)	.00† (0.0%)†	−.12 (6.0%)	−.11 (2.7%)	−.03 (0.9%)	−.13 (6.5%)	61.3
Woodcock–Johnson III Test of Achievement (direct assessment)							
Letter–Word Identification ($n = 2700$)	−.10 (2.9%)	−.01† (0.1%)†	−.17 (3.3%)	−.15 (4.6%)	−.09 (3.1%)	−.19 (4.7%)	75.6
Applied Problems ($n = 2683$)	−.12 (2.9%)	−.04† (0.2%)†	−.17 (4.0%)	−.15 (3.0%)	−.11 (1.8%)	−.19 (4.8%)	78.4
Pre-Academic Skills ($n = 2683$)	−.15 (4.0%)	−.05† (0.4%)†	−.19 (6.2%)	−.18 (5.7%)	−.14 (4.0%)	−.22 (8.1%)	78.0
Spelling ($n = 2701$)	−.16 (3.3%)	−.06 (0.6%)	−.15 (4.0%)	−.14 (3.3%)	−.15 (3.6%)	−.17 (5.5%)	85.0

Note. Nonparenthetical entries are Pearson product moment correlations. Parenthetical entries indicate the percentage of variance in the respective criterion measure scores between children within classrooms that is accounted for by a given ASETS scale score. Values equal 1 − reduction in the intraclass correlation (100) as estimated via hierarchical linear modeling. Each two-level random coefficients model entered a given ASETS scale as the covariate. All correlations and fixed effects associated with ASETS scales are significant statistically at $p < .01$ unless indicated † (nonsignificant). ASETS = Adjustment Scales for Early Transition in Schooling.

[a] Total percentage of potentially explainable variance between children within classrooms. Values equal 1 − intraclass correlation (100), where the intraclass correlation was estimated via hierarchical linear modeling. Each two-level, unconditional means model applied random intercepts for classrooms, where the random effect was significant at $p < .001$.

Table 5
Relationships between First Grade spring ASETS scores and concurrent criterion measures.

Criterion measure	ASETS scale						
	Aggression	Attention Seeking	Reticence/ Withdrawal	Low Energy	Overactivity	Underactivity	% explainable variance[a]
Student–Teacher Relationship Scale (teacher rating)							
Closeness ($n=3058$)	−.24 (8.0%)	.02† (3.5%)†	−.41 (32.5%)	−.35 (22.0%)	−.21 (10.0%)	−.44 (40.7%)	76.7
Conflict ($n=3050$)	.74 (59.4%)	.49 (30.2%)	.16 (1.4%)†	.35 (8.4%)	.73 (63.4%)	.28 (11.2%)	82.2
Positive Relationship ($n=3059$)	−.64 (39.4%)	−.34 (12.7%)	−.30 (8.7%)	−.41 (18.8%)	−.62 (42.1%)	−.40 (20.8%)	77.9
Parent rating							
Total Behavior Problems ($n=2900$)	.26 (6.3%)	.16 (3.9%)	.12 (6.6%)	.21 (7.2%)	.26 (5.4%)	.18 (10.8%)	88.4
Peabody Picture Vocabulary Test, Third Edition (direct assessment)							
Receptive Vocabulary ($n=2883$)	−.00† (0.7%)†	.01† (0.7%)†	−.13 (1.1%)†	−.11 (4.4%)	−.01† (1.4%)	−.14 (2.7%)	56.8
Woodcock–Johnson III Test of Achievement (direct assessment)							
Basic Reading Skills fiction ($n=2873$)	−.17 (7.2%)	−.09 (1.2%)	−.19 (11.0%)	−.23 (19.6%)	−.16 (2.5%)	−.24 (19.7%)	49.2
Word Attack ($n=2875$)	−.16 (4.5%)	−.09 (0.7%)†	−.19 (13.6%)	−.21 (13.4%)	−.15 (1.7%)	−.22 (19.0%)	56.3
Quantitative Concepts ($n=2877$)	−.13 (0.8%)†	−.07 (0.3%)†	−.20 (4.7%)	−.23 (14.0%)	−.12 (0.8%)†	−.24 (14.1%)	72.9

Note. Nonparenthetical entries are Pearson product moment correlations. Parenthetical entries indicate the percentage of variance in the respective criterion measure scores between children within classrooms that is accounted for by a given ASETS scale score. Values equal 1 − reduction in the intraclass correlation (100) as estimated via hierarchical linear modeling. Each two-level random coefficients model entered a given ASETS scale as the covariate. All correlations and fixed effects associated with ASETS scales are significant statistically at $p<.01$ unless indicated † (nonsignificant). ASETS = Adjustment Scales for Early Transition in Schooling.

[a] Total percentage of potentially explainable variance between children within classrooms. Values equal 1 − intraclass correlation (100), where the intraclass correlation was estimated via hierarchical linear modeling. Each two-level, unconditional means model applied random intercepts for classrooms, where the random effect was significant at $p<.001$.

deviation above the mean are sufficiently accurate. This accuracy is particularly important because, as ordinarily applied, users would want to be able to discriminate between adjustment and maladjustment and between varied levels of maladjustment (e.g., *SS* 60–69 are considered at risk or subclinical and *SS* ≥ 70 maladjusted; McDermott & Weiss, 1995).

Table 6
Increased risk of first-grade teacher-reported academic nonproficiency associated with Prekindergarten 2 ASETS scores.

