Received: 6 December 2017 | Accepted: 23 May 2018

DOI: 10.1002/pits.22150

WILEY

RESEARCH ARTICLE

Initial assessment versus gradual change in early childhood behavior problems–Which better foretells the future?

Paul A. McDermott[1] | Michael J. Rovine[1] | Katharine W. Buek[1] | Roland S. Reyes[1] | Jessica L. Chao[2] | Marley W. Watkins[3]

[1]University of Pennsylvania

[2]American Speech-Language-Hearing Association

[3]Baylor University

Correspondence
Paul A. McDermott, University of Pennsylvania, Graduate School of Education, 3700 Walnut Street, Philadelphia, PA 19104-6216.
Email:drpaul4@verizon.net

This research was conducted with the cooperation of the U.S. Department of Health and Human Services, Administration for Children and Families.

Abstract

Leading research argues the distinct importance of earliest detection of childhood behavior problems and the value of discovering subsequent change patterns as children transition through the early education years. This study examined the relative contributions of earliest assessments of children's problem behaviors as compared to the changes in those behaviors over time for the prediction of important later outcomes. Focusing on the representative national sample from the Head Start Impact Study (n = 3,827), classroom behavior problems were assessed across 4 years spanning prekindergarten through first grade. Individual child indices were derived in multilevel growth modeling to reflect initial assessments and subsequent change patterns. These indices were thereafter applied in multilevel logistic regression and receiver operating characteristic curve analyses to predict later academic and social difficulties. Both children's initial assessments and their transitional changes proved to be good predictors of most outcomes, where the accuracy for initial assessments and transitional changes was effectively equivocal. The evidence clarifies that initial assessment of behavior problems is sufficient to predict later outcomes; additional assessments did not augment forecasting accuracy nor did the combination of both initial assessment and information about subsequent change improve accuracy. Implications are discussed for assessment theory and practice.

KEYWORDS

behavior problems, developmental transition, early childhood education, early onset, Head Start, multilevel generalized linear modeling, multilevel growth curve modeling, receiver operating characteristic curve

1 | INTRODUCTION

National prevalence rates for preschool emotional and behavioral problems now approach 20% (Dougherty et al., 2015), with early and untreated problems undermining critical developmental processes and portending more serious and sometimes intractable disorders at later ages (Campbell & James, 2007; Feeney-Kettler, Kratochwill, & Kettler, 2011; Kataoka, Zhang, & Wells, 2002). Fewer than one in five identified children actually receive intervention services and these services are usually delayed until the elementary school years (Feeney-Kettler et al., 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). Unattended emotional and behavioral disorders also tend to cascade forward and manifest in future academic deficits, school adjustment problems, special needs enrollment, retention, suspension, absenteeism, and dropout (Bornstein, Hahn, & Suwalsky, 2013; Buhs & Ladd, 2001; Darney, Reinke, Herman, Stormont, & Ialongo, 2013; Obradović, Burt, & Masten, 2009; Searle, Sawyer, Miller-Lewis, & Baghurst, 2014; Wright, Morgan, Coyne, Beaver, & Barnes, 2014).

This situation has motivated a variety of responses, at the center of which are large-scale assessment systems to early identify and differentiate manifestations of preschool socioemotional distress (Campbell & James, 2007; Feeney-Kettler et al., 2011; McDermott, Watkins, Rovine, & Rikoon, 2013, 2014; Rescorla et al., 2011), the intention being to clarify the distinct nature of problems in such a way that might lead to preventative or restorative intervention. The assessment systems characteristically embrace either of two related but distinctly different approaches–a static approach or a transitional approach. The former approach is the more common and features application of standardized child behavior rating scales focusing on a given setting (school or home) at a given point in time. The latter approach also concentrates on a given setting (school or home) but requires repeated assessments of child behavior, particularly as the child develops through the critical early childhood transition years, spanning preschool entry through arrival in formal schooling. Each approach offers some advantage; the static approach, if applied early enough in the child's development, enables prompt identification and differentiation of any problem behaviors and the transitional approach informs the degree and direction of change as time passes. For both approaches, a primary measure of criterion and consequential validity is whether they are able to predict what will ultimately transpire for a child in the future (academic success, school adjustment, home adjustment). One might intuitively argue for the use of both approaches in tandem, a thorough assessment at the earliest opportunity to be followed by a succession of later assessments across the transition years. But this thinking presumes that the combined approaches are better than one in isolation and obfuscates the fundamental question as to whether one approach is better than the other and quite sufficient.

1.1 | PROS and CONS

Mental health policy and research emphasize the importance of early assessment to detect childhood emotional and behavioral disturbances (Glascoe, 2000; Poulou, 2015). The prime rationale is that prompt detection enables prompt intervention (either preventive or corrective) to preempt progressive deterioration, evolution of secondary morbidity, functional interference with natural developmental processes, with schooling and socialization, and with the moral imperative to reduce or eliminate personal suffering (National Research Council and Institute of Medicine, 2009). Timely identification is thought to be cost-beneficial because it can forestall maladies at their most seminal stages and permit subsequent avoidance of risk factors (e.g., child neglect, bullying) and introduce protective factors (e.g., family guidance, special needs services). Moreover, some disturbances have reciprocal relationships between early age-of-onset and the rapidity of deterioration and the resistance to delayed intervention (Kessler et al., 2005).

On the other hand, not all children enter a day-care or prekindergarten environment where adults can properly assess the relative social adjustment of interacting children. Children from underresourced families, in particular, are more likely to attend schools of lower quality and to have teachers with less training (Currie & Thomas, 2000; Isenberg et al., 2016; Lee & Loeb, 1995; Zhai, Raver, & Jones, 2012). In these settings, it is less likely that the classroom environment will provide the optimal context for accurate assessment of the child's behavior, and that valid and reliable assessments will be administered because of a lack of resources, training, or both (McDermott et al., 2011,

2017). Additionally, low-income children are likely to enter prekindergarten programs later than other children, or not at all. For example, among children from the most underresourced households, some enter preschool when they are approximately 3 years old, but many more enter a year later, and some children are not seen in a school setting until kindergarten or first grade when attendance becomes compulsory (U.S. Department of Health and Human Services, 2010). Hence, the timing of earliest assessment will differ among children.

