Psychometric Analyses of Child Outcome Measures with American Indian and Alaska Native Preschoolers:

Initial Evidence from AI/AN FACES 2015

Technical Report

OPRE Report 2018-21
February 2018

ACKNOWLEDGMENTS:

The authors would like to express their appreciation to our Project Officers Meryl Barofsky and Mary Mueggenborg. We thank members of the AI/AN FACES 2015 Workgroup including federal officials, tribal Head Start directors, and tribal early childhood researchers. In particular, we appreciate the input of staff at OPRE and the Office of Head Start: Maria Woolverton, Laura Hoard, Aleta Meyer, WJ Strickland, Robert Bialas, and Angie Godfrey. Additionally, we are also grateful for the contributions of Jerry West, former Mathematica FACES project director, and of our partners at the Tribal Early Childhood Research Center, Michelle Sarche and Nancy Rumbaugh Whitesell for earlier comments on this work. We also acknowledge our partners Michelle Najarian and Fred Yan at Educational Testing Service. We thank the Mathematica team, including Louisa Tarullo, Nikki Aikens, Emily Moiduddin, Barbara Lepidus Carlson, Annalee Kelly, Melissa Hepburn, Joseph Baker, Michael Cavanaugh, John Carsley, Dan O'Connor, Gavin Fugere, Erin Slyne, Giuseppe Scaturro, John Chen, Leonard Brown, Andrew Hurwitz, Matthew Kalish, Ryan McInerney, Bill Reeves, Linda Mendenko, Scilla Albanese, Kathleen Feeney, Maya Reid, Charles Bush, Molly Crofton, Miriam Loewenberg, Serge Lukashanets, Anne Bloomenthal, Cheri Vogel, William Garrett, Kimberly Ruffin, Brigette Tran as well as Joan Gutierrez, Richard Godwin, Julio Garcia, Sheryl Safran, Marcia Tobias, Sue Golden, and Karen Markowski at the Survey Operations Center and all of the Mathematica field and telephone staff who collected the data. We would also like to thank the three Region XI AI/AN programs who participated during piloting and training activities. Most of all, we offer our gratitude to the staff, families and children of the 21 Region XI AI/AN Head Start programs, who opened their doors and shared their time with us.

Psychometric Analyses of Child Outcome Measures with American Indian and Alaska Native Preschoolers: Initial Evidence from AI/AN FACES 2015

Technical Report
OPRE Report 2018-21
February 2018

Submitted to:
Meryl Barofsky, Project Officer
Mary Mueggenborg, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Lizabeth Malone
Sara Bernstein
Sally Atkins-Burnett
Yange Xue
Mathematica Policy Research

Project Director:
Lizabeth Malone
Mathematica Policy Research
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221

Contract Number: HHSP23320095642WC/HHSP23337052T

Mathematica Reference Number: 40290.66j

Suggested citation: L. Malone, S. Bernstein, S. Atkins-Burnett, and Y. Xue (2018). *Psychometric Analyses of Child Outcome Measures with American Indian and Alaska Native Preschoolers: Initial Evidence from AI/AN FACES 2015*. OPRE Report 2018-21, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This page has been left blank for double-sided copying.

OVERVIEW

Head Start is a national program that aims to promote school readiness by enhancing the social and cognitive development of children through the provision of educational, health, nutritional, social, and other services to enrolled children and families. The program places special emphasis on helping preschoolers develop the reading, language, social-emotional, mathematics, and science skills they need to be successful in school. It also seeks to engage parents in their children's learning and to promote their progress toward their own educational goals (Administration for Children and Families [ACF] 2017). It also offers supports related to children's home or Native language and culture based on community needs and priorities. The Head Start program aims to achieve these goals by providing comprehensive child development services to economically disadvantaged children and families through grants to local public and private nonprofit and for profit agencies.

To date, the Head Start Family and Child Experiences Survey (FACES) has been a major source of descriptive information on Head Start and preschool children ages 3 to 5 years old who attend the program, with the most recent round in 2014. There are 12 regions for federal management of Head Start. FACES gathers data on Head Start programs, staff, children, and families from Regions I through X, which are the 10 geographically based Head Start regions nationwide. Regions XI and XII are not geographically based and instead are defined by the populations served. In 2015, a new study—the American Indian and Alaska Native Head Start Family and Child Experiences Survey (AI/AN FACES 2015)—focused on Region XI, which are programs operated by federally recognized American Indian and Alaska Native tribes.

Introduction

AI/AN FACES 2015 is the first national study of Region XI AI/AN Head Start children and their families, classrooms, and programs. The study is conducted by Mathematica Policy Research and its partner—Educational Testing Service—under contract to the Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. The study design, implementation, and dissemination has been informed by extensive collaboration with a workgroup comprised of Head Start directors from Region XI programs, early childhood researchers with experience working with tribal communities, Mathematica researchers, and federal government officials.

The AI/AN FACES 2015 study presents a new opportunity to explore the psychometric performance of commonly used measures of preschoolers' cognitive and social-emotional development. The reliability and validity of a measure are not inherent but depend on its use. Norming samples for most child assessment measures do not include large numbers of AI/AN children and as a result little is known about measure performance when administered to AI/AN children. Concerns exist about whether scores from these measures accurately reflect the children's abilities, skills, and knowledge. Previous smaller studies have used these measures with AI/AN children, but none were large enough to test the measures' psychometric performance. Child outcomes measures in AI/AN FACES 2015 were aligned with those in FACES 2014. Therefore, this alignment allows us to learn how standardized child development measures performed when administered to a large sample of AI/AN children.

This technical report describes the performance of cognitive and social-emotional measures of preschoolers' development for AI/AN children, using recent data from AI/AN FACES 2015 and FACES 2014.

Research Question

- What is the psychometric performance of cognitive and social-emotional measures of preschoolers' development with AI/AN children?

Purpose

The purpose of this technical report is to present findings from analyses of how preschool cognitive and social-emotional measures performed in AI/AN FACES 2015. We examined the internal consistency of measures when administered to AI/AN children, reviewed descriptive statistics as context of difference in mean ability across groups in the AI/AN FACES 2015 and FACES 2014 samples, conducted analyses of differential item functioning (DIF) within cognitive measures to compare the performance of AI/AN children and White children (including data from FACES 2014), and examined the strength of bivariate correlations between measures of similar constructs and different constructs to assess evidence of concurrent and discriminant validity. The findings, therefore, provide initial evidence on the reliability and validity of the measures for AI/AN preschoolers.

Findings

For most of the measures, findings from these analyses suggest that it is appropriate to report the AI/AN FACES 2015 preschool child outcomes scores, the exception being one of the two measures of executive function (Heads-Toes-Knees-Shoulders or HTKS, which was added to AI/AN FACES 2015 to expand measurement of this construct beyond what is used in FACES 2014).

- All measures demonstrated acceptable reliability with alphas of 0.70 or above.
- The strength of correlations between measures is in an expected pattern. Correlations are stronger between measures of similar constructs (for example, receptive and expressive language) than between different constructs (for example, social behavior and language).
- Among six cognitive measures flagged across reviews, none warrant additional follow-up based on the DIF analyses. Most cognitive measures did not show evidence of performing differently across groups based on DIF analysis. Two cognitive measures (Peabody Picture Vocabulary Test-Fourth Edition and Expressive One-Word Picture Vocabulary Test-Fourth Edition) had items demonstrating DIF; however, the number of items with DIF was close to or less than the number we would expect by chance and were balanced overall with some easier for AI/AN children and others easier for White children.
- None of the teacher- and assessor-reported social-emotional measures exhibited performance concerns based on the current review.
- Examination of the executive function measures indicated that the pencil tapping task is an appropriate measure for this sample. However, a floor problem was found with the HTKS, indicating the measure provided limited information to distinguish the children in this sample.

These analyses are based on a specific sample of children—AI/AN children in Head Start programs operated by federally recognized tribes. While this information provides initial evidence of the reliability and validity for these measures of cognitive and social-emotional development, researchers should keep in mind the diversity of tribal communities and the AI/AN population nationwide and in Head Start more generally as compared to Region XI AI/AN Head Start when considering the use of these measures with other AI/AN children.

Methods

The AI/AN FACES 2015 sample provides information about Region XI Head Start children, their families, classrooms, centers, and programs. We selected a sample of Region XI Head Start programs from the 2012-2013 Head Start Program Information Report, selecting one to two centers per program and two to four classrooms per center. Within each classroom, all children (both AI/AN and non-AI/AN) were selected for the study. Twenty-one programs, 37 centers, 73 classrooms, and 1,049 children participated in the study.