ASETS scale	Odds ratio	Odds ratio 95% confidence limits (lower/upper)	% Risk increment[a]
First-Grade Language and Literacy Ability($n=2188$, estimated variance between children = 84.1%)[b]			
Aggression	1.03*	1.02/1.04	3.1
Attention Seeking	1.01	0.99/1.02	0.4
Reticence/Withdrawal	1.05*	1.03/1.05	4.0
Low Energy	1.07*	1.05/1.09	7.2
Overactivity	1.03*	1.02/1.04	2.8
Underactivity	1.05*	1.03/1.06	4.8
First-Grade Mathematics Ability($n=2182$, estimated variance between children = 87.4%)[b]			
Aggression	1.04*	1.02/1.05	3.6
Attention Seeking	1.01	0.99/1.03	1.3
Reticence/Withdrawal	1.04*	1.02/1.05	3.9
Low Energy	1.06*	1.04/1.08	6.2
Overactivity	1.03*	1.02/1.04	3.2
Underactivity	1.04*	1.03/1.06	4.5
First-Grade Social Science Ability($n=2177$, estimated variance between children = 84.8%)[b]			
Aggression	1.04*	1.03/1.06	4.4
Attention Seeking	1.01	0.99/1.02	0.7
Reticence/Withdrawal	1.04*	1.03/1.06	4.4
Low Energy	1.07*	1.05/1.10	7.4
Overactivity	1.04*	1.02/1.05	3.7
Underactivity	1.05*	1.04/1.07	5.3

Note. Entries are based on generalized multilevel logistic regression modeling using adaptive quadratures with parameter starting values drawn from pseudo-likelihood estimation of the multilevel generalized linear model. A separate model was constructed for each academic performance area. ASETS = Adjustment Scales for Early Transition in Schooling.

[a] Values = (odds ratio − 1)100 and express the percentage increase in risk of future academic nonproficiency per each 1 scaled score increase in the respective ASETS scale.

[b] Based on unconditional models, values = the intraclass correlation (100), where the intraclass correlation = estimated coefficient for random intercepts / (estimated coefficient for random effects + estimated coefficient for residuals).

* $p<.001$.

3.5. Criterion-related validity

Table 4 shows concurrent relations between spring, PreK 2 ASETS scores and other relevant measures, while Table 5 shows similar relations for spring 1st Grade students. Whereas all statistically significant correlations are in the direction that would be expected, ASETS scores evince moderate to strong relations with other teacher measures and weak relations with parent measures and direct assessments of achievement. Given the nested nature of the data, the last column in each table lists the percentage of criterion measure variance that actually pertains to children's individual differences, whereas the parenthetical values reveal how much of that variance is accounted for by a given ASETS scale. Thus, for instance, while Table 4's last column entry for the STRS Closeness scale indicates that 74.9% of score variance evolves from children's individual differences (rather than teacher or classroom characteristics), it is found that 33.6% of that variance is predictable from children's ASETS general Underactivity scores, and more particularly, 28.1% is predictable from Reticence/Withdrawal and 21.3% is predictable from Low Energy scores. Parents seem to be somewhat more sensitive to reporting underactive-type problems, and teachers to overactive-type problems. Also, ASETS Reticence/Withdrawal and Low Energy scores are somewhat more effective in accounting for individual differences in academic performance.

Table 6 shows the degree to which ASETS scores forecast increased risk of teacher-assessed academic nonproficiency 2 years later, at the close of 1st Grade. Again, because outcomes are nested within teachers/classrooms, odds ratios were estimated through multilevel modeling. For each outcome (Language and Literacy, Mathematics, Social Science), all ASETS scales except Attention Seeking were able to indicate significant risk for subsequent nonproficiency. To interpret results, one refers to the last column where, for example, the 3.1 entry for Aggression means that, for every 1 *SS* point increase in ASETS Aggression during spring PreK 2, there is a 3.1% increment in the risk of Language and Literacy nonproficiency at the conclusion of 1st Grade. Similarly, the 7.4 entry for Low Energy means that, for every 1 *SS* point increase in ASETS Low Energy during spring PreK 2, there is a 7.4% increment in the risk of Social Science nonproficiency at the conclusion of 1st Grade.

3.6. Change detection

A measure that would be useful for following children across early school transitions must be reasonably sensitive to changes as time passes. Multilevel individual growth-curve modeling was applied to test sensitivity to change and to reveal the direction and trends for change for each scale. Table 7 shows the statistically significant change parameters. Each value indicates the estimated magnitude and direction of change in ASETS *SS* points per developmental level (year). To illustrate, the entry for Aggression shows that change is linear over time, with a decrease of 0.60 *SS* point per year. Characteristically, most scales, on average, exhibit a steady drop in problem behavior as children move through the transitions, while Attention Seeking displays a constant increase. Most scales also quadratic and cubic trajectories, as well as linear. Aggression manifests distinctly linear change, as does Overactivity, as would follow from the fact that most Overactivity items are from the Aggression scale. In summation, it is clear that each scale is capable of detecting change, most of that change being complex higher-order rather than

Table 7
Linear and higher-order growth parameters of the Adjustment Scales for Early Transition in Schooling (ASETS) over four years.