Also, evidence from an early assessment can often be misleading because most early childhood disturbances do not follow a straight pathway over time. Rather, the preponderance of early-detected self-control problems (e.g., physical aggression) and reticence/withdrawal problems will decrease markedly in frequency and intensity as children move through the early transition years, whereas other problems will characteristically intensify (Barker & Maughan, 2009; McDermott et al., 2017; Pine & Fox, 2015; Tremblay, 2010). Thus, results of early assessment cannot be presumed emblematic of things to come.

The transitional view toward assessment has become increasingly popular, especially as it pertains to children in their early years. Early childhood educators contend that the child's responses to the movement from Piagetian preoperational discovery learning to common structured curricula, from individual interests to group cooperation, and from nurturing acceptance to performance grading will influence substantively the child's attitudes toward and adaptations to long-term academic and interpersonal challenges (Gurin, Day, Hurtado, & Gurin, 2002; Heckman, 2006; Pianta, Cox, & Snow, 2007).

The transitional nature of the early childhood educational period is such that adaptation, like growth itself, can only be understood as a developmental progression. This requires a clear picture of the course of change in any given area of performance. To acquire understanding of the role of problem behaviors, it follows that practitioners would need to follow their developmental pathways as children move through the milestone transitions from preschool to formal schooling. This perspective reveals whether observed disturbances remain stable over time, whether they tend to normalize as children navigate through the socializing influences of common schooling, whether they are likely to worsen with exit of preschool and entry into the more competitive setting of formal schooling, and when and where observed problems will generalize to other contexts as time progresses (Barker & Maughan, 2009; Bub, McCartney, & Willett, 2007; Cote, Vaillancourt, Barker, Nagin, & Tremblay, 2007; McDermott et al., 2013, 2014, 2017; Morgan, Farkas, & Wu, 2009; Trembley, 2010). However, organizing and maintaining a longitudinal assessment program is logistically complex and the requirements for well-designed longitudinal measures of behavior problems are psychometrically demanding. One cannot simply repetitively apply a given rating scale across contiguous developmental periods without doing the understandably costly groundwork to ensure construct continuity and linked vertical scaling. Additionally, this approach demands the opportunity and patience to wait long enough after first detection of a problem to uncover the developmental course that a child has embarked upon, a delay that may preclude timely intervention.

1.2 | Methodological approach

Our research is designed to contrast the consequential and predictive validity of earliest assessment versus transitional change approaches. To accomplish this, we demonstrate the relative predictive accuracy of each method for distal outcomes measured at the close of first grade (direct assessments of academic achievement, parents' assessments of home behavioral adjustment, and teachers' assessments of academic proficiency and school social adjustment). We regress each of these outcomes on children's earliest assessments for behavior disturbance and on indicators of the degree and direction of change in those behavior problems across the transition years. Drawn from individual growth models, child-level parameters will serve as the proxy for children's initial level of performance (random intercepts from first assessment) and child-level change parameters (random slopes) as the indicator of transitional change. To maximize the generalizability of results, we apply a nationally representative sample of children from the Head Start Impact Study ([HSIS]; U.S. Department of Health and Human Services, 2010). Because respondent teachers tended to observe more than one child per classroom, the data obtained on children's problem behaviors were nested within classrooms reflecting both classroom contextual and teacher observational effects, thereby necessitating the estimation of multilevel growth-curve models. Additionally, the information pertaining to distal outcomes at the end of first

grade was similarly nested and required appropriate multilevel estimation of the relative contributions of children's initial assessments versus their transitional changes.

1.3 | Research questions

This research seeks to answer two questions examining each of four nationally standardized measures of early childhood behavior problems (aggression, attention seeking, reticence/withdrawal, low energy behavior) assessed over 2 years of prekindergarten and 1 year each of kindergarten and first grade. First, are there appropriate multilevel growth models yielding statistically significant random effects variance components, reflecting variability in the individual children's relative performance at initial assessment and their change across the transition years? Second, what is the relative contribution and forecasting accuracy of children's individual initial assessment and transitional change?

2 | METHOD

2.1 | Participants

HSIS was a nationwide randomized investigation designed to estimate the relative effectiveness of Head Start. Sampling for HSIS was conducted using a clustered, multistage stratified design wherein 84 Head Start grantees were randomly selected from the Northeast, North Central, South, Plains, and West regions of the country. From these grantees, 383 Head Start centers were randomly selected. From among the pool of families applying for enrollment in the selected centers that were eligible for Head Start enrollment (defined by federal income criteria), children were randomly permitted either to enroll in Head Start or a non–Head Start program. Classroom teachers assessed each child's behavioral adjustment toward the end of the first and second years of prekindergarten (PreK 1 and PreK 2, respectively), kindergarten (K), and first grade (1st grade).

Because not all children selected for prekindergarten enrollment entered school for PreK 1, and because some enrolled in noneducational settings or were not enrolled until kindergarten or later, the sample size increased as children moved from PreK 1 to 1st grade (i.e., PreK 1, $n = 1{,}377$; PreK 2, $n = 2{,}764$; K, $n = 2{,}873$; 1st grade, $n = 3{,}077$). As reported below (see Sensitivity Analyses section), sensitivity analyses assessing the effects of missing data supported the assumption that patterns of missing data were essentially random and unrelated to levels or changes in the focal longitudinal variables.

Since there was no significant effect of participation in Head Start on measures of problem behavior or 1st-grade outcomes (U.S. Department of Health and Human Services, 2010), we employ the full study sample from both Head Start and non–Head Start conditions in this analysis. The full national sample assessed in this study contained 3,827 children, where mean age at study entry was 4.0 years ($SD = .5$), with 49.6% females, 37.8% Hispanic, 29.5% African-American, 32.7% White or other race/ethnicity, 25.7% primarily Spanish-speaking, 12.8% identified with special needs, and 82.7% residing in urban areas. During PreK 1, children attended 540 preschool centers (867 classrooms) and during PreK 2 1,032 centers (1,815 classrooms). During kindergarten, children attended 1,469 schools (2,280 classrooms) and during first grade 1,617 schools (2,576 classrooms). During PreK, approximately 80% of classrooms were not affiliated with conventional schools (~60% being day care or other nonschool centers), with about 90% of post-PreK classrooms affiliated with public schools.