The FACES 2014 sample provides information at the national level about Head Start programs, centers, classrooms, and the children and families they serve. We selected a sample of Head Start programs from the 2012-2013 Head Start Program Information Report, with two centers per program and two classrooms per center selected for participation. Within each classroom, we randomly selected 12 children for the study. One-hundred seventy-six programs, 346 centers, 667 classrooms, and 2,206 children (in 60 programs) were still study participants in spring 2015.

This page has been left blank for double-sided copying.

CONTENTS

This page has been left blank for double-sided copying.

TABLES

This page has been left blank for double-sided copying.

I. INTRODUCTION

The American Indian and Alaska Native Head Start Family and Child Experiences Survey (AI/AN FACES) is the first national descriptive study of children and families enrolled in Head Start programs operated by federally-recognized tribes (known as Region XI AI/AN Head Start).[1] Since 1997, the Head Start Family and Child Experiences Survey (FACES) has been a major source of descriptive information on Head Start and the preschool children ages 3 to 5 years old who attend the program, but historically, FACES has not included Region XI due to the intensive community-based planning required to successfully carry out a study in partnership with tribal Head Start programs and communities. AI/AN FACES 2015, first conducted during the 2015-2016 program year fills this gap. It provides data to assess the service needs of the children and families in Region XI and to help inform policies and practices for addressing those needs. Region XI includes nearly 150 tribally run Head Start programs across the United States, which serve approximately 20,000 children, the majority of whom (85 percent) are American Indian and Alaska Native (AI/AN).

The AI/AN FACES 2015 study presents a new opportunity to explore the psychometric performance of commonly used measures of preschoolers' cognitive and social-emotional development. The reliability and validity of a measure are not inherent but depend on its use. Norming samples for most child assessment measures do not include large numbers of AI/AN children and as a result little is known about measure performance when administered to AI/AN children. Concerns exist about whether scores from these measures accurately reflect the children's abilities, skills, and knowledge. Previous smaller studies have used these measures with AI/AN children, but none were large enough to test the measures' psychometric performance. Pilot testing of the measures was conducted prior to the fielding of AI/AN FACES 2015, focusing on administration with fewer than 10 children. Child outcomes measures in AI/AN FACES 2015 were aligned with those in FACES 2014. Therefore, this alignment allows us to learn how standardized child development measures performed when administered to a large sample of AI/AN children.

This technical report presents findings from analyses of how the cognitive and social-emotional measures performed in AI/AN FACES 2015. We examined the internal consistency of measures when administered to children in AI/AN FACES 2015. We conducted analyses of differential item functioning (DIF) within measures comparing the performance of White and AI/AN children on the items relative to the overall ability of the children. We examined the strength of bivariate correlations between measures of similar constructs and different constructs to assess evidence of concurrent and discriminant validity.

In the remainder of this paper, we present an overview of AI/AN FACES 2015 and our analytic approach (Chapter I), key analysis findings to include an overview of findings and detailed findings on the performance of cognitive, social-emotional, and executive function

[1] In this document, we use the terms American Indian and Alaska Native (AI/AN), tribal, tribe, and Native to refer inclusively to the broad and diverse groups of American Indian and Alaska Native tribes, villages, communities, corporations, and populations in the United States, acknowledging that each tribe, village, community, corporation, and population is unique from others with respect to language, culture, history, geography, political and/or legal structure or status, and contemporary context.

measures and findings on the correlations among measures (Chapter II), and a summary and implications for researchers (Chapter III). For most of the measures, findings from these analyses suggest that it is appropriate to report the AI/AN FACES 2015 preschool child outcomes scores, the exception being one of the two measures of executive function (Heads-Toes-Knees-Shoulders, which was added to AI/AN FACES 2015 to explore expanding measurement of this construct).

A. Overview of AI/AN FACES

AI/AN FACES 2015 is a descriptive study of children and families who attend Region XI AI/AN Head Start programs. It was conducted in fall and spring of the 2015–2016 program year. At both time points, the study assessed the school readiness skills of children, surveyed parents about their family characteristics and home and community experiences, and asked teachers to rate children's social and emotional skills, classroom behavior, and approaches to learning, and to report on any concerns about the children and how the concerns were addressed. In spring 2016, observations of children's classrooms took place, and teachers, center directors, and program directors completed surveys about their backgrounds and the Head Start classrooms and programs (for example, classroom activities, culture and language resources, and staffing).

The AI/AN FACES 2015 study consists of a nationally representative sample of Region XI AI/AN Head Start programs, classrooms, and children. It represents all children—AI/AN and non-AI/AN—in Region XI. A total of 1,049 children and their families participated in AI/AN FACES 2015 from 73 classrooms in 21 Region XI Head Start programs. By design, the AI/AN FACES 2015 study provides a picture of the AI/AN children who attend Head Start programs in Region XI only, which serves 49 percent of all AI/AN children in Head Start. [2] AI/AN FACES 2015 is not representative of AI/AN children in Head Start in Regions I through X, or Region XII. The sample represents all children enrolled in Region XI AI/AN Head Start in the fall of 2015, including those who are attending for the first time and those who are attending a second year of the program. The study follows them through the spring of a single program year. Further, the study provides information about Region XI as a whole and not individual programs or tribes. In other words, the data support analyses at the Office of Head Start (OHS) region level, but **not** at geographic zones within Region XI or at the program or tribal level. AI/AN FACES 2015 data should not be considered representative of AI/AN children and families nationally, nor should they be considered representative of Head Start beyond Region XI.

Informed by the principles of tribal participatory research (Fisher and Ball 2003, for example), nearly two years of extensive planning preceded AI/AN FACES 2015, with advice from members of a workgroup that consisted of tribal Head Start directors, researchers, and federal government officials. Together, members of the AI/AN FACES Workgroup discussed and provided input on the AI/AN FACES 2015 design, its implementation, and how the findings would be disseminated with tribal voices at the forefront. The most recent round of FACES, conducted in fall 2014 through spring 2015 (Kopack Klein et al. 2017), served as a foundation

[2] AI/AN children served in Regions I through X are included in FACES; however, because they represent only a small percentage of all children in Head Start, the number of AI/AN children in the FACES sample is too small to provide reliable estimates. Additionally, although there are AI/AN children in Region XII (migrant and seasonal programs), the structure and nature of service delivery in Region XII programs is substantially different than in Regions I–XI so we excluded programs in Region XII.

for the study design, with modifications and additions to reflect Region XI AI/AN Head Start programs, families, and communities. As part of the collaborative design process, members provided advice on (1) the key research questions and information needs; (2) the population of interest (contributing to the overall sample design); (3) appropriate measures to assess growth and development of children Region XI AI/AN Head Start programs serve and to describe characteristics of children's homes and families, Head Start classrooms, and programs; and (4) research methods and practices that would be culturally grounded and effective in tribal communities. The Workgroup members also identified dissemination priorities relating to key analysis topics and the target audiences, and the formats best suited for presenting findings and providing appropriate context for tribal data. As it relates to measurement, in determining what measures to use or questions to ask, the study considered *aligning* with those used in FACES, *adapting* those used in FACES, or *adding* measures to address FACES measurement gaps relative to the priorities AI/AN FACES 2015 Workgroup members identified. For the assessment of child outcomes, the Workgroup members provided input for aligning with FACES.

We followed a multistage sampling approach, starting with the 21 programs.[3] We generally sampled two centers per program and two classrooms per center, though sometimes we sampled fewer (when there was only one center or classroom available to sample) or more than two to achieve sample targets.[4] Within each sampled classroom, we selected all children. Across the 2015–2016 program year, if a child in Region XI Head Start left the program at any time, he or she was no longer considered part of the study population.

In fall 2015 and spring 2016, AI/AN FACES 2015 used two major instruments to measure children's cognitive skills, physical outcomes, executive function, and social-emotional development:

(1) a direct child assessment—an untimed, one-on-one assessment, measuring each child's cognitive (language, literacy, and mathematics), physical (height and weight), and executive function outcomes. The assessment used standardized test material such as images from the Peabody Picture Vocabulary Test–Fourth Edition (PPVT–4; Dunn et al. 2006) and from the Woodcock-Johnson III measures (Woodcock et al. 2001, 2004). Web-based, computer-assisted personal interviewing (CAPI) facilitated the transition from one measure to the next without requiring the assessor to calculate stopping or starting points. Assessors asked children questions and showed them corresponding pictures on a second computer screen (separate from the computer screen viewed by the assessor). Assessors then entered the children's responses into the laptop, using software that ensured adherence to all basal and ceiling rules.[5] Assessors also rated children's behavior in the test situation.

The direct assessment includes two language paths: assessed in English and assessed in English, shortened assessment battery. Children are routed to the assessed in English path if the

[3] One program participated in spring 2016 only because of the time required for tribal approval.

[4] Due to a large proportion of Region XI AI/AN Head Start programs with only one center, we selected four classrooms in single-center programs whenever possible.