ASETS scale	Parameter estimate for change (standard error)		
	Linear	Quadratic	Cubic
Aggression	−0.5991***		
	(0.0716)		
Attention Seeking	1.3919***	1.3009***	0.2344*
	(0.4856)	(0.4361)	(0.0975)
Reticence/Withdrawal	−1.8166***	−2.2514***	−0.6269***
	(0.4865)	(0.4358)	(0.0974)
Low Energy	−0.9750***	−2.1628***	−0.5585***
	(0.4698)	(0.4219)	(0.0944)
Overactivity	−0.4494***		
	(0.0751)		
Underactivity	−1.6173	−2.4517***	−0.6893***
	(0.5241)	(0.4637)	(0.1033)

Note. Values are estimated through multilevel individual growth-curve modeling. Models for the 6 ASETS scales incorporated statistically significant random coefficients for intercepts and linear slopes. Higher-order random slopes were uniformly nonsignificant and thus excluded. Only statistically significant fixed effects parameters are reported unless nonsignificant linear estimates appear as requisite for subsequent sequential *F* tests associated with higher-order estimates. Specification for the full model was $\hat{Y}_{ijk} = \gamma_{000} + \gamma_{100}\text{Time}_{ijk} + \gamma_{200}\text{Time}^2_{ijk} + \gamma_{300}\text{Time}^3_{ijk} + (\mu_{00k} + \mu_{10k}\text{Time}_{ijk}) + (\mu_{0jk} + \mu_{1jk}\text{Time}_{ijk}) + r_{ijk}$, although terms associated with nondifficant fixed effects for a given model were dropped. Parameters reflect change in ASETS scaled scores per year through 4 years spanning Prekindergarten 1 to First Grade. $N = 3077$.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

simple linear. Moreover, problem behavior will tend to decrease temporally, except increases are more common for Attention Seeking.

3.7. Differential change detection

The ability to detect change opens up many paths to practitioners and researchers. Here we select one rich avenue of inquiry that demonstrates how ASETS scores can be used to discover the nature of long-term socioemotional adjustment trajectories that distinguish children who reach successful outcomes versus those who reach unsuccessful outcomes at the end of 1st Grade. For this purpose, we expanded each of the unconditional multilevel models discussed above to break out the average change trajectories per different blocking variables. For example, the model for aggression is supplemented by entry of a blocking variable reflecting whether children are in the lowest quartile of WJ III Basic Reading skills (i.e. nonproficient) versus above the lowest quartile (i.e. proficient). The results are illustrated in Figs. 2a–f. The change trajectories reflect the interaction between the blocking variable and time, and are controlled for the effects of children's sex, ethnicity, primary language and special needs status, and urban residence. Age was not found to be a statistically significant predictor in any model.

Fig. 2a shows what occurs when Aggression trajectories are distinguished by whether children perform nonproficiently (the lowest quartile) versus proficiently in spring 1st Grade on WJ III Basic Reading Skills. The eventual nonproficient children showed more aggression all along (effect size [*ES*] = 0.42; see Fantuzzo, Gadsden, & McDermott, 2011, p. 789, on computation), with a

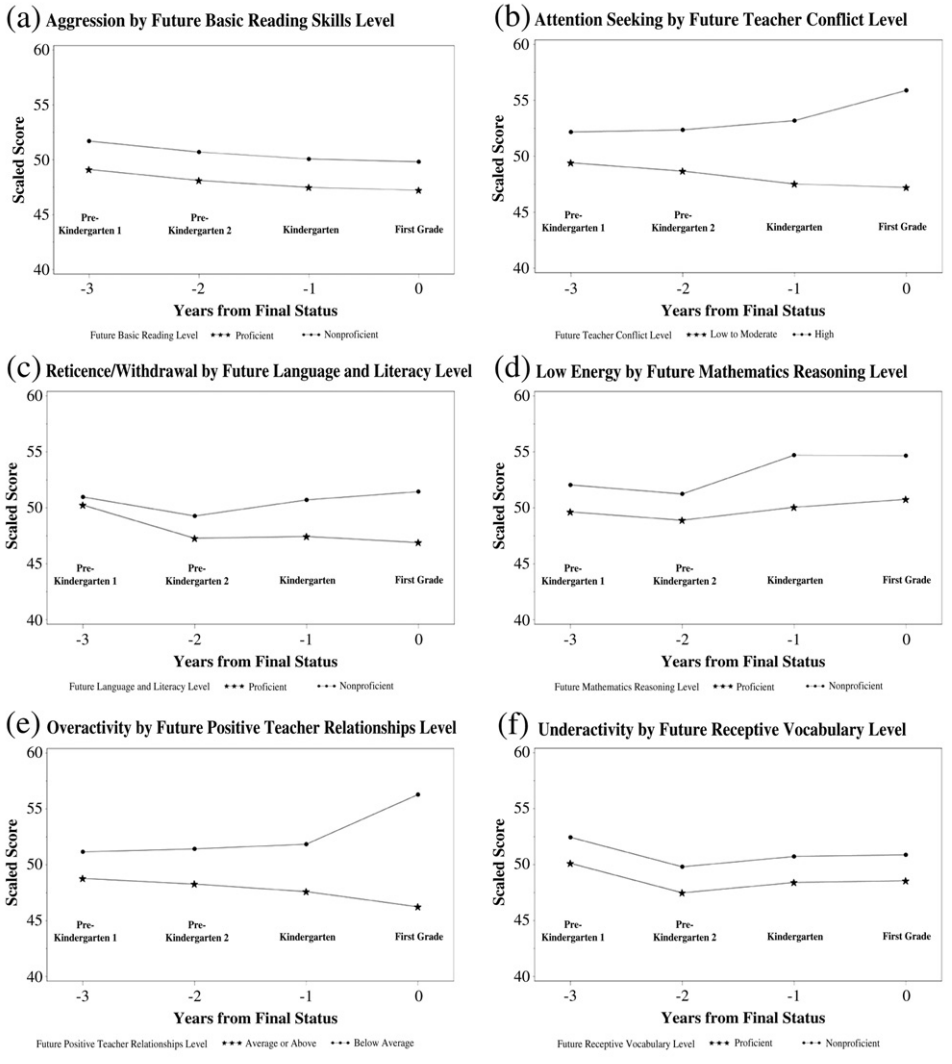

Fig. 2. Estimated average growth trajectories for ASETS scales.