2.2 | Longitudinal measures

The HSIS employed the Adjustment Scales for Early Transition in Schooling ([ASETS]; McDermott et al., 2013), a standardized teacher rating scale, for longitudinal assessment of classroom behavior problems over the four years of the study. The ASETS is comprised of four scales of problem behaviors: Aggression, Attention Seeking, Reticence/Withdrawal, and Low Energy. The aggression scale comprises 32 items related to physical aggression, bullying, disruption, cheating, theft, and domineering behavior, and includes items such as “Physically aggressive in peer

conflicts" and "Starts fights and rough play during games." The Attention-Seeking scale comprises 12 items relating to clinging, feigned helplessness, loudness, and crying, including such behaviors as "Insists on sitting next to teacher" and "Tells on others to gain teacher's favor." The Reticence/Withdrawal scale describes behaviors that characterize timidity, disengagement, and asociality. It includes 24 items including "Freezes up and doesn't answer questions" and "Ignores all other children." Finally, the Low Energy scale contains 12 items that relate to disinterest, lethargy, and lack of motivation. Example items include "Cannot work up energy to face anything new" and "Too lethargic to ask for help." Items are dichotomous, indicating the presence or absence of a given behavior over the past month, and are presented in classroom situations (involving teacher, age-mates, learning activities, organized play, group activities) to provide context around observed problem behaviors. Additionally, to diminish negative response sets, at least one item in each situation describes normal or commonplace behavior. It should be noted that Aggression and Attention Seeking are phenotypically regarded as externalizing forms of problem behavior and Reticence/Withdrawal and Low Energy are regarded as internalizing forms of behavior.

Each ASETS scale is based on longitudinal exploratory and confirmatory structural analyses across the four academic years, with item response theory (IRT) calibration under the two-parameter logistic model, vertical equating using nonbiased linking items, and scaled scores (*SS*s) via expected a posteriori (EAP) Bayesian estimation where the population *SS* $M = 50$ and $SD = 10$ at PreK 1, the reference year. Internal consistency as derived directly from the IRT EAP scores and their standard errors was .96 for Aggression, .87 for Attention Seeking, .92 for Reticence/Withdrawal, and .77 for Low Energy. Substantial evidence for concurrent and predictive criterion validity, as well as sensitivity to linear and higher-order growth detection is described in McDermott et al. (2013).

2.3 | Distal outcome measures

ASETS problem behavior scores were used to predict scores on direct assessments of achievement, parent ratings of home adjustment, and teacher ratings of classroom adjustment and achievement at the close of first grade.

2.3.1 | Direct assessments

The Basic Reading Skills cluster (letter and word reading and writing, phonemic and structural analysis) and Mathematics Reasoning cluster (quantitative concepts, counting, problem solving) of the Woodcock–Johnson III Tests of Achievement ([WJ]; Woodcock, McGrew, & Mather, 2002) were administered. HSIS first-grade population internal consistency for Basic Reading Skills was .91 and Mathematics Reasoning .78 (U.S. Department of Health and Human Services, 2010). Ample validity support has been reported for the two WJ achievement clusters (Dumont & Willis, 2006; McGrew, Woodcock, & Schrank, 2007; Salvia, Ysseldyke, & Witmer, 2017). Additionally, the Peabody Picture Vocabulary Test (Dunn, Dunn, & Dunn, 1997) was used to assess listening comprehension (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 (U.S. Department of Health and Human Services, 2010). Criterion validity evidence is abundant (e.g., Dumont & Willis, 2006; Salvia, Ysseldyke, & Bolt, 2007) and internal consistency was .78 for the HSIS population.

2.3.2 | 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 contains 14 dichotomous items, such as "Is disobedient at home," "Can't concentrate, can't pay attention for long," and "Feels worthless or inferior." Development and validity evidence are provided for the FACES national study in U.S. Department of Health and Human Services (2001, p. 2.27) and for HSIS in the U.S. Department of Health and Human Services (2010). Additional validity evidence has been reported by other researchers (e.g., Vaden-Kiernan et al., 2010; Ziv, Alva, & Zill, 2010). For the HSIS first-grade sample, internal consistency ranged from .78 to .79.

2.3.3 | Teacher ratings

The Pianta Student–Teacher Relationships Scale (Pianta, 1996) features 15 items, such as, "I share an affectionate, warm relationship with this child" and "This child easily becomes angry at me," rated on a 5-point scale ranging from 1 = *definitely does not apply* to 5 = *definitely applies*. The items are divided into two subscales: Closeness (seven items) and Conflict (eight items). Concurrent and predictive validity evidence for the scale is provided in Pianta (2001) and Pianta and Stuhlman (2004) and internal consistency for the relevant HSIS population = .82 for Closeness and .89 for Conflict (U.S. Department of Health and Human Services, 2010). Teacher report of Academic Ability was rated at the end of the first grade for Language and Literacy, Mathematics, and Social Science, based on attainment of multiple skills compared to the attainment of peers (U.S. Department of Health and Human Services, 2010). Initially rated on a 5-point scale from 1 = *far below average* to 5 = *proficient*, the data were subsequently reduced by the Federal Government to a simple binary scale (1–2 vs. 3–5) to enhance parsimony and reliability. Because each measure is essentially a single index, internal consistency estimates are infeasible. Thus, the appropriate standard error of the *M* is reported here, where (assuming binary scores = 0 vs. 1) SE_M for Language and Literacy = .008, Mathematics = .008, and Social Science = .007.

2.4 | Procedure

The research questions respectively motivated the applications of multilevel individual growth-curve modeling to estimate each child's intercept and slope parameters for scores in Aggression, Attention Seeking, Reticence/Withdrawal, and Low Energy behavior across the four transition years, and multilevel generalized linear regression to regress the various distal outcomes on those individual child parameters. The method corresponds to the two-step strategy demonstrated by Horrocks and van Den Heuvel (2009), Maruyama, Takahashi, and Takeuchi (2009), and Wang, Wang, and Wang (2000), except that the second step applies a multilevel procedure to accommodate the nested variance of distal outcomes.