[5] Each measure is structured with specific starting points based on age and with specific rules for moving to earlier items (referred to as setting the basal or base of a child's ability) as well as on rules for stopping (referred to as establishing the ceiling or upper limit of a child's ability).

child uses English most often at home, or if they use a non-English language most often at home and make 12 or fewer errors on the preLAS (a warmup and language screener). Children who use a non-English language most often at home and who make 12 or more errors on the preLAS are assessed in English with the shortened assessment battery (language and physical measures only). In spring 2016, 934 children followed the assessed in English path (918 who used English most often at home, and 16 who used a non-English language most often at home), and 2 children followed the assessed in English, shortened assessment battery path.

(2) a Teacher Child Report (TCR)—teachers provided reports of children's school readiness skills and development. As part of the TCR, teachers described children's developmental outcomes by using web-based questionnaires or, if they preferred, paper questionnaires.

For more information on the AI/AN FACES 2015 study design, sample, and methodology see the AI/AN FACES 2015 User's Manual (Malone et al. 2018).

B. Overview of analytic approach

To examine measure performance in AI/AN FACES 2015, we conducted three analyses. We (1) estimated and reviewed descriptive statistics for each measure to determine the adequacy and extent of the variance in ability of the samples, as well as the reliability of the measure with each sample, (2) examined item functioning within a subset of measures (item fit, model fit, and DIF), and (3) computed bivariate correlations between measures. Descriptive statistics included the reliability, means, and variation found in AI/AN FACES 2015 and the FACES 2014 samples. This section provides an overview of these analyses.[6] In Table 1, we identify the measures included in the descriptive analyses and analyses of item functioning. All measures were included in the correlations. For information on the measures, please review the AI/AN FACES 2015 (Bernstein et al. 2018) and FACES 2014 (Aikens et al. 2017a) study reports with descriptive data tables.

Table 1. Approach to examining child outcomes measures in AI/AN FACES 2015

Measure	Instrument	Approach[a]
Language screener		
Simon Says (PreLAS 2000; Duncan and DeAvila 2000)	Child assessment	Review of descriptive statistics
Art Show (PreLAS 2000; Duncan and DeAvila 2000)	Child assessment	Review of descriptive statistics
Language development—receptive language		
Peabody Picture Vocabulary Test—Fourth Edition (Dunn et al. 2006)	Child assessment	Review of descriptive statistics; examine differential item functioning (DIF)

[6] These analyses are an initial exploration; additional future analyses to assess concurrent and predictive validity could include whether there are expected associations with other validated measures and expected longer term outcomes.

Table 1 (continued)

Measure	Instrument	Approach[a]
Language development—expressive language		
Expressive One Word Picture Vocabulary Test—4 (Martin and Brownell 2010)	Child assessment	Review of descriptive statistics; examine DIF
Literacy knowledge and skills—early writing		
Spelling (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics; examine DIF
Literacy knowledge and skills—alphabet knowledge and phonological awareness		
Letter-Word Identification (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics; examine DIF
Letter-Sounds items from the Early Childhood Longitudinal Study—Birth Cohort PreK version (ECLS-B) (http://nces.ed.gov/ecls/)	Child assessment	Review of descriptive statistics; examine DIF
Mathematics knowledge and skills		
Applied Problems (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics; examine DIF
Mathematics assessment items from ECLS-B and ECLS-K (kindergarten version) (http://nces.ed.gov/ecls/)	Child assessment	Review of descriptive statistics; examine DIF
Social-emotional development and approaches to learning		
26 items from Behavior Problems Index (Peterson and Zill 1986), Personal Maturity Scale (Entwisle et al. 1987), Social Skills Rating Scale (Gresham and Elliott 1990)	Teacher Child Report	Review of descriptive statistics
ECLS-K Approaches to Learning Scale	Teacher Child Report	Review of descriptive statistics
Leiter Examiner Ratings: (1) attention, (2) organization/impulse control, (3) activity level, (4) sociability (Leiter International Performance Scale Revised, Examiner Rating Scale; Roid and Miller 1997)	Assessor rating	Review of descriptive statistics
Executive function		
Pencil Tapping (Smith-Donald et al. 2007; Blair 2002; Diamond and Taylor 1996)	Child assessment	Review of descriptive statistics
Heads-Toes-Knees-Shoulders (Ponitz et al. 2009; McClelland et al. 2014)	Child assessment	Review of descriptive statistics

[a]Review of descriptive statistics includes review alongside the same statistics from FACES 2014 except for the HTKS given it was not administered in FACES 2014.

Descriptive statistics review. As a first step in understanding how AI/AN children are performing, we wanted to determine as context if there was a difference in mean ability across groups and the variance within the AI/AN FACES 2015 sample and the FACES 2014 sample. To evaluate how measures of cognitive and social-emotional development and executive function performed in AI/AN FACES 2015, we reviewed descriptive statistics for each measure (reported response ranges, weighted and unweighted means and standard deviations, and Cronbach's alphas[7]). We used spring data so that all children in the sample had the opportunity to learn the content taught in the preschool year, with one exception. For the preLAS (Simon Says and Art Show), we looked at fall data because in FACES 2014 the spring measures were administered to a subsample of children who did not demonstrate English proficiency in the fall. We compared these statistics for AI/AN children in AI/AN FACES 2015 (n = 718) to all children in FACES 2014 (n = 1,921) by using study reports (Bernstein et al. 2018, Aikens et al. 2017a) and unpublished frequencies and tabulations. We followed the approach used in previous FACES reporting of identifying meaningful differences where means differed by more than 0.25 standard deviations or percentages differed by more than 5 percentage points. For means, we examined differences using both AI/AN FACES 2015 and FACES 2014 standard deviations and if either of them was greater than 0.25 standard deviation we indicated this in the table. We did not conduct statistical tests for significance given this was an initial exploration. Any differences identified in the descriptive review are not indicators of bias or potential differential item functioning (or lack thereof) but provide context on differences in the groups' ability (mean and the variation in ability) as assessed by these measures.

Differential item functioning (DIF) analysis. In addition to the descriptive statistics review, for most cognitive measures, we examined DIF for AI/AN children compared with non-AI/AN children.[8] We did not conduct DIF analyses for two measures. First, we did not examine the language screener (*pre*LAS 2000 Simon Says and Art Show subtests) because we have done extensive analysis of these measures for FACES (Aikens et al. 2014). Also the primary purpose of this measure is as a language screener to assess English proficiency, and other studies previously examined DIF between speakers of different languages (Rainelli et al. 2017). Over 90 percent of children in Region XI speak English. Second, we did not examine DIF for a direct assessment of children's executive function—pencil tapping—given it is largely non-verbal, and the rule of the task is the same for each item, so item-level DIF is not appropriate. For the remaining cognitive measures, the analytic sample included 768 AI/AN children from AI/AN FACES 2015 with assessment data. Data on AI/AN children came from AI/AN FACES 2015 only, given that the sample of AI/AN children in FACES 2014 is too small to produce reliable estimates for this group alone. For purposes of the DIF analysis, non-AI/AN children were limited to White, non-Hispanic children from FACES 2014 and AI/AN FACES 2015 (a total of 638 children, 513 and 125 children, respectively).[9] This approach focuses on the comparison of AI/AN children to the majority group used in most publisher norming samples—White children (for example, more than 60 percent of the sample in the 3–5 age group in the PPVT–4 normative group are White, non-Hispanic).[10] We excluded White children who are Hispanic because that

[7] Cronbach's alphas of 0.70 or above are generally considered to be in the acceptable range.

[8] American Indian and Alaska Native children are those whose parents reported that they were American Indian or Alaska Native only or in combination with another race or Hispanic ethnicity.

[9] The sample of White, non-Hispanic children was pooled from AI/AN FACES 2015 and FACES 2014 to allow an adequate sample size for the DIF analyses (the number of White, non-Hispanic children from AI/AN FACES 2015 alone would have been too small).

[10] DIF analyses comparing White and Black and White and Hispanic children are conducted for both the PPVT and EOWPVT as part of their measure development process (described in their technical manuals: Dunn et al. 2006;

would likely add linguistic issues unrelated to the cultural differences and/or skills being assessed. In the remainder of this report, we refer to White, non-Hispanic, non-AI/AN children as White children.