common decrease over time for all children. Fig. 2b tracks Attention Seeking changes for children who end up manifesting high levels of Conflict with teachers (highest quartile on the STRS) versus low or moderate Conflict. The trajectories begin to separate with the PreK 1 to PreK 2 transition and ultimately produce a large departure by the end of 1st Grade ($ES = 0.91$). Figs. 2c and d display somewhat similar patterns. Fig. 2c tracks Reticence/Withdrawal as differentiated by children who are eventually reported by teachers as nonproficient versus proficient in Language and Literacy, and Fig. 2d tracks Low Energy behavior by whether nonproficiency versus proficiency is realized for WJ III Mathematics Reasoning. Both models show a steady decrement in problem behavior across the PreK years, but a marked elevation in problem behaviors with the K transition, finally attaining $ES = 0.91$ for Reticence/Withdrawal and 1.09 for Low Energy. Fig. 2e depicts that children, in general, will tend to decline in observed Overactivity (although those later seen as having Positive Relationships with 1st Grade teachers also display fewer overactivity problems all along), but that with transition to 1st Grade, overactivity spikes noticeably for those ultimately having relatively poor teacher relationships ($ES = 1.39$). Finally, in following Underactivity over the transitions, where children are blocked by their future performance on the PPVT-III, both the nonproficient (lowest quartile) and proficient in receptive vocabulary display a unique cubic change pattern, with Underactivity decreasing over the PreK 1–PreK 2 transition, then leveling off as children proceed into K, and decreasing again over the K-1st Grade transition ($ES = 0.40$ at each developmental level). All of these findings illustrate that ASETS measures have the capacity to detect differential socioemotional change patterns across the transitions, even when controlled for important alterative factors, such as sex, ethnicity, and special needs.

4. Discussion

The ASETS dimensional structure does not resemble that of the Head Start ASPI (Noone-Lutz et al., 2002) except for the fact that both exhibit a hierarchical structure with higher-order Overactivity and Underactivity factors. Any disparities between the Head Start ASPI and the ASETS structures are unsurprising inasmuch as the populations, purposes, and methodologies are very different. ASPI was developed exclusively for Head Start as based on a single time point assessment with 829 children in one urban location. The 46 respondent teachers assessed all of the children in their full-day, public school affiliated, classrooms ($M = 18.0$ children per classroom). The average child age was 4.6 years with 75% being African American. In contrast, ASETS employs longitudinal assessments for a much larger and nationally representative sample of both Head Start and non-Head Start prekindergarteners, kindergarteners, and first graders. Per national trend, the ASETS sample was more diverse (with 70% non-African American and 60% part-day PreK obtained from over 2300 classrooms affiliated with both public and private organizations). The HSIS sampling design also limited ASETS administration to an average of only 1.6 randomly selected PreK children per classroom, but it extended to a larger proportion of younger children, so that the average age at entry was 4.0 years.

The methodological distinctions between this work and the original ASPI are perhaps the most striking because ASETS was able to take advantage of the longitudinal perspective over the transition years. It also was able to benefit from the technology emergent from correlational estimates assuming underlying item-level continua, confirmation through structural equations modeling, IRT vertical scale equating, and Bayesian score estimates. Moreover, various multilevel modeling methods made it possible to produce more accurate tests of concurrent and predictive validity and tests of sensitivity to differential change over the transition periods.

One might ponder the practical utility of having an instrument that features both broad-band second-order dimensions (Overactivity and Underactivity) and narrow-band first-order dimensions (Aggression, Attention Seeking, Reticence/Withdrawal, Low Energy). The theoretical propriety of the second-order dimensions is long established (Achenbach, 2009; McDermott, 1986). Omnibus behavioral and personality measures have traditionally bifurcated in this fashion, much in the way that cognitive ability measures have traditionally reified a single dimension of general conceptual ability (Lubinski, 2009), but the broad-band dimensions also enable a more steady framework for diversifying instruments across cultural and ethnolinguistic boundaries. Whereas it is much more likely to discover disparities in the exact nature of more specific, narrow-band dimensions across substantially different cultural populations (sampling error alone would tend to militate this variability at the item/behavioral level), it is also more probable that the broad-band Overactivity–Underactivity dichotomy would tend to offer fair promise of generalization across populations (e.g., see Canivez & Beran, 2009; Canivez & Bohan, 2006; George, McDermott, Watkins, Worrell, & Hall, 2012). This finding has potential implications, for example, when an instrument like ASETS is translated to another respondent language and one seeks to statistically equate the devices for common field application and research.