In the first step, which addressed the first research question, we estimated the appropriate growth model for each of the four types of longitudinal problem behavior, where models incorporated statistically significant linear, quadratic, and cubic fixed effects as per Type 1 sequential *F* tests and significant random intercepts at the classroom level and random intercepts and slopes at the child level. General model specification was $\hat{Y}_{ijk} = \gamma_{000} + \gamma_{100}\text{Time}_{ijk} + \gamma_{200}\text{Time}^2{}_{ijk} + \gamma_{300}\text{Time}^3{}_{ijk} + (\mu_{00k} + \mu_{10k}\text{Time}_{ijk}) + (\mu_{0jk} + \mu_{1jk}\text{Time}_{ijk}) + r_{ijk}$. Note that the procedure simultaneously estimated both child-level random intercepts and slopes coefficients such that their natural reduction in the restricted maximum-likelihood process would be relative and unbiased (Sayers et al., 2014). Resultant child-level intercept and slope values were extracted for use as predictors in step 2 analyses. As a measure to avoid confusion in step 2 applications, each child's intercept value is referred to as the child's level (estimated level of problem behavior at first assessment, i.e., PreK 1) and slope value is referred to as the child's change (estimated degree and direction of change over the transition years, i.e., PreK 1 to 1st grade).

For these models, we used full information maximum likelihood (FIML) estimation based on the raw data likelihood. This approach models all available data by constructing the likelihood based on raw data values rather than summary statistics (means, covariances, etc.). Here, missing data are not imputed, they are part of the model. Researchers have shown that FIML estimates are equivalent to empirical Bayes estimates (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Other simulation studies have shown that multiple imputation estimates approach FIML estimates as the number of imputations increases (Collins, Schafer, & Kam, 2001). The assumptions for all three of the approaches mentioned are identical (Allison, 2012; Little & Rubin, 2002).

For the second step, which addressed the second research question, binary distal outcomes were generated and regressed on children's individual level and change values for problem behavior. Binary outcomes were deemed appropriate because (1) the alternative WJ normal-curve equivalent scores and parent and teacher ratings were significantly abnormally distributed and differentially skewed; (2) the WJ item-domain representation was relatively sparse below the 25th percentile with punctuated rather than graduated changes in item difficulty (a problem common to

commercial tests; see McDermott et al., 2009); (3) teachers' assessments of child performance in Language and Literacy, Mathematics, and Social Science had already been bifurcated by the federal government for psychometric efficiency; and (4) dichotomously scaled outcomes would yield relative probabilities of desirable versus undesirable outcomes in first grade as related to children's initial level and gradual change in problem behavior. Thus, an outcome reflecting Reading Proficiency versus Nonproficiency was formed from WJ Basic Reading Skills scores where Proficiency comprised performance in the upper four quintiles (scored 0) and Nonproficiency the lowest quintile (scored 1). Similarly, a Mathematics Nonproficiency variable was formed from WJ Mathematics Reasoning (lowest quintile = 1) as were variables formed from the Peabody Picture Vocabulary Test, Pianta Closeness with Teacher, and teacher assessments of Language and Literacy, Mathematics, and Social Science (lower quintile = 1). Conversely, a parent Home Problem Behaviors indicator and Pianta Conflict with Teacher indicator were formed from the respective scales (upper quintile = 1). Quintiles were preferred because they provided the necessary statistical power for reliable point separation in multilevel generalized (logistic) modeling (Stokes, Davis, & Koch, 2001).

Using children's problem behavior level and change values as explanatory predictors, step 2 applied multilevel generalized linear modeling, with adaptive quadratures, the Bernoulli response distribution, and logit link function to estimate the likelihood of each undesirable outcome (reading nonproficiency, conflict with teacher, etc.). Model specification was $\text{Nonproficiency}_{\Theta[\text{logit}]ij}=\gamma_{00}+\gamma_{10}\text{Level}_i+\gamma_{20}\text{Change}_i+\mu_{0j.}$, where i indexes children and j represents teachers/classrooms. The relative risk increment or reduction for better versus poorer outcomes associated with children's initial problem behavior level and gradual change was estimated through the odds ratio. Finally, the accuracy of predictions was summarized through receiver operating characteristic (ROC) curve models (Swets, Dawes, & Monahan, 2000) based on the response probabilities for undesirable outcomes as adjusted for classroom nesting through the multilevel generalized linear models.

3 | RESULTS

3.1 | Sensitivity analyses

Given that multilevel individual growth-curve models estimate values for missing data, all analyses were repeated as based solely on those children who were present at PreK 1. These sensitivity analyses produced results for both the step 1 multilevel growth models and step 2 multilevel logistic models and ROC analyses that were uniformly consistent with results based on the full data set in terms of significant fixed and random effects and relative magnitude of accuracy under ROC curve analyses. This supported the assumption that the children in the full sample were observed at random with some data missing at random and unrelated to levels of or changes in the longitudinal variables (Little & Rubin, 2002; Marini, Olsen, & Rubin, 1979).

3.2 | Growth models

Table 1 presents parameters for the growth models pertaining to each form of longitudinal behavior problems. A quadratic model was preferable for Aggression, whereas the other forms of problem behavior required cubic models. All of the models featured negative linear change. Only linear random slopes could be estimated for Aggression and Attention Seeking, whereas both linear and quadratic random slopes could be estimated for Reticence/Withdrawal and Low Energy behaviors. Recall that Aggression and Attention Seeking represent externalizing behaviors while Reticence/Withdrawal and Low Energy are considered internalizing forms of behavior. Thus, the evidence shows that change patterns vary linearly among children with respect to externalizing problems but vary among children both linearly and quadratically for internalizing problems.