To examine DIF, we used item response theory (IRT) for each cognitive outcome measure. IRT models estimate the level of an underlying trait in a person (that is, the estimated ability) and how difficult it is to answer a question correctly (that is, item difficulties). Both are estimated on the same scale, so the model provides the probability that a person with a given ability will correctly answer a given item in an assessment. DIF occurs when children with the same estimated ability have a different probability of giving a correct answer, indicating that the item could be biased.[11] However, DIF alone is not proof of bias or that the item is unfair to one group of children or another, as there could be real differences between the two groups that account for their responses to an item. For example, one group of children could be exposed to math concepts that differ from the math concepts presented to children in another group. We used spring data instead of fall data for these analyses so that all children in the sample had the opportunity to learn the content taught in the preschool year, but this would not account for the fact that curricula differ across centers. Some content in the assessment[12] may not have been emphasized in one or more classrooms. For the PPVT–4, EOWPVT–4, and the WJ III Spelling, we used a Rasch (one-parameter IRT) model. A three-parameter model was used to examine DIF for the following four measures as part of their score construction: WJ III Letter-Word Identification, Letter-Sounds items from the ECLS–B preschool assessment, WJ III Applied Problems, and ECLS Mathematics. DIF analyses are conducted at the item level, so the size of the sample for each item will vary because not all children receive every item (due to publisher basal and ceiling rules). DIF analyzes the items, rather than the children, and requires that there is a large enough sample of respondents (children) for variation in ability. DIF is conducted with raw data, and the sample size of children needed depends on the number of parameters estimated. For the Rasch analysis (a one-parameter model), we included only those items that had at least 100 children in each group (AI/AN and White). Because most of the assessments are adaptive, children do not take every item by design. IRT uses information from all of the cases and all of the items to iteratively estimate the difficulty of the items and ability of the children on that construct. If a child skipped an item (did not respond), the Rasch model would estimate the

Martin and Brownell 2010). Previous rounds of FACES also compared these groups when developing scores for FACES, with no evidence of bias in those comparisons of the WJ III or ECLS-B measures.

[11] An item is considered problematic if it qualifies as level C DIF, or a DIF greater than 1.645 (Delta units) based on Mantel-Haenszel (Zwick 2012) or 0.64 (logits) using PROX (Linacre and Wright 1989). DIF is an indicator of potential bias and requires further evaluation to determine if it is a true difference in the dimension measured by that item or if the items is not fair for a subgroup due to something unrelated to the construct.

[12] Given the small number of centers in the AI/AN FACES 2015 sample, if one center taught content in a given domain and other centers did not include that content, the analysis could detect DIF. It is important to note that DIF analyses do not adjust for the clustering of children in classrooms within centers. In AI/AN FACES 2015, we selected all children in two to four classrooms for a given center, so center differences in instruction would likely influence the results. For example, children in centers that spend more time on number concepts would be more likely to correctly answer items addressing number concepts than children with a similar underlying ability in mathematics in centers that do not spend much time on number concepts. The concentration of AI/AN children is high in some centers, potentially biasing the results if these are the centers that spend more time on mathematics. However, by focusing on spring outcomes, we minimized the differences in exposure within the centers that are tied to differences in the sequencing of the curriculum.

probability of the child answering correctly based on how other children of similar ability responded to that item. For the three-parameter models, in addition to the minimum of 100 in each group, a combined sample of 400 children on an item was required given the need to estimate additional parameters. While the descriptive statistics review included both weighted and unweighted estimates, the DIF analyses was conducted unweighted. The goal of DIF analyses is not to produce representative estimates of children's ability, but examine group differences in the difficulty of the items.

Analytic considerations. One caveat about these analyses is that it includes data from both FACES 2014 and AI/AN FACES 2015, which were collected at different times. FACES 2014 data were collected during the 2014–2015 program year, but AI/AN FACES 2015 data were collected one year later (2015–2016 program year). We do not think this was an issue for our analysis given that IRT models estimate ability as an underlying trait, which should not change with respect to the year in which data were collected. However, curricular emphases and life experiences could change from one program year to the next, potentially affecting some item difficulties. For example, an item with content that involves a rarely occurring event might have a different difficulty if such an event happened in the intervening year (for example, an item that talks about earthquakes or flooding in the year before and in the year after a major earthquake or flood). Similarly, a change in national emphasis on one domain such as measurement or another domain, such as spatial reasoning/geometry, in mathematics instruction could also affect the difficulty of related items. We do not believe that this was an issue in these analyses. For example, none of the items for which DIF was identified is aligned with major events in the 2014–2015 or 2015–2016 program years, nor is it aligned with changes in early childhood standards.

In this section we provide a summary of findings across the descriptive statistics review and analysis of item functioning (Table 2). We also provide more detail about findings on cognitive measures and on social-emotional and executive function measures, respectively.

For understanding AI/AN children's performance, the descriptive statistics showed differences in means of greater than 0.25 standard deviations for 9 of 18 measures between AI/AN children in AI/AN FACES 2015 and all children in FACES 2014. These differences do not necessarily indicate problems with item or measure functioning but provide context on differences in mean ability. We considered weighted and unweighted means and standard deviations as part of the review of descriptive statistics, but any meaningful differences in means were consistent across weighted and unweighted statistics. For cognitive measures, four of nine measures showed differences, with three favoring children from FACES 2014. For social-emotional measures, five of nine measures showed differences, with four favoring AI/AN children. For one measure, Head-Toes-Knees-Shoulders (HTKS), we could not compare to FACES, but the descriptive statistics show limited variability, discussed further in Chapter II.

For understanding measure performance, we found no evidence of DIF for most measures (five of seven measures). It should be noted that although the analyses uncovered evidence of DIF for some items in two of the measures, this is not unusual because we analyzed so many items,[13] and the sample was large. DIF relies on a statistic that is more sensitive when the samples are large, so even a small difference can be statistically significant. The number of items showing evidence of DIF was close to or less than what would be expected due to chance. In our analyses with these samples DIF was identified for the EOWPVT–4 and the PPVT–4; the DIF favored the AI/AN children for some of the items and the White children for other items in the same measure. The direction of the DIF was balanced across groups and so does not suggest that there is systematic bias in the AI/AN FACES 2015 scores. Both of these measures assess children's vocabulary, and one would expect that some children may be exposed to some vocabulary while other children would have been exposed to other vocabulary at the same level of difficulty. Based on the results of our analyses, we include the raw, standard, and W and/or IRT scores for the measures administered in AI/AN FACES 2015.

[13] We may expect differences in at least 5 percent of items due to chance ($p \leq 0.05$). Corrections for multiple comparisons were not employed.

Table 2. Summary of findings on child outcomes measures in AI/AN FACES 2015: Spring 2016

Measure	Instrument	Analytic approach[a]	Summary	Cronbach's alpha[b]
Cognitive measures				
Language screener[c]				
Simon Says (*pre*LAS 2000; Duncan and DeAvila 2000)	Child assessment	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.90
Art Show (*pre*LAS 2000; Duncan and DeAvila 2000)	Child assessment	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from AI/AN FACES 2015.	0.80
Language development—receptive language				
Peabody Picture Vocabulary Test, Fourth Edition (Dunn et al. 2006)	Child assessment	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.97
		Examine DIF	DIF analyses identified 9 items with DIF between AI/AN and White children (4.7% of items).	
Language development—expressive language				
Expressive One Word Picture Vocabulary Test – 4 (Martin and Brownell 2010)	Child assessment	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.96
		Examine DIF	DIF analyses identified 8 items with DIF between AI/AN and White children (5.9% of items).	
Literacy knowledge and skills—early writing				
Spelling (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from FACES 2014.	0.80
		Examine DIF	No DIF detected.	

Table 2 (continued)

Measure	Instrument	Analytic approach[a]	Summary	Cronbach's alpha[b]
Literacy knowledge and skills—alphabet knowledge and phonological awareness				
Letter-Word Identification (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from FACES 2014.	0.86
		Examine DIF	No DIF detected.	
Letter-sounds knowledge (ECLS-B letter-sounds IRT score) (http://nces.ed.gov/ecls/)	Child assessment	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from FACES 2014.	0.89[d]
		Examine DIF	No DIF detected.	
Mathematics knowledge and skills				
Applied Problems (Woodcock-Johnson III Tests of Achievement; Woodcock et al. 2001)	Child assessment	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.88
		Examine DIF	No DIF detected.	
Early math (ECLS-B math IRT score) (http://nces.ed.gov/ecls/)	Child assessment	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.76[d]
		Examine DIF	No DIF detected.	
Social-emotional and executive function measures				
Social-emotional development and approaches to learning				
26 items from Behavior Problems Index (Peterson and Zill 1986), Personal Maturity Scale (Entwisle et al. 1987), Social Skills Rating Scale (Gresham and Elliott 1990)				
Total behavior problems	Teacher Child Report	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.87
Social skills	Teacher Child Report	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.90

Table 2 (continued)