On the other hand, the four first-order scales are particularly useful for both practice and research applications with children at-risk in the general U.S. population and especially where there is an intent to forecast adaptation to the early school transitions and to follow children over the critical years. These scales retain a sufficient amount of unique and reliable variance to support differential evaluation for each child. Aggression particularly is important because it emerges frequently in other research (LeBoeuf et al., 2010; Petitclere et al., 2008) as the primary phenotype for maladjustment. Now with the benefit of a vertically-equated scale, we may observe the common developmental trajectory of aggression to be one of steady decline through 1st Grade. This result makes sense if one considers the socialization effects of behavioral extinction; that is, disruptive activity gets children into trouble and discipline by authorities and adverse reactions of peers will tend to diminish the aggressive inclination. ASETS growth trajectories also show that a fair portion of children sustain aggressiveness over time and these children are more inclined to reap negative outcomes for some future achievements. There is a great deal of controversy in the literature as to whether children in organized care increase or decrease their aggressive tendencies over the transitions (Magnuson et al., 2007; Phillips & Lowenstein, 2011), with the result depending to some extent on the nature and quality of care (McCartney et al., 2010).

It is rather difficult to reconcile all of the counterpoising claims, but inquiry into these claims should profit from instrumentation specifically tailored for the longitudinal perspective.

The behavior assessed by the other first-order ASETS dimensions follows much more complicated, curvilinear pathways as compared to Aggression. Attention Seeking is uniquely incremental over time, whereas both Reticence/Withdrawal and Low Energy predominantly dissipate over time. The characteristic increase in Attention Seeking is curious. As a variant of acting-out behavior, it is quite possible that this phenotype is sometimes conflated with acting out that is distinctly aggressive in nature. This conflation might explain why numerous researchers link increased aggression with early organized child care (Magnuson et al., 2007; Phillips & Lowenstein, 2011). PreK, as in the HSIS study, is a type of child care, and the oft associated later disruptive behavior may actually be the detection of excessive attention seeking that steady exposure to organized care encourages.

Discovery of a Low Energy dimension is reminiscent of recent work by Garner, Marceaux, Mrug, Patterson, and Hodgens (2010). They refer to a constellation of behaviors termed *sluggish cognitive tempo* that is defined by behaviors very similar to Low Energy. The constellation was earlier identified in clinical samples (McBurnett, Pfiffner, & Grivk, 2001) and describes apathetic, unmotivated, and slow reactivity that impairs socialization and cognitive achievement, much as Low Energy portends more limited engagement with teachers and ineffective learning.

4.1. Limitations and future research

The present work is drawn from data intended originally for a national randomized controlled experiment, where it was imperative to sample in regions of the country that would allow sufficient availability of Head Start-eligible children who would be unable to attend Head Start. This fact could be considered a limitation to the extent that certain locations in the country were not sampled because their populations were too small or they were servicing children that were from migratory families or tribal subpopulations. HSIS also excluded children with prior exposure to Early Head Start. One may also consider the variety of ASETS dimensions to be somewhat limited in that they do not cover types of psychopathology frequently studied in clinical or psychiatric populations (e.g. obsessive compulsive disorder, autism). However, we see this not as a limitation but rather as a reflection of the discovery that children in a very large nonclinical population manifest different behavioral patterns, especially as relates to their transition into conventional schooling.

ASETS may open several lines of research. First, to the extent that ASETS used the same item content and format as ASPI, it enables secondary analyses and application of ASPI data using the new scales and scoring routines. Such continuity makes possible the nomothetic comparison of available ASPI data to the broader national population. Second, whereas McDermott et al. (2006) demonstrated that the K-12 ASCA classroom behavioral contexts (rather than items themselves) also retained a meaningful dimensional structure that augmented item-level information, Bulotsky-Shearer, Fantuzzo, and McDermott (2008) also showed that ASPI's situational contexts could yield similar information for Head Start children. Thus, a natural next step for ASETS research is the investigation of the dimensional nature of the contextual structure of problem behaviors over time and their ability to inform developmental trajectories and outcomes. Finally, the multidimensional and transitional features of ASETS invite research on the typological (latent class transition) change of early childhood socioemotional adjustment, as grounded in longitudinally continuous attributes (the ASETS scales).

4.2. Conclusion

ASETS was developed in part as a response to the NRC (2008) call to provide appreciable quality, technically enhanced, and purposeful assessment devices for use in early childhood education. It further attempted to address the serious impediment caused by assessment devices that were not built specifically for populations of children at high risk, devices that are laden with clinical jargon and absent the kinds of item content that can appreciate the subtle nuances of more commonplace nonclinical or subclinical socioemotional phenomena, and devices that are intended for preschoolers but have no psychometric continuity with those intended for school-age children. Most importantly, it is the transition story that ASETS is designed to tell.

References

Achenbach, T. M. (2009). The Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications.. Burlington, VT: University of Vermont.

Aud, S., Husser, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., et al. (2011). The conditions of education 2011 (NCES 2011-033).. Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*, 45–87.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102.

Bauer, D. J., & Curran, P. J. (2006). Multilevel modeling of hierarchical and longitudinal data using SAS: Course notes.. Cary, NC: SAS Institute.

Benner, A. D., & Crosnoe, R. (2011). The racial/ethnic composition of elementary schools and young children's academic and socioemotional functioning. *American Educational Research Journal*, *48*, 621–646.

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.

Briggs-Gowan, M. J., Carter, A. S., Bosson-Heenan, J., Guyer, A. E., & Horowitz, S. M. (2006). Are infant–toddler social–emotional and behavioral problems transient? *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*, 849–858.

Bulotsky-Shearer, R., & Fantuzzo, J. (2004). Adjustment scales for preschool intervention: Extending validity and relevance across multiple perspectives. *Psychology in the Schools*, *41*, 725–736.