Considering the variance decomposition, 55.0% of the Aggression and 53.0% of Low Energy variation is between children, while 43.0% is between children for Attention Seeking and 65.0% for Reticence/Withdrawal. This is the focal variance for this study. Of this variance, more than 90% is attributable to differences on levels

TABLE 1 Properties of multilevel individual growth-curve models for early childhood behavior problems across 4 years

Effect	Aggression	Attention Seeking	Reticence/ Withdrawal	Low Energy
Fixed effects parameter estimates (and *SE*)[a]				
Intercept	49.81 (.21)	49.83 (.20)	50.44 (.21)	50.06 (.18)
Linear slope	−1.35 (.24)	−.13 (.54)	−5.23 (.55)	−2.93 (.50)
Quadratic slope	.22 (.07)	−.78 (.44)	3.39 (.45)	2.77 (.42)
Cubic slope		.23 (.09)	−.63 (.10)	−.54 (.09)
Random effects parameter estimates (and *SE*)[b]				
Variance/covariance				
Classroom intercepts	7.76 (2.75)	9.47 (2.79)	18.31 (2.68)	7.51 (2.40)
Child intercepts	45.13 (2.64)	21.92 (2.09)	27.38 (2.96)	10.38 (2.02)
Child linear slopes	2.58 (.44)	2.03 (.45)	17.15 (5.91)	18.69 (5.00)
Child quadratic slopes			1.15 (.56)	1.79 (.48)
Child intercepts by linear slopes	−4.13 (.91)	−2.15 (.84)	−10.46 (3.28)	−8.86 (2.64)
Child intercepts by quadratic slopes			2.15 (.94)	2.61 (.76)
Child linear by quadratic slopes			−4.22 (1.77)	−5.29 (1.51)
Residual	31.15 (2.66)	31.75 (2.70)	24.57 (2.70)	27.48 (2.48)
Random effects variance decomposition[c]				
% Variance				
Between classrooms	9.0	15.0	26.0	13.0
Between children	55.0	43.0	65.0	53.0
Level (viz., intercept)	94.6	91.5	59.9	33.6
Change (viz., slope)	5.4	8.5	40.1	66.4
Within children	36.0	42.0	39.0	68.0

Note: $n = 3{,}827$.

[a]Values are based on Type I (sequential) *F* tests, with statistical significance at $p < .03$. Unreported higher-order effects indicate statistical nonsignificance and exclusion from a model.

[b]Values are based on restricted maximum-likelihood estimation, with statistical significance at $p < .03$.

[c]For a given area of behavior problems (column), the values for between classrooms, between children, and within children variance sum to 100.0% of all estimated variance. Bold entries pertain to the focal variance components for the overall study. For a given area of behavior problems (column), the values for level and change sum to 100.0% of the estimated variance between children.

(initial assessments) and less than 10% to changes (slopes) for Aggression and Attention Seeking (the two externalizing types). Alternatively, much less variation in levels (59.9% and 33.6%, respectively) and more variation in changes (40.1% and 66.4%, respectively) are associated with Reticence/Withdrawal and Low Energy behavior (the two internalizing types). Thus, the change patterns for internalizing problems are more complex and more heterogeneous in terms of differences between children. (Note that level and change variance constitute the effects of interest in this study and are shown in bold in Table 1.)

3.3 | Level versus change

The level (indicating a child's problem behavior level at first assessment) and change values (indicating degree and direction of change over time), as derived from the multilevel growth models, were applied as simultaneous predictors in multilevel logistic regression to estimate outcomes at the end of the first grade. For each distal outcome, Table 2 reports the corresponding odds ratio and risk increment (%) for a child's inclusion in the least desirable quintile for a

TABLE 2 Multilevel adjusted risk odds for negative distal outcomes as associated with initial level and emergent change in early childhood behavior problems

	Aggression		Attention Seeking		Reticence/Withdrawal		Low Energy	
	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment
First-grade reading nonproficiency (direct assessment, ICC = .57, *n* = 2,873)								
Level	1.13 (1.09/1.17)	12.9	1.10 (1.05/1.16)	10.1	1.22 (1.13/1.31)	21.7	1.78 (1.40/2.24)	77.5
Linear change	1.09 (.88/1.34)†		1.18 (.91/1.53)†		2.26 (1.43/3.57)	126.0	1.75 (1.41/2.17)	75.0
Quadratic change					9.67 (1.51/61.86)	867.3	2.29 (1.21/4.34)	129.2
First-grade mathematics nonproficiency (direct assessment, ICC = .51, *n* = 2,879)								
Level	1.10 (1.07/1.13)	9.6	1.06 (1.01/1.11)	5.9	1.24 (1.17/1.31)	24.0	1.74 (1.47/2.06)	74.3
Linear change	1.07 (.87/1.31)†		1.32 (1.03/1.69)†		2.11 (1.47/3.02)	110.9	1.64 (1.40/1.91)	63.6
Quadratic change					6.49 (1.45/28.99)	549.2	1.77 (1.07/2.93)	76.6
First-grade receptive vocabulary nonproficiency (direct assessment, ICC = .40, *n* = 2,883)								
Level	1.00 (.98/1.03)†		.96 (.92/1.00)†		1.13 (1.08/1.17)	12.7	1.26 (1.10/1.44)	25.8
Linear change	.89 (.74/1.08)†		1.11 (.88/1.40)†		1.05 (.78/1.41)†		1.10 (.96/1.25)†	
Quadratic change					.74 (.20/2.73)†		1.21 (.76/1.93)†	
First-grade home behavior problems (parent rating, ICC = .19, *n* = 2,900)								
Level	1.11 (1.09/1.14)	11.4	1.10 (1.06/1.14)	9.6	1.08 (1.05/1.12)	8.3	1.32 (1.19/1.47)	32.1
Linear change	1.30 (1.11/1.53)	30.2	1.39 (1.16/1.68)	39.3	1.45 (1.12/1.86)	44.7	1.24 (1.11/1.37)	23.7
Quadratic change					4.34 (1.43/13.13)	334.0	1.52 (1.05/2.21)	51.9
First-grade conflict with teacher (teacher rating, ICC = .19, *n* = 3,050)								
Level	1.45 (1.35/1.56)	45.1	1.35 (1.25/1.39)	32.0	1.03 (1.00/1.06)†		1.06 (.94/1.19)†	
Linear change	10.41 (6.61/16.39)	940.7	4.26 (3.19/5.69)	326.0	2.40 (1.83/3.14)	139.7	1.82 (1.60/2.06)	81.8
Quadratic change					24.29 (7.65/77.14)	>999	6.97 (4.49/10.84)	597.2
First-grade lack of closeness with teacher (teacher rating, ICC = .29, *n* = 3,053)								
Level	1.09 (1.06/1.11)	8.6	1.00 (.96/1.04)†		1.32 (1.24/1.40)	32.0	1.21 (1.07/1.35)	21.3
Linear change	1.65 (1.38/1.96)	64.6	.93 (.76/1.14)†		7.43 (4.80/11.52)	643.0	1.60 (1.40/1.83)	60.2
Quadratic change					>99 (494.47/> 999)	>999	6.17 (3.75/10.16)	517.0