Measure	Instrument	Analytic approach[a]	Summary	Cronbach's alpha[b]
ECLS-K Approaches to Learning Scale	Teacher Child Report	Review of descriptive statistics	No issues identified in review of descriptive statistics from AI/AN FACES 2015 and FACES 2014.	0.92
Leiter Examiner Ratings (Leiter International Performance Scale Revised, Examiner Rating Scale; Roid and Miller 1997)				
Attention	Assessor rating	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from AI/AN FACES 2015.	0.97
Organization/impulse control	Assessor rating	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from AI/AN FACES 2015.	0.95
Activity level	Assessor rating	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from AI/AN FACES 2015.	0.89
Sociability	Assessor rating	Review of descriptive statistics	Means differ by more than .25 standard deviations between AI/AN FACES 2015 and FACES 2014, with higher scores from AI/AN FACES 2015.	0.90
Executive function				
Pencil Tapping (Smith-Donald et al. 2007; Blair 2002; Diamond and Taylor 1996)	Child assessment	Review of descriptive statistics	Means differ by more than five percentage points between AI/AN FACES 2015 and FACES 2014, with higher scores from FACES 2014.	0.92
Heads-Toes-Knees-Shoulders (Ponitz et al. 2009; McClelland et al. 2014)	Child assessment	Review of descriptive statistics[e]	Taken in the context of other studies using Head-Toes-Knees-Shoulders (HTKS), mean scores from AI/AN FACES 2015 are lower than scores from other studies using the same scoring approach. However, the majority of children scored zero in AI/AN FACES 2015 indicating a floor effect.	0.95

[a] Review of descriptive statistics included review alongside the same statistics from FACES 2014.
[b] Cronbach's alphas are for AI/AN children. The AI/AN FACES 2015 descriptive tables and study design report (Bernstein et al. 2018) present alphas for all children in AI/AN FACES 2015.
[c] We examined scores for Simon Says and Art Show to help determine if AI/AN children were being nonresponsive because they were being assessed by an unfamiliar adult (which would have implications for the EOWPVT–4). This did not appear to be an issue. For these measures, we examined fall scores given in the spring only children who did not pass the screener in the fall received the Art Show subtest. Cronbach's alphas provided in this table are based on fall 2015.
[d] For these IRT scores, we present the reliability coefficient of the number right of the items that a measure contributed to the combined IRT score. The reliability of the IRT score is only available for a combined ECLS-WJ score and is based on the reliability of theta and applies to both letter-sounds (0.77) or early math (0.88) IRT scores. The IRT model is estimated for all children, so there is no separate IRT score reliability for AI/AN children only.
[e] Note that because this measure was not used in FACES 2014, we cannot compare descriptive statistics to FACES 2014. Instead, we compared to other studies of preschool children that scores HTKS on the same 0-60 scale used in AI/AN FACES 2015.

A. Cognitive measures findings

All cognitive measures demonstrated acceptable reliability with alphas of 0.70 or above. We further reviewed descriptive statistics for all nine cognitive measures and conducted a DIF analysis for seven of the nine (excluding the *pre*LAS Simon Says and Art Show as described in Chapter I on our analytic approach). We focus here on the DIF results. Table 3 provides information from the descriptive statistics review. As described earlier, comparisons of means are between AI/AN children from AI/AN FACES 2015 and all children from FACES 2014 and are considered meaningful if greater than 0.25 standard deviations or 5 percentage points.[14] Differences in means do not indicate the presence of bias, and we do not consider the differences in means for these measures to be a cause for concern as no items within the measures were identified as having DIF.

Table 3. Summary statistics for children's cognitive measures identified to have context differences in descriptive statistics review: Spring 2015 and Spring 2016

	AI/AN children in AI/AN FACES 2015		All children in FACES 2014	
Construct (Measure)	**n**	**Mean and Range**	**n**	**Mean and Range**
Early writing (WJ III Spelling standard score)	684		1,704	
Mean (SD)		83.8 (16.3)		90.4 (16.7)
Range		32-122		29-133
Letter-word knowledge (WJ III Letter-Word Identification standard score)	680		1,699	
Mean (SD)		90.3 (12.2)		95.8 (13.6)
Range		62-134		53-167
Letter-sounds knowledge (ECLS-B letter-sounds IRT score)	251		1,047	
Mean (SD)		0.5 (0.7)		1.7 (2.2)
Range		0-3		0-9.3

Source: Spring 2016 AI/AN FACES 2015 Assessor Rating and Direct Child Assessment and Spring 2015 FACES 2014 Assessor Rating and Direct Child Assessment.

Note: AI/AN FACES 2015 weighted estimates in this table are representative of all AI/AN children enrolled in Region XI Head Start programs in fall 2015 and who were still enrolled in spring 2016. FACES 2014 statistics are weighted to represent all children enrolled in Head Start in fall 2015 and who were still enrolled in spring 2016.

The n columns in this table include unweighted sample sizes to identify the number of children with valid data on each of the constructs.

Standard scores in this table reflect an individual's performance relative to English-speaking children of the same age nationally. These scores have a mean of 100 and a standard deviation of 15. IRT-based scores provide information on children's absolute performance at a specific point in time.

Differences were defined meaningful for context when they were at least 0.25 standard deviations (based on either study's standard deviation) or 5 percentage points but were not statistically tested.

[14] Differences refer to means more than 0.25 standard deviations apart. We did not conduct formal significance tests given the analysis was for context purposes. Weighted statistics are reported, but differences in means are consistent across weighted and unweighted statistics (greater than 0.25 standard deviations).

Based on the DIF analysis, we found no evidence that five of seven measures examined performed differently across groups: (1) WJ III Spelling, (2) WJ III Letter-Word Identification, (3) ECLS-B letter sounds, (4) WJ III Applied Problems, and (5) ECLS mathematics. The DIF analyses found that AI/AN and White children differ on some items in the PPVT–4 and EOWPVT–4.

- **PPVT–4.** Across the samples, the first 108 items had at least 100 children from each group complete the item that we could conduct DIF analysis. The highest item number administered to any child was 144 for AI/AN children and 192 for White children.
  - Nine of the 108 items examined in the PPVT–4 demonstrated DIF (that is, 8.3 percent of items attempted by at least 100 children). We would expect to find DIF for 5 percent of the items, or about 6 items in the PPVT–4, just by chance. Thus, the number of items with DIF is close to what we would expect by chance.
  - Given that PPVT–4 items are administered in sets of 12, we further examined the pattern of DIF within sets by using a more conservative criterion of |0.50| logits. Through this approach, we identified 24 out of 108 items with DIF between AI/AN children and White children. Please note that an entire set must be administered even if the ceiling rule is met before reaching all 12 items. Eight sets (with one to six items per set) contained items identified with DIF. Within each set, some items favored AI/AN children and some items favored White children. With the more conservative criteria, 14 items were easier for AI/AN children, and 10 items were easier for White children.
  - The mean square infit statistics are all in the acceptable range (i.e., below 1.3). The good infit indicates that children responded to items close to their ability level in expected ways. Both the difficulty of the items and ability of the children are estimated on the same scale.
  - The outfit statistics are mostly in the acceptable range (i.e., below 1.3), with the exception of five items used to establish a basal score. Outfit statistics identify items that had unexpected responses on items that are farther from the child’s ability. The outfit for the early items indicates that unexpected incorrect responses on the items were more frequent for children whose vocabulary ability is far higher than the difficulty level of the items. Items at the extremes of difficulty (very easy or very difficult) are more likely to have poorer fit. For example, the probability that a high ability student would get an easy item correct is very high and so an incorrect answer on a very easy item by these high performers would signal misfit. The DIF contrast on these items indicates that the items were easier for AI/AN children compared to White children.
  - The pattern for items in the mid-range of the sets that have DIF is balanced. Some items were easier for AI/AN children, and some were easier for White children. All but one of these sets has a mix of items favoring AI/AN and White children. Therefore, there is no evidence that items in the mid-range are more difficult overall for AI/AN children than they are for White children.