Bulotsky-Shearer, R., Fantuzzo, J. F., & McDermott, P. A. (2008). An investigation of classroom situational dimensions of emotional and behavioral adjustment and cognitive and social outcomes for Head Start children. *Developmental Psychology*, *44*, 139–154.
Buss, K. A. (2011). Which fearful toddlers should we worry about? Context, fear regulation, and anxiety risk. *Developmental Psychology*, *47*, 804–819.
Campbell, S. B. (2001). Behavior problems in preschool children.. New York, NY: Guilford.
Campbell, J. M., & James, C. L. (2007). Assessment of social and emotional development in preschool children. In B. A. Bracken, & J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (pp. 111–135) (4th ed.). . Mahwah, NJ: Erlbaum.
Canivez, G. L., & Beran, T. N. (2009). Adjustment Scales for Children and Adolescents: Factorial validity in a Canadian sample. *Canadian Journal of School Psychology*, *24*, 284–302.
Canivez, G. L., & Bohan, K. J. (2006). Adjustment Scales for Children and Adolescents and Native American Indians: Factorial validity generalization for Yavapai Apache youths. *Journal of Psychoeducational Assessment*, *24*, 329–341.
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464–504.
Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*, 754–761.
Dobbs, J., Doctoroff, G. L., Fisher, P. H., & Arnold, D. H. (2006). The association between preschool children's socio-emotional functioning and their mathematical skills. *Applied Developmental Psychology*, *27*, 97–108.
Drotar, D., Stein, R. E. K., & Perrin, E. C. (1995). Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology*, *24*, 184–192.
du Toit, M. (Ed.). (2003). *IRT from SSI, BILOG-MG, MULTILOG, PARSCALE, TESTFACT.*. Lincolnwood, IL: Scientific Software International.
Dumont, R., & Willis, J. O. (2006). Test descriptions and reviews. In E. Fletcher-Janzen, & C. R. Reynolds (Eds.), *The special education almanac* (pp. 39–146). Hoboken, NJ: Wiley.
Dunn, L. M., Dunn, L. L., & Dunn, D. M. (1997). Peabody Picture Vocabulary Test (3rd ed.). .. Circle Pines, MN: American Guidance.
Egger, H. L., & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry*, *47*, 313–337.
Eivers, A. R., Brendgen, M., & Borge, A. I. H. (2010). Stability and change in personality and antisocial behavior across the transition to school: Teacher and peer perspectives. *Early Education and Development*, *21*, 843–864.
Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*, *19*, 401–423.
Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2005). First grade and educational attainment by age 22: A new story. *American Journal of Sociology*, *110*, 1458–1502.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.
Fantuzzo, J. W., Bulotsky, R., McDermott, P. A., Mosca, S., & Noone-Lutz, M. (2003). A multivariate analysis of emotional and behavioral adjustment to preschool and preschool educational outcomes. *School Psychology Review*, *32*, 185–203.
Fantuzzo, J. W., Bulotsky-Shearer, R., Frye, D., McDermott, P. A., McWayne, C., & Perlman, S. (2007). Investigation of social, emotional, and behavioral dimensions of school readiness for low-income, urban preschool children. *School Psychology Review*, *36*, 44–62.
Fantuzzo, J. W., Bulotsky-Shearer, R., Fusco, R. A., & McWayne, C. (2005). An investigation of preschool emotional and behavioral adjustment problems and social–emotional school readiness competencies. *Early Childhood Research Quarterly*, *20*, 259–275.
Fantuzzo, J. W., Gadsden, V. L., & McDermott, P. A. (2011). An integrated curriculum to improve mathematics, language, and literacy for Head Start children. *American Educational Research Journal*, *48*, 763–793.
Fantuzzo, J. W., Manz, P. H., & McDermott, P. A. (1998). Preschool version of the Social Skills Rating System: An empirical analysis of its use with low-income children. *Journal of School Psychology*, *36*, 199–214.
Fantuzzo, J. W., McDermott, P. A., Manz, P. H., Hampton, V., & Burdick, N. A. (1996). Preschool scales of perceived competence and acceptance: Do they work with low-income urban children? *Child Development*, *67*, 1071–1084.
Federal Interagency Forum on Child and Family Statistics (2008). America's children in brief: Key national indicators of well-being, 2008.. Washington, DC: U.S. Government Printing Office.
Feeney-Kettler, K. A., Kratochwill, T. R., & Kettler, R. J. (2011). Identification of preschool children at risk for emotional and behavioral disorders: Development and validation of a universal screening system. *Journal of School Psychology*, *49*, 197–216.
Furr, R. M. (2011). Scale construction and psychometrics for social and personality psychology.. Thousand Oaks, CA: Sage.
Garner, A. A., Marceaux, J. C., Mrug, S., Patterson, C., & Hodgens, B. (2010). Dimensions and correlates of attention deficit/hyperactivity disorder and sluggish cognitive tempo. *Journal of Abnormal Child Psychology*, *38*, 1097–1107.
George, S., McDermott, P. A., Watkins, M. W., Worrell, F., & Hall, T. E. (2012). The assessment of youth psychopathology in Trinidad and Tobago: A cross-cultural construct validity study of the Adjustment Scales for Children and Adolescents (ASCA). *International Journal of Educational and Psychological Assessment*, *10*, 159–178.
Goldsmith, H. G., & Davidson, R. J. (2004). Disambiguating the components of emotion regulation. *Child Development*, *75*, 361–365.
Guilford, J. P. (1956). Psychometric methods.. New York, NY: McGraw-Hill.
Gurin, P., Day, E. L., Hurtado, S., & Gurin, G. (2002). Diversity and higher education: Theory and impact on educational outcomes. *Harvard Educational Review*, *72*, 330–366.
Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly*, *21*, 431–454.
Hall, R. J., Snell, A. F., & Singer Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, *2*, 232–256.
Hau, K. -T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology*, *57*, 327–351.
Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902.
Hemmeter, M. L., & Ostrosky, M. (2006). Social and emotional foundations for early learning: A conceptual model for intervention. *School Psychology Review*, *33*, 583–601.
Horn, W. F., Wagner, A. E., & Ialongo, N. (1989). Sex differences in school-aged children with pervasive attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology*, *17*, 109–125.
Hoyt, W. T., & Kems, M. -D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.
Huston, A. C., & Bentley, A. C. (2010). Human development in social context. *Annual Review of Psychology*, *61*, 411–437.
Kataoka, S. H., Zhang, L., & Wells, K. B. (2002). Unmet need for mental health care among U.S. children. Variation by ethnicity and insurance status. *American Journal of Psychiatry*, *159*, 1548–1555.
Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representatives parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement*, *54*, 757–765.
Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457–477.
Kolker, J., Osborne, D., & Schnurer, E. (2004). Early child care and education: The need for a national policy.. Washington, DC: Center for National Policy.
LeBoeuf, W. A., Fantuzzo, J. W., & Lopez, M. L. (2010). Measurement and population miss-fits: A case study on the importance of using appropriate measures to evaluate early childhood interventions. *Applied Developmental Science*, *14*, 45–53.
Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the questions, weighing the merits. *Structural Equation Modeling*, *9*, 151–173.