(Continues)

TABLE 2 (Continued)

	Aggression		Attention Seeking		Reticence/Withdrawal		Low Energy	
	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment	Odds Ratio (95% CLs)	% Risk Increment
First-grade language and literacy nonproficiency (teacher rating, ICC = .15, *n* = 3,042)								
Level	1.08 (1.06/1.10)	8.1	1.04 (1.01/1.07)†		1.19 (1.15/1.23)	18.6	1.47 (1.31/1.65)	47.0
Linear change	1.43 (1.23/1.65)	42.5	1.44 (1.21/1.70)	43.6	3.24 (2.49/4.23)	224.7	1.94 (1.70/2.21)	93.5
Quadratic change					68.46 (22.68/ > 99)	>999	5.84 (3.81/8.97)	484.3
First-grade mathematics nonproficiency (teacher assessment, ICC = .20, *n* = 3,029)								
Level	1.10 (1.07/1.12)	9.6	1.06 (1.02/1.09)	5.7	1.21 (1.16/1.26)	20.8	1.46 (1.27/1.68)	45.8
Linear change	1.39 (1.19/1.63)	39.2	1.37 (1.14/1.65)	37.0	3.42 (2.53/4.61)	214.6	2.13 (1.79/2.54)	113.2
Quadratic change					79.65 (23.19/ > 99)	>999	7.94 (4.53/13.93)	694.1
First-grade social studies and science nonproficiency (teacher assessment, ICC = .26, *n* = 3,020)								
Level	1.12 (1.09/1.15)	11.7	1.05 (1.01/1.09)†		1.22 (1.17/1.28)	22.3	1.51 (1.28/1.78)	51.0
Linear change	1.51 (1.25/1.82)	50.8	1.66 (1.33/2.08)	66.2	4.68 (3.25/6.73)	367.5	2.09 (1.73/2.52)	108.5
Quadratic change					>99 (65.60/ > 999)	>999	9.00 (4.84/16.74)	799.9

Note: Odds ratios and confidence limits (CLs) are derived from parameter estimates obtained through multilevel generalized linear modeling applying adaptive quadratures, the Bernoulli response distribution, and logit link function. Percentage risk increment = odds ratio − 1 (100). All values are statistically significant at $p < .03$ unless indicated † for nonsignificance. ICC (intraclass correlation coefficient) = proportion of between-classrooms variance estimated for a given distal outcome variable at the end of the first grade.

given outcome. The effects of child level and change are controlled for one another. As shown, in predicting first-grade reading nonproficiency from children's level and change in Aggression, only the odds ratio for level is statistically significant and shows a 12.9% increment in the risk of nonproficiency for each additional point in Aggression at first assessment. Changes in Aggression after the first assessment make no apparent difference. Alternatively, viewing the results for use of level and change in Reticence/Withdrawal in predicting reading nonproficiency, both level and change are significant. Specifically, as a child's initial level of Reticence/Withdrawal increases by 1 point, there is a 21.7% increase in risk for nonproficiency while as the linear change increases by 1 point, there is a 126.0% increase in the risk of nonproficiency. As a child's quadratic change increases by 1 point, there is an 867.3% increase in that risk.

For nearly every type of distal outcome (direct assessments of achievement, teacher assessments of achievement, parent ratings of home adjustment, teacher ratings of classroom adjustment), children's initial levels, and subsequent change in problem behavior make a difference in risk increment. Yet, a conspicuous absence of effect is found for transitional changes (slopes) in Aggression or Attention Seeking (both phenotypically considered externalizing forms of problem behavior) in predicting any type of direct assessment (reading, mathematics, receptive vocabulary), and there is a general resistance of receptive vocabulary to prediction, except when using children's initial assessments in Reticence/Withdrawal and Low Energy behaviors (both internalizing behavior problems).

Comparative performance of levels and changes is best evaluated through ROC analyses, where the area under the curve (AUC) is estimated based on the relative ability of levels versus changes to accurately predict an outcome. (An AUC of 1.00 would indicate perfect predictive accuracy and .50 mere chance accuracy, where values ≥.90 indicate high accuracy.) Based on those models where level effects were statistically significant (29 models), the AUC $M = .958$ ($SD = .038$; range = .874–.997) and change effects (27 models) AUC $M = .950$ ($SD = .039$; range = .858–.997). The differences between level and change forecasting accuracy are inconsequential. Joint forecasting accuracy (27 models) AUC $M = .947$ ($SD = .037$; range = .867–.998).

4 | DISCUSSION

This research sought to determine the relative contributions of earliest assessments of children's problem behaviors as compared to change in those behaviors for the prediction of later outcomes. For a nationally representative sample of children from low-income families, individual child indices were derived in multilevel growth modeling to reflect initial assessments and change patterns. These indices of initial level of problem behavior and change in behavior over the early education transition years were thereafter applied in multilevel logistic regression and ROC analyses to predict later academic and social difficulties. Both children's initial assessments and their transitional changes proved to be good predictors of most outcomes, where the accuracy for initial assessments and transitional changes was effectively equivocal.

The results are noteworthy because they indicate, at least as pertains to the kinds of childhood behavior disturbances covered in this national study, that it does not matter whether prognostications are based on first discovery of a problem or on protracted reassessments of that problem. The resultant predictions will be quite accurate. As such, interventions could be grounded on earliest detection, thus avoiding delay of treatment in anticipation of later assessment results. To the extent that effective interventions are available, earliest detection and intervention could minimize children's distress and associated academic and socioemotional sequelae and maximize the cost benefits for schools and communities (Bornstein et al., 2013; Campbell & James, 2007; Rescorla et al., 2011).