- **Expressive One Word Picture Vocabulary Test–4.** Across the samples, 83 items had at least 100 children from each group complete the item that we could conduct DIF analysis. The highest item administered was 135 for AI/AN children and 136 for White children.
  - Analyses of 83 items indicated that there were 8 items with DIF between AI/AN children and White children (or 9.6 percent of items attempted by at least 100 children in each group). We would expect to find DIF for 5 percent of items, or about 5 items in the EOWPVT–4, just by chance. Thus, the number of items with DIF is close to what we would expect to see by chance.
  - Three of the EOWPVT–4 items with detectable differences were easier for AI/AN children (as indicated by negative DIF contrast), and five were more difficult for AI/AN children than for White children. The majority of the items that were more difficult for the AI/AN children were those attempted by fewer than 375 of these children, and they appear to be the items that were used to confirm the stop rule for these children.
  - The infit statistics showed good fit for all 8 items, with the mean square infit below 1.3. The good infit indicates that children responded to the items close to their ability level in expected ways.
  - Overall, these items also had good outfit. Only two items had an outfit mean square of greater than 1.3, and these are among the easier items. The outfit for the early items indicates that unexpected incorrect responses on the items were more frequent for children whose vocabulary ability is far higher than the difficulty level of the items.
  - Most of the DIF was at the extreme ends of difficulty for the items administered in the EOWPVT–4, with the DIF on more difficult items favoring White children. This suggests that additional analyses would be warranted if these items are used in future studies occurring in kindergarten and/or grade 1. The number of difficult items with DIF is low, and so the estimates from the EOWPVT are acceptable for current use during Head Start.[15]

B. Social-emotional and executive function measures findings

As initial evidence of validity, we reviewed descriptive statistics for nine social-emotional and executive function measures, focusing on the summary or mean scores and variation. We did not conduct additional analyses at the item level for DIF as individual differences would be expected due to the context in which children are being rated; therefore, looking at summary scores provides more information on whether the measure is an accurate assessment of children across different groups. Furthermore, on these teacher-rated measures, there is a clustering effect of responses within teachers, and differences might reflect individual differences between teachers rather than between the different groups.

[15] The items with DIF at the high end of the range given to children in AI/AN FACES 2015 and in FACES 2014 are words that are learned in school. If future researchers use this measure with school-age AI/AN children, DIF analyses should be done to see if these items continue to show evidence of DIF in kindergarten and grade 1. If so, then they should be looked at more closely by a panel of culture and content experts.

Our review of descriptive statistics did not show any differences between AI/AN children in AI/AN FACES 2015 and all children in FACES 2014 with respect to the three of the nine social-emotional and executive function measures listed below. Alphas were 0.70 or above, and means were within 0.25 standard deviations of each other.

- Total problem behaviors
- Social skills
- Approaches to learning

These measures are based on the TCR. In the design phase of the study, tribal Head Start directors in the AI/AN FACES 2015 Workgroup were asked to review the TCR, and they noted they did not expect teachers to have trouble using the items. Nor did the directors express concerns about appropriateness, as they felt that they asked teachers to consider the same types of behaviors.

We found differences in means of greater than 0.25 standard deviations (or 5 percentage points) in the five measures below.

- Leiter-R attention
- Leiter-R organization/impulse control
- Leiter-R activity
- Leiter-R sociability
- Pencil tapping

The first four measures are from the Leiter Examiner Ratings completed by assessors, and the last is a measure of executive function from the direct child assessment. Descriptive statistics for these five measures are reported in Table 4. The Leiter scores favor AI/AN children. Because assessor differences account for some of the variance in ratings, we did not conduct item level analyses with these ratings. We did examine the correlations with teacher ratings and pencil tapping (see Section C in this chapter).

For pencil tapping, in the spring, there is a difference of more than 5 percentage points between AI/AN children in AI/AN FACES 2015 and all children from FACES 2014. The spring estimates are based on all children, including those who did not take the assessment in the fall such as those who were 3 years old in the fall and not old enough for this task. Therefore, we further examined scores in fall and the change across the program year. In the fall, AI/AN children averaged 37.7 percent on pencil tapping, with a standard deviation of 31.7 (compared with 46.2 percent and 35.2 percent, respectively, in FACES 2014) (Bernstein et al. 2018; Aikens et al. 2017b). Looking at change, for children who had both fall and spring data, we see that the scores for both groups increased from fall to spring, providing initial evidence that the pencil tapping task is sensitive to change for both groups even when there are differences in mean scores.[16] Further, for both groups, the fall scores are below what we would expect to see by chance (that is, getting 50 percent of the trials correct), and the gains are similar, with fall-to-spring change of 14 percentage points for AI/AN FACES 2015 and 18 percentage points for the FACES 2014 sample. This sensitivity to change with increased age (and the presence of Head Start support), therefore, provides some initial evidence of validity. Additionally, the measure has prior evidence of validity for White and Hispanic children (e.g., Bierman et al. 2008), and we see in the current study the measure is performing in a similar way (that is, measuring age-related change) for AI/AN children.

Table 4. Summary statistics for children's social-emotional measures identified to have context differences in descriptive statistics review: Spring 2015 and Spring 2016

	AI/AN children in AI/AN FACES 2015		All children in FACES 2014	
Construct (Measure)	n	Mean and Range	n	Mean and Range
Assessor rating during direct assessment				
Attention (Leiter-R)	686		1,828	
Mean (SD)		26.0 (6.0)		23.1 (6.7)
Range		0-30		0-30
Organization/impulse control (Leiter-R)	686		1,828	
Mean (SD)		21.1 (4.7)		18.7 (5.2)
Range		0-24		0-24
Activity (Leiter-R)[a, b]	686		1,828	
Mean (SD)		10.0 (2.7)		9.3 (2.9)
Range		0-12		0-12
Sociability (Leiter-R)	686		1,828	.
Mean (SD)		13.9 (2.2)		13.2 (2.7)
Range		0-15		0-15
Direct child assessment				
Executive function (pencil tapping[c])	554		1,530	
Mean (SD)		48.1 (32.9)		59.3 (34.7)
Range		0-100		0-100

Source: Spring 2016 AI/AN FACES 2015 Assessor Rating and Direct Child Assessment and Spring 2015 FACES 2014 Assessor Rating and Direct Child Assessment.

Note: AI/AN FACES 2015 weighted estimates in this table are representative of all AI/AN children enrolled in Region XI Head Start programs in fall 2015 and who were still enrolled in spring 2016. FACES 2014 statistics are weighted to represent all children enrolled in Head Start in fall 2015 and who were still enrolled in spring 2016.

The n columns in this table include unweighted sample sizes to identify the number of children with valid data on each of the constructs.

Raw scores are reported unless noted otherwise.

Differences were defined meaningful for context when they were at least 0.25 standard deviations (based on either study's standard deviation) or 5 percentage points but were not statistically tested.

[a] Means are not within 0.25 standard deviations based on the AI/AN FACES 2015 standard deviation only.

[b] Higher scores indicate better activity level.

[c]In the Pencil Tapping task, children are asked to inhibit the natural response to imitate the adult assessor exactly (or to tap repeatedly) and instead to keep in mind that the rule is to do the opposite of what the assessor does. Reported scores reflect the percentage of times the child tapped correctly. They can take on any value from 0 to 100, with higher scores indicating better skills on the task. The task is only administered to children age 4 and older at the time of the direct assessment.

[16] For AI/AN FACES 2015, children completing pencil tapping at both time points tapped correctly 38.5 percent of the time on average in the fall and 52.6 percent in the spring. For FACES 2014, children completing pencil tapping at both time points tapped correctly 47.1 percent of the time on average in the fall and 65.0 percent in the spring.

HTKS was not administered in FACES 2014, so we could not compare descriptive statistics for AI/AN FACES 2015 to FACES 2014, but the AI/AN FACES 2015 statistics raise concerns about this measure for this group of children in the given age range. HTKS was added to the assessment battery to explore expanding the measurement of executive function. Its use in low-income samples has been limited, and its inclusion in the AI/AN FACES 2015 battery was discussed during the design phase as exploratory. In terms of actual performance in AI/AN FACES 2015, we find that over two-thirds of children scored a zero on HTKS in the spring, indicating it is not providing information for most children (also known as a serious floor problem). Further, it is difficult to find appropriate comparisons, as there is no normative group. Studies using the HTKS vary by the version used (the HTKS was administered in AI/AN FACES 2015 in 3 parts with 10 items each) and by the scoring approach (allowing partial credit or not). In other studies using the same approach used in AI/AN FACES 2015, the scores tend to be higher than in AI/AN FACES 2015. For example, in a sample of children from 31 classrooms (with 51 percent of study children attending Head Start classrooms), the average score at the end of the preschool year was 23.0 with a possible score of 60 and a standard deviation of 18.6 (Schmitt et al. 2014). In AI/AN FACES 2015, the mean in the spring is 5.9 and standard deviation is 12.1.[17]

While no concerns were raised during the study design phase as part of the AI/AN FACES 2015 Workgroup's review, it is possible that children from cultures with strong respect for elders would find this task particularly difficult. A task in which a child is asked to do the opposite of what an adult says will be strange to any child, but it could be that this is especially difficult given AI/AN children's cultural context, in which respect for adults is an important shared value (Lynch and Hanson 2004).[18] Another factor explaining the low scores might be that executive function measures do have a cognitive component, and as noted above, the means for some cognitive measures appear to be lower for AI/AN children in AI/AN FACES 2015 than for all children in FACES 2014. However, given that the majority of children received a score of zero, indicating that HTKS was not a good measure for the group of children in the given age range, AI/AN FACES 2015 does not include HTKS in reporting or data products. The measure developer noted the use of the practice items in scoring can help to add variability, but as the child is learning the task and receiving feedback during those items, we focused on the test items for scoring (McCelland, personal communication, August 2017).