Lubinski, D. (2009). Exceptional cognitive ability: The phenotype. *Behavioral Genetics*, *39*, 350–358.
Lynch, R. G. (2004). Exceptional returns: Economic, fiscal, and social benefits of investment in early childhood development.. Washington, DC: Economic Policy Institute.
Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preparation and performance? *Economics of Education Review*, *26*, 33–51.
Manz, P. H., Fantuzzo, J. W., & McDermott, P. A. (1999). The parent version of the Preschool Social Skills Rating Scale: An analysis of its use with low-income, ethnic minority children. *School Psychology Review*, *28*, 493–504.
Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29*, 322–346.
Mathiesen, K. S., Sanson, A., Stoolmiller, M., & Karevold, E. (2009). The nature and predictors of undercontrolled and internalizing problem trajectories across early childhood. *Journal of Abnormal Child Psychology*, *37*, 209–222.
McBurnett, K., Pfiffner, L. J., & Grivk, P. J. (2001). Symptom properties as a function of ADHD type: An argument for continued study of sluggish cognitive tempo. *Journal of Abnormal Child Psychology*, *29*, 207–213.
McCartney, K., Burchinal, M., Clarke-Stewart, A., Bub, K. L., Owen, M. T., & NICHD Early Child Care Research Network (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology*, *46*, 1–17.
McDermott, P. A. (1986). The observation and classification of exceptional child behavior. In R. T. Brown, & C. R. Reynolds (Eds.), *Psychological perspectives on childhood exceptionality: A handbook* (pp. 136–180). New York, NY: Wiley.
McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment*, *5*, 413–424.
McDermott, P. A. (1994). *National profiles in youth psychopathology: Manual of Adjustment Scales for Children and Adolescents.* Philadelphia, PA: Edumetric and Clinical Science.
McDermott, P. A., Fantuzzo, J. W., Waterman, C., Angelo, L. E., Warley, H. P., Gadsden, V. L., et al. (2009). Measuring preschool cognitive growth while it's still happening: The Learning Express. *Journal of School Psychology*, *47*, 337–366.
McDermott, P. A., Steinberg, C. M., & Angelo, L. E. (2006). Situational specificity makes the difference in assessment of youth behavior disorders. *Psychology in the Schools*, *42*, 121–136.
McDermott, P. A., & Weiss, R. V. (1995). A normative typology of healthy, subclinical, and clinical behavior styles among American children and adolescents. *Psychological Assessment*, *7*, 162–170.
McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27*, 82–99.
McWayne, C., & Chang, K. (2009). A picture of strength: Preschool competencies mediate the effects of early behavior problems on later academic and social adjustment for Head Start children. *Journal of Applied Developmental Psychology*, *30*, 273–283.
Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3–31.
Mooijaart, A. (1983). Two kinds of factor analysis for ordered categorical variables. *Multivariate Behavioral Research*, *18*, 423–441.
Muthén, B. (1987). LISCOMP: Analysis of linear structural equations with a comprehensive measurement model.. Chicago, IL: Scientific Software International.
Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, *63*, 729–757.
National Research Council (2008). Early childhood assessment: Why, what, and how.. Washington, DC: National Academies Press.
Noone-Lutz, M., Fantuzzo, J. F., & McDermott, P. A. (2002). Multidimensional assessment of emotional and behavioral adjustment of low-income preschool children: Development and initial validation. *Early Childhood Research Quarterly*, *17*(338–355), 596–601.
Nunnally, J. C. (1978). Psychometric theory (2nd ed.). .. New York, NY: McGraw-Hill.
Oades-Sese, G. V., Kaliski, P. K., & Weiss, K. (2010). Factor structure of the Devereux Early Childhood Assessment clinical form in low-income Hispanic American bilingual preschool children. *Journal of Psychoeducational Assessment*, *28*, 357–372.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Pychometrika*, *44*, 443–460.
Petitclere, A., Boivin, M., Dionne, G., Zoccolillo, M., & Tremblay, R. E. (2008). Disregard for rules: The early development and predictors of a specific dimension of disruptive behavior disorders. *Journal of Child Psychology and Psychiatry*, *50*, 1477–1484.
Phillips, D. A., & Lowenstein, A. E. (2011). Early care, education, and child development. *Annual Review of Psychology*, *62*, 483–500.
Pianta, R. C. (1996). Student–Teacher Relationship Scale.. Charlottesville, VA: University of Virginia.
Pianta, R. C. (2001). Student–Teacher Relationship Scale professional manual.. Lutz, FL: Psychological Assessment Resources, Inc.
Pianta, R. C., Cox, M. J., & Snow, K. L. (Eds.). (2007). *School readiness and the transition to kindergarten in the era of accountability*.. Baltimore, MD: Paul H. Brookes Publishing Co.
Pianta, R. C., & Stuhlman, M. W. (2004). Teacher–child relationships and children's success in the first years of school. *School Psychology Review*, *33*, 444–458.
Pihlakoski, L., Sourander, A., Aromaa, M., Rautava, P., Helenius, H., & Silanpaa, M. (2004). The continuity of psychopathology from early childhood to adolescence: A prospective cohort study of 3-12-year-olds. *Journal of Child and Adolescent Psychiatry*, *15*, 400–417.
President's New Freedom Commission on Mental Health (2003). Achieving the priorities: Transforming mental health care in America.. Rockville, MD: U.S. Department of Health and Human Services.
Race to the Top Fund (2009). *Final Rule*, *34*. (pp. 59798): C.F.R.
Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Harder, V. S., Otten, L., Bilenberg, N., et al. (2011). International comparisons of behavioral and emotional problems in preschool children: Parents' reports from 24 societies. *Journal of Clinical Child & Adolescent Psychology*, *40*, 456–467.
Salvia, J., Ysseldyke, J., & Bolt, S. (2007). Assessment in special and inclusive education (11th ed.). .. Belmont, CA: Wadsworth.
Sameroff, A. J., & Fiese, B. H. (2000). Models of development and developmental risk. In C. H. ZeamahJr. (Ed.), *Handbook of infant mental health* (pp. 3–19) (2nd ed.). . New York, NY: Guilford.
SAS (2011). Statistical analysis system (version 9.3).. Cary, NC: SAS Institute, Inc [Software].
Sass, D. A., & Smith, P. L. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling*, *13*, 566–586.
Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.
Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 162–171.
Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*, 148–154.
Strickland, J., Keller, J., Lavigne, J. V., Gouze, K., Hopkins, J., & LeBailly, S. (2011). The structure of psychopathology in a community sample of preschoolers. *Journal of Abnormal Child Psychology*, *39*, 601–610.
Thissen, D., & Wainer, H. (2001). Test scoring.. Mahwah, NJ: Erlbaum.
Thompson, B., & Melancon, J. (November,1996). Using item “testlets/parcels” in confirmatory factor analysis: An example using the PPDP-78. *Paper presented at the annual meeting of the Mid-South Educational Research Association, Tuscaloosa, AL* (ERIC Document Reproduction Service No. ED 404 349).
U.S. Department of Health and Human Services (2001). *Head Start FACES: Longitudinal findings on program performance.* Third progress report.. Washington, DC: Administration for Children and Families.
U.S. Department of Health and Human Services (2003). Strengthening Head Start: What the evidence shows. Retrieved from: http://aspe.hhs.gov/hsp/strengthenheadstart03/
U.S. Department of Health and Human Services (2010a). Head Start Impact Study final report.. Washington DC: Administration for Children and Families.
U.S. Department of Health and Human Services (2010b). Head Start Impact Study technical report.. Washington DC: Administration for Children and Families.
Vaden-Kiernan, M., D'Elio, M. A., O'Brien, R. W., Tarullo, L. B., Zill, N., & Hubbell-McKey, R. (2010). Neighborhood as a developmental context: A multilevel analysis of neighborhood effects on Head Start families and children. *American Journal of Community Psychology*, *45*, 49–67.
Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