4.1 | Limitations

It is important to emphasize that this research presumes a certain population homogeneity and relative constancy for any given form of problem behavior. To illustrate, it has been shown by Cote et al. (2007) that, within the general population of children assessed for physical aggression, and by McDermott et al. (2017) for reticence/withdrawal, there exist three distinct latent subpopulations, each with a characteristically different problem level at earliest assessment

and change pattern over time. This demonstrates that there can be distinctive heterogeneity in a given population. However, for both the Cote et al. (2007) and McDermott et al. (2017) examples, one subpopulation is relatively more adjusted across time and one subpopulation is relatively more maladjusted, such that relative constancy is maintained. In such instances, information from the earliest assessment is likely to serve adequately for predicting subsequent outcomes because the different initial problem levels will differentially predict. In these studies, even though the child population is heterogeneous with respect to level and change in problem behavior, the problem level at first assessment serves as a good signal that the child manifests a certain intensity of problem behavior early on that, irrespective of some change over time, will identify that child as a relatively adjusted or maladjusted case.

However, it is also possible that relative constancy may not accompany population heterogeneity. Barker and Maughan (2009) have found for early childhood hyperactivity and peer problems and Feng, Shaw, and Silk (2008) for anxiety that, whereas latent subpopulations of initial level and gradual change exist (thus heterogeneity), the typical change patterns for some subpopulations are so dramatic as to cross over (intersect) the patterns of other subpopulations over time. In such cases, a child's initial assessment would not serve well as a predictor of distal outcomes. Here it is necessary to reassess children over time and allow change to inform any prediction. The discovery of population homogeneity (no distinctive subpopulations) versus heterogeneity (two or more distinct subpopulations) and relative constancy versus inconstancy in change can be resolved through application of latent growth mixture modeling. Exemplary software is provided with Mplus (Version 7.3; Muthen & Muthen, 2015) and PROC TRAJ (Jones, Nagin, & Roeder, 2001).

There are also limitations related to the generalizability of this research. First, in terms of generalizability to populations of children, it should be noted that because the HSIS focused on a low-income, Head Start–eligible population, these results pertain specifically to this demographic and cannot be generalized to other socioeconomic populations. Second, with regard to generalizability to other behavioral constructs, the HSIS and this analysis were primarily focused on problem behaviors that can be directly observed in a classroom context. As such, results do not necessarily generalize to all forms of child behavioral disturbance. As is necessary with all large-scale studies attempting to assess a number of performance areas, strategic decisions are made regarding which areas to measure.

Because of time and resource constraints, most of the measures used in HSIS are relatively brief, leaving open the question as to whether the constructs of interest are fully covered by the measures chosen. For example, we have not examined a longitudinal measure of parent-rated behavior in this study. Whereas it is often desirable to consider both parent and teacher measures of behavior problems in prediction of later outcomes, HSIS was designed around Head Start's legislative mandate, which prioritizes the development of learning-related and classroom-specific behaviors. As such, the HSIS sampling strategy was designed to produce a nationally representative sample of schools and classrooms, and not necessarily of parents or families. Thus, we determined that it was most appropriate to focus on teacher-rated measures of child behavior, while at the same time using parent ratings of first-grade behavior as a distal outcome. While recognizing that the measures used in these analyses are not able to capture all of the relevant facets of a construct, we have utilized the most rigorously validated and reliable measures of classroom behavior available and examined outcomes of various types with different raters and methods.

Finally, we note that while the results of this study indicate that both initial assessment and change over time are very good predictors of first-grade outcomes, this does not imply that these predictors explain all of the variance in outcomes. However, the ROC results provide some confidence in the quality of the models used here.

5 | CONCLUSION

The definition of a developmental trajectory naturally encompasses both the concept of a starting point (initial assessment) and a gradual change (growth rate). This research examined the relative utility of each concept for prediction of important future events that would inform the timing of intervention for early childhood behavior problems. The investigation clarifies that, for the types of problems studied, it suffices to know a child's initial level of behavioral

adjustment. Additional assessments did not augment forecasting accuracy nor did the combination of both initial assessment and information about subsequent change improve accuracy. Rather than undermining the importance of developmental transition studies to understand the nature of change and the contextual circumstances that surround change, the evidence highlights the crucial importance of problem detection at the earliest opportunity. It points to timely detection as a reliable and cost-effective alternative to delayed intervention, multiple reassessments, and the prospect of more serious and more generalized problem development. Repeated assessment, however, may be indicated when the goal is to uncover the natural developmental course of problem behaviors or to evaluate response to interventions.

ORCID

Katharine W. Buek http://orcid.org/0000-0002-2671-5536
Roland S. Reyes http://orcid.org/0000-0002-9595-8283
Jessica L. Chao http://orcid.org/0000-0002-7187-7227

REFERENCES

Allison, P. (2012). *Handling missing data by maximum-likelihood*. SAS Global Forum. Cary, NC: SAS Institute.

Barker, E. D., & Maughan, B. (2009). Differentiating early-onset persistent versus childhood—Limited conduct problem youth. *American Journal of Psychiatry*, *166*, 900–908.

Bornstein, M. H., Hahn, C.-S., & Suwalsky, J. T. D. (2013). Developmental pathways among adaptive functioning and externalizing and internalizing behavioral problems: Cascades from childhood into adolescence. *Applied Developmental Science*, *17*, 76–87.

Bub, K. L., McCartney, K., & Willett, J. B. (2007). Behavior problem trajectories and first-grade cognitive ability and achievement skills: A latent growth curve analysis. *Journal of Educational Psychology*, *99*, 653–670.

Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology*, *37*, 550–560.

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* (4th ed.). 111–135.). Mahwah, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Cote, S. M., Vaillancourt, T., Barker, E. D., Nagin, D., & Tremblay, R. E. (2007). The joint development of physical and indirect aggression: Predictors of continuity and change during childhood. *Development and Psychopathology*, *19*, 37–55.

Currie, J., & Thomas, D. (2000). School quality and longer-term effects of Head Start. *The Journal of Human Resources*, *35*, 755–744.

Darney, D., Reinke, W. M., Herman, K. C., Stormont, M., & Ialongo, N. S. (2013). Children with co-occurring academic and behavior problems in first grade: Distal outcomes in twelfth grade. *Journal of School Psychology*, *51*, 117–128.