C. Correlations across measures

We also examined the correlations between the cognitive and social-emotional outcome measures for AI/AN children in AI/AN FACES 2015. It is important to keep in mind that measures gathered using the same method (for example, direct assessments or teacher reports) have slightly stronger correlations due to shared method variance. Therefore, correlations will be higher among measures such as the PPVT–4 and WJ III Letter Word Identification that are direct

[17] The mean age at assessment for AI/AN children was 69 months, and the mean age at assessment in Schmitt et al. was 61 months.

[18] Further underscoring these differences is the fact that when we look at *all* children from AI/AN FACES 2015, the mean rises to 7.3 (SD=13.9). This includes the AI/AN children, so the increase in mean score is the result of adding only about 80 non-AI/AN children to the sample.

adaptive assessments than those with different methods (for example, pencil tapping and teacher-reported approaches to learning).

We examined bivariate correlations of W scores or, if W scores were not available, raw scores to allow for variation across age. Standard scores account for age and attenuate the correlation across measures. For AI/AN FACES 2015−, we find that the strength of correlations between measures is in an expected pattern. These results provide evidence of concurrent and discriminant validity. Correlations are stronger between measures of similar constructs (for example, receptive and expressive language) than between different constructs (for example, social behavior and language):

- Among the cognitive measures (Appendix Table A.1), measures assessing similar constructs have high correlations (0.74 to 0.95). For example, the receptive vocabulary and expressive vocabulary measures (PPVT–4 and EOWPVT–4) have a correlation of 0.74. These language measures are moderately correlated with literacy measures (e.g., WJ III Letter-Word Identification and ECLS Letter Sounds), ranging from 0.34 to 0.48, and with math measures (e.g., WJ III Applied Problems and ECLS Math), ranging from 0.64 to 0.71, the latter indicating overlap of basic concepts and language abilities in mathematics assessments for this age group. Literacy and math measures are also moderately correlated (0.36 to 0.64). Similar patterns have been found in the past for FACES 2009 (Kopack Klein et al. 2017).
- Among the teacher-reported social-emotional measures (Appendix Table A.2), we find moderate to high correlations in expected directions. Positive skills and problem behaviors are negatively correlated (-0.45 to -0.64); social skills and approaches to learning are positively correlated (0.76).
- Across teacher and assessor report of social-emotional development (Appendix Table A.2), correlations are weak to moderate. Teachers rate based on behavior observed in the classroom, and assessors rate based on observations during brief assessments. Given these very different sources of information, it is not surprising that the correlations are relatively weak. Assessor and teacher reports of positive behavior are positively correlated (0.28 to 0.34). Assessor report of positive social-cognitive behaviors and teacher-reported problem behaviors are negatively correlated (-0.21 to -0.33) at similar strength of correlation.
- The correlations between cognitive measures and assessor-report of behavior with the Leiter (Appendix Table A.1) are weak to moderate (0.11 to 0.35). The Leiter scores are based on reports of cognitive-social behavior during the testing situation. The more moderate correlations reflect that there is a cognitive aspect of children’s attention and organization as measured by the Leiter.
- Across the teacher-report social-emotional measures and direct cognitive measures, correlations are weak to moderate (Appendix Table A.3). Teacher-reported problem behaviors are negatively correlated with direct cognitive measures (-0.10 to -0.37); teacher reports of positive behavior are positively correlated with direct cognitive measures (0.14 to 0.41), with the higher estimates being with math measures.

This page has been left blank for double-sided copying.

Our review and analyses of 18 cognitive and social-emotional development measures administered in AI/AN FACES 2015 support a recommendation that the scores are valid for this sample of Region XI AI/AN Head Start preschoolers on all measures except for the HTKS.

- All measures demonstrated acceptable reliability with alphas of 0.70 or above.
- The strength of correlations between measures is in an expected pattern. Correlations are stronger between measures of similar constructs (for example, receptive and expressive language) than between different constructs (for example, social behavior and language).
- Among the six cognitive measures flagged across reviews, none warrant additional follow-up. Although four cognitive measures show potential differences in mean ability, no items within these measures demonstrated DIF. On the two measures with items demonstrating DIF, the number of items with DIF was close to or less than the number we would expect by chance. The DIF among the items were balanced overall with some easier for AI/AN children and others easier for White children.
- None of the teacher- and assessor-reported social-emotional measures exhibit performance concerns.
- Examination of the executive function measures indicated that the pencil tapping task is an appropriate measure for this sample. While the means on the pencil tapping task were lower for AI/AN children in AI/AN FACES 2015 than for all children in FACES 2014, both were below chance in the fall (that is, tapping correctly for less than 50 percent of the trials) and the change across the program year was similar suggesting sensitivity to developmental change for both groups. In addition, the pencil tapping task was correlated in expected ways with cognitive and social-emotional measures indicating concurrent and discriminant validity.
- A serious floor problem was found with the HTKS, indicating the measure provided limited information to distinguish the children in this sample. The majority of children in AI/AN FACES 2015 did not respond correctly to any of the trials (items).

It is important to remember these analyses are based on a specific sample of children—AI/AN children in Head Start programs operated by federally recognized tribes. While this information provides initial evidence of the reliability and validity for these measures of cognitive and social-emotional development, researchers should keep in mind the diversity of tribal communities and the AI/AN population nationwide and in Head Start more generally as compared to Region XI AI/AN Head Start when considering the use of these measures with other AI/AN children. In other words, AI/AN FACES 2015 data should not be considered representative of AI/AN children and families nationally, nor should they be considered representative of Head Start beyond Region XI.

An additional consideration for future studies is that the expressive vocabulary measure (EOWPVT–4) indicated potential concerns for older children. The items with DIF at the high end of the range given to children in AI/AN FACES 2015 and in FACES 2014 favored White children. These items with higher difficulty were at the ceiling for the AI/AN group while many in the White group of children continued to subsequent sets before reaching a ceiling. The DIF may be related to lower overall ability. These more difficult words include many words typically learned in school. If future studies use this measure with school-age AI/AN children, DIF analyses should be done to see if these items continue to show evidence of DIF in kindergarten and grade 1. If so, then they should be examined more closely by a panel of culture and content experts.

Finally, the reliability and validity of a measure are not inherent but depend on its use. The analyses here provide initial evidence that the measures provide reliable and valid measures of children's cognitive and social-emotional development in AI/AN FACES 2015.

Aikens, N., A. Kopack Klein, E. Knas, J. Hartog, M. Manley, L. Malone, L. Tarullo, and S. Lukashanets. *Child and Family Outcomes During the Head Start Year: FACES 2014-2015 Data Tables and Study Design.* OPRE Report 2017-100. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017a.

Aikens, N. A. Kopack Klein, E. Knas, M. Reid, A. Mraz Esposito, M. Manley, L. Malone, L. Tarullo, S. Lukashanets, and J. West. *Descriptive Data on Head Start Children and Families from FACES 2014: Fall 2014 Data Tables and Study Design.* OPRE Report 2017-97. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017b.

Aikens, N., E. Moiduddin, J. West, S. Atkins-Burnett, Y. Xue, and S. Lukashanets. "Examining Procedures for Assessing Dual Language Learners: Findings from FACES and UPCOS." Report submitted to the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Washington, DC: Mathematica Policy Research, 2014.

Bernstein, S., L. Malone, A. Kopack Klein, C. Bush, K. Feeney, M. Reid, S. Lukashanets, and N. Aikens. *Descriptive Data on Region XI Head Start Children and Families: AI/AN FACES Fall 2015-Spring 2016 Data Tables and Study Design.* OPRE Report 2018-26. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2018.

Bierman, K.L., R.L. Nix, M.T. Greenberg, C. Blair, and C.E. Domitrovich. "Executive Functions and School Readiness Intervention: Impact, Moderation, and Mediation in the Head Start REDI Program." *Development and Psychopathology*, vol. 20, no. 3, 2008, pp. 821-843.

Blair, C. "School Readiness: Integrating Cognition and Emotion in a Neurobiological Conceptualization of Children's Functioning at School Entry." *American Psychologist*, vol. 57, 2002, pp. 111-127.

Diamond, A., and C. Taylor. "Development of an Aspect of Executive Control: Development of the Abilities to Remember What I Said and to 'Do as I Say, Not as I Do'." *Developmental Psychobiology*, vol. 29, 1996, pp. 315-334.

Duncan, S., and E. DeAvila. *Preschool Language Assessment Survey 2000 Examiner's Manual: English Forms C and D.* Monterey, CA: CTB/McGraw-Hill, 1998.

Dunn, L.M., L.L. Dunn, and D.M. Dunn. *Peabody Picture Vocabulary Test, Fourth Edition Examiner's Manual and Norms Booklet*. Circle Pines, MN: American Guidance Service, 2006.