von Suchodoletz, A., Trommsdorff, G., Heikamp, T., Wieber, F., & Gollwitzer, P. M. (2009). Transition to school: The role of kindergarten children's behavior regulation. *Learning and Individual Differences*, *19*, 561–566.
Waller, N. G. (2001). MicroFACT 2.0: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems.. St. Paul, MN: Assessment Systems.
Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway? *Early Childhood Research Quarterly*, *27*, 46–54.
Wilkinson, W. W. (2007). The structure of the Lawrence locus of control scale in young adults: Comparing item and parcel indicator models. *Personality and Individual Differences*, *43*, 416–425.
Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2002). TESTFACT (ver. 4.0) [computer program].. Lincolnway, IL: Scientific Software International.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2002). Woodcock–Johnson III Tests of Achievement.. Itasca, IL: Riverside.
Yates, A. (1987). Multivariate exploratory data analysis: A prospective on exploratory factor analysis.. Albany, NY: State University of New York Press.
Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1999). BILOG-MG (ver. 3.0) [computer program].. Lincolnway, IL: Scientific Software International.
Ziv, Y., Alva, S., & Zill, N. (2010). Understanding Head Start children's problem behaviors in the context of arrest or incarceration of household members. *Early Childhood Research Quarterly*, *25*, 396–408.