Dougherty, L. R., Leppert, K. A., Merwin, S. M., Smith, V. C., Bufferd, S. J., & Kushner, M. R. (2015). Advances and directions in preschool mental health research. *Child Development Perspectives*, *9*, 14–19.

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.

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.

Feng, X., Shaw, D. S., & Silk, J. S. (2008). Developmental trajectories of anxiety symptoms among boys across early and middle childhood. *Journal of Abnormal Psychology*, *117*, 32–47.

Glascoe, F. P. (2000). Early detection of developmental and behavioral problems. *Pediatrics in Review*, *21*, 272–280.

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.

Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902.

Horrocks, J., & van Den Heuvel, M. J. (2009). Prediction of pregnancy: A joint model for longitudinal and binary data. *Bayesian Analysis*, *4*, 523–538.

Isenberg, E., Max, J., Gleason, J., Johnson, M., Deutsch, J., Hansen, M., & Angelo, L. (2016). *Do low-income students have equal access to effective teachers? Evidence from 26 districts* (Report No. NCEE 2017–4007). Retrieved from https://ies.ed.gov/ncee/pubs/20174008/pdf/20174007.pdf

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research*, *29*, 374–393.

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.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*, 593–602.

Lee, V. E., & Loeb, S. (1995). Where do Head Start attendees end up? One reason why preschool effects fade-out. *Educational Evaluation and Policy Analysis*, *17*, 62–82.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models.* Cary, NC: SAS Institute.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Marini, M. M., Olsen, A. R., & Rubin, D. B. (1979). Maximum likelihood estimation in panel studies with missing data. In K. Schuessler (Ed.), *Sociological methodology 1980.* San Francisco, CA: Jossey Bass.

Maruyama, N., Takahashi, F., & Takeuchi, M. (2009). Prediction of an outcome using trajectories estimated from a linear mixed model. *Journal of Biopharmaceutical Statistics*, *19*, 779–790.

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

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

McDermott, P. A., Watkins, M. W., Rovine, M. J., & Rikoon, S. H. (2013). Assessing changes in socioemotional adjustment across the early school transitions–New national scales for children at risk. *Journal of School Psychology*, *51*, 97–115.

McDermott, P. A., Watkins, M. W., Rovine, M. J., & Rikoon, S. H. (2014). Informing context and change in young children's sociobehavioral development–The national Adjustment Scales for Early Transition in Schooling (ASETS). *Early Childhood Research Quarterly*, *29*, 255–267.

McDermott, P. A., Rovine, M. J., Watkins, M. W., Chao, J. L., Irwin, C. W., & Reyes, R. (2017). Latent national subpopulations of early education classroom disengagement of children from underresourced families. *Journal of School Psychology*, *65*, 69–82.

McGrew, K. S., Woodcock, R. W., & Schrank, K. A. (2007). Woodcock–Johnson III normative update technical manual. Itasca, IL: Riverside.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Kindergarten predictors of recurring externalizing and internalizing psychopathology in the third and fifth grades. *Journal of Emotional and Behavioral Disorders*, *17*, 67–79.

Muthen, L. K., & Muthen, B. O. (2015). *Mplus (Version 7.3) [Computer software].* Los Angeles, CA: Authors.

National Research Council and Institute of Medicine. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities.* Washington, DC: The National Academies Press.

Obradović, J., Burt, K. B., & Masten, A. S. (2009). Testing a dual cascade model linking competence and symptoms over 20 years from childhood to adulthood. *Journal of Clinical Child & Adolescent Psychology*, *39*, 90–102.

Pianta, R. G. (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.

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. Retrieved from https://www.nasponline.org

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.

Pine, D. S., & Fox, N. A. (2015). Childhood antecedents and risk for adult mental disorders. *Annual Review of Psychology*, *66*, 459–485.

Poulou, M. S. (2015). Emotional and behavioural difficulties in preschool. *Journal of Child and Family Studies*, *24*, 225–236.

Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Harder, V. S., Otten, L., Bilenberg, N., … Verhulst, F. C. (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.

Salvia, J., Ysseldyke, J. E., & Witmer, S. (2017). *Assessment in special and inclusive education* (13th ed.). Boston, MA: Cengage Learning.

Sayers, A., Heron, J., Smith, A. D. A. C., Macdonald-Wallis, C., Gilthorpe, M. S., Steele, F., & Tilling, K. (2014). Joint modeling compared with two stage methods for analyzing longitudinal and prospective outcomes: A simulation study of childhood growth and BP. *Statistical Methods in Medical Research*, *26*(1)437–452.

Searle, A. K., Sawyer, M. C., Miller-Lewis, L. R., & Baghurst, P. A. (2014). Prospective associations between children's preschool emotional and behavioral problems and kindergarten classroom engagement, and the role of gender. *The Elementary School Journal*, *114*, 380–405.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2001). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–24.

Tremblay, R. B. (2010). Developmental origins of disruptive behaviour problems: The "original sin" hypothesis, epigenetics and their consequence for prevention. *Journal of Child Psychology and Psychiatry*, *51*, 341–367.

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. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/perform_3rd_rpt.pdf

U.S. Department of Health and Human Services. (2010). *Head Start impact study technical report*. Washington DC: Administration for Children and Families. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_tech_rpt.pdf

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.

Wang, C. Y., Wang, N., & Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, *56*, 487–495.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2002). Woodcock–Johnson III tests of achievement. Itasca, IL: Riverside.

Wright, J. P., Morgan, M. A., Coyne, M. A., Beaver, K. M., & Barnes, J. C. (2014). Prior problem behavior accounts for the racial gap in school suspensions. *Journal of Criminal Justice*, *42*, 257–266.

Zhai, F., Raver, C. C., & Jones, S. M. (2012). Academic performance of subsequent schools and impacts of early interventions: Evidence from a randomized controlled trial in Head Start settings. *Children and Youth Services Review*, *34*, 946–954.

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.

How to cite this article: McDermott PA, Rovine MJ, Buek KW, Reyes RS, Chao JL, Watkins MW. Initial assessment versus gradual change in early childhood behavior problems–Which better foretells the future?. *Psychol Schs*. 2018;55:1071–1085. https://doi.org/10.1002/pits.22150