Entwisle, D.R., K.L. Alexander, D. Cadigan, and P.M. Pallis. "The Emergent Academic Self-Image of First Graders: Its Response to Social Structure." *Child Development*, vol. 58, 1987, pp. 1190-1206.

Fisher, P.A., and T.J. Ball. “Tribal Participatory Research: Mechanisms of a Collaborative Model.” *American Journal of Community Psychology,* vol. 32, no. 3, 2003, pp. 207-216.

Gresham, F.M., and S.N. Elliot. *Social Skills Rating System.* Circle Pines, MN: American Guidance Service, 1990.

Kopack Klein, A., L. Malone, J. West, N. Aikens, and L. Tarullo. *Data Tables for FACES 2009 Report: Head Start Family and Classroom Supports for Kindergarten Achievement.* OPRE Report 2017-70. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation, 2017.

Lynch, W., and M.J. Hanson. *Developing Cross-Cultural Competence: A Guide for Working with Young Children and Their Families*. Towson, MD: Paul H. Brookes Publishing, 1992.

Malone, L., A. Kopack Klein, S. Bernstein, B. Lepidus Carlson, S. Albanese, A. Bloomenthal, J. Baker, M. Cavanaugh, M. Reid, K. Feeney, R. Sutton-Heisey, M. Hepburn, A. Kelly, and G. Lim. “American Indian and Alaska Native Head Start Family and Child Experiences Survey 2015: User’s Manual.” Report submitted to the U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. Washington, DC: Mathematica Policy Research, 2018.

Martin, N., and R. Brownell. *Expressive One-Word Picture Vocabulary Test-4th Edition*. Novato, CA: Academic Therapy Publications, 2010.

McClelland, M.M., C.E. Cameron, R. Duncan, R.P. Bowles, A.C. Acock, A. Miao, and M.E. Pratt. “Predictors of Early Growth in Academic Achievement: The Head-Toes-Knees-Shoulders Task.” *Frontiers in Psychology*, vol. 5, 2014, 599.

Peterson, J., and N. Zill. “Marital Disruption, Parent-Child Relationships, and Behavior Problems in Children.” *Journal of Marriage and the Family*, vol. 48, 1986, pp. 295-307.

Ponitz, C.C., M.M. McClelland, J.S. Matthews, and F.J. Morrison. “A Structured Observation of Behavioral Self-Regulation and Its Contribution to Kindergarten Outcomes.” *Developmental Psychology*, vol. 45, 2009, pp. 605-619.

Rainelli, S., R.J. Bulotsky-Shearera, V.A. Fernandez, D.B. Greenfield, and M. López. ““Examining the Validity of Behavioral Self-Regulation Tools in Predicting Preschoolers’ Academic Achievement.” *Early Childhood Research Quarterly,* vol. 38, 2017, pp. 10-22.

Roid, G.H., and L.J. Miller. *Leiter International Performance Scale Revised, Examiner Rating Scale (Leiter-R).* Lutz, FL: Psychological Assessment Resources, Inc., 1997.

Schmitt, S.A., M.E. Pratt, and M.M. McClelland. “Examining the Validity of Behavioral Self-Regulation Tools in Predicting Preschoolers’ Academic Achievement.” *Early Education and Development*, vol. 25, no. 5, 2014, pp. 641-660.

Smith-Donald, R., C. Raver, T. Hayes, and B. Richardson. “Preliminary Construct and Concurrent Validity of the Preschool Self-Regulation Assessment (PSRA) for Field-Based Research.” *Early Childhood Research Quarterly*, vol. 22, 2007, pp. 173-187.

Woodcock, R., K. McGrew, and N. Mather. *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing, 2001.

Linacre, J.M., and B.D. Wright. “Mantel-Haenszel DIF and PROX Are Equivalent!” *Rasch Measurement Transactions*, vol. 3, no. 2, 1989, pp. 52-53.

Zwick, R.J. “A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement.” *ETS Research Report Series*, vol. 2012, issue 1, 2012, pp. i-30.

This page has been left blank for double-sided copying.

APPENDIX A

CORRELATIONS AMONG AI/AN FACES 2015 CHILD OUTCOME MEASURES

This page has been left blank for double-sided copying.

Table A.1. Correlations of AI/AN FACES 2015 direct child assessment measures: Spring 2016

		PPVT- 4 W (GSV) Score	EOWPVT-4 Total Raw Score	WJ Letter Word W Score	WJ Applied Problems W Score	ECLS-B math ability estimate	ECLS-B Letter ability estimate[a]	Pencil tapping[b]	Leiter Cognitive-Social Total Score	Leiter-Attention	Leiter-Organization/ impulse control	Leiter-Activity	Leiter-Sociability
PPVT- 4 W (GSV) Score	Pearson Correlation	1	.743**	.480**	.699**	.706**	.341**	.425**	.332**	.328**	.334**	.268**	.280**
	N	683	683	678	679	681	251	553	682	682	682	682	682
EOWPVT-4 Total Raw Score	Pearson Correlation	.743**	1	.469**	.642**	.677**	.340**	.373**	.299**	.294**	.296**	.260**	.247**
	N	683	687	680	683	682	251	554	686	686	686	686	686
WJ Letter Word W Score	Pearson Correlation	.480**	.469**	1	.501**	.636**	.954**	.366**	.294**	.279**	.295**	.259**	.256**
	N	678	680	680	676	677	251	552	679	679	679	679	679
WJ Applied Problems W Score	Pearson Correlation	.699**	.642**	.501**	1	.900**	.364**	.446**	.352**	.345**	.346**	.300**	.305**
	N	679	683	676	683	678	251	551	682	682	682	682	682
ECLS-B math ability estimate	Pearson Correlation	.706**	.677**	.636**	.900**	1	.501**	.532**	.334**	.327**	.324**	.292**	.284**
	N	681	682	677	678	682	251	553	681	681	681	681	681
ECLS-B Letter ability estimate[a]	Pearson Correlation	.341**	.340**	.954**	.364**	.501**	1	.229**	.152*	.150*	.147*	.111	.115
	N	251	251	251	251	251	251	224	250	250	250	250	250
Pencil Tapping[b]	Pearson Correlation	.425**	.373**	.366**	.446**	.532**	.229**	1	.254**	.252**	.231**	.271**	.184**
	N	553	554	552	551	553	224	554	554	554	554	554	554
Leiter-Cognitive-Social Total Score	Pearson Correlation	.332**	.299**	.294**	.352**	.334**	.152*	.254**	1	.973**	.968**	.902**	.867**
	N	682	686	679	682	681	250	554	686	686	686	686	686
Leiter-Attention	Pearson Correlation	.328**	.294**	.279**	.345**	.327**	.150*	.252**	.973**	1	.919**	.841**	.794**
	N	682	686	679	682	681	250	554	686	686	686	686	686
Leiter-Organization/ impulse control	Pearson Correlation	.334**	.296**	.295**	.346**	.324**	.147*	.231**	.968**	.919**	1	.834**	.815**
	N	682	686	679	682	681	250	554	686	686	686	686	686
Leiter-Activity	Pearson Correlation	.268**	.260**	.259**	.300**	.292**	.111	.271**	.902**	.841**	.834**	1	.731**
	N	682	686	679	682	681	250	554	686	686	686	686	686
Leiter-Sociability	Pearson Correlation	.280**	.247**	.256**	.305**	.284**	.115	.184**	.867**	.794**	.815**	.731**	1
	N	682	686	679	682	681	250	554	686	686	686	686	686

Source: Spring 2016 AI/AN FACES Direct Child Assessment.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

[a]This task is administered only to children who meet a certain threshold on the WJ III Letter-Word Identification subtest.

[b]This task is administered only to children age 4 and older at the time of assessment.

Table A.2. Correlations of AI/AN FACES 2015 teacher and assessor report of children's behavior: Spring 2016

		Teacher Child Report					Assessor report from child assessment			
		Approaches to learning	Social skills	Problem behaviors-aggressive	Problem behaviors-hyperactive	Problem behaviors-withdrawn	Leiter-Attention	Leiter-Organization/ impulse control	Leiter-Activity	Leiter-Sociability
Teacher Child Report										
Approaches to learning	Pearson Correlation	1	.764**	-.565**	-.635**	-.453**	.289**	.320**	.316**	.277**
	N	706	706	706	704	706	672	672	672	672
Social skills	Pearson Correlation	.764**	1	-.611**	-.584**	-.447**	.309**	.337**	.302**	.316**
	N	706	707	706	704	706	673	673	673	673
Problem behaviors-aggressive	Pearson Correlation	-.565**	-.611**	1	.664**	.406**	-.323**	-.334**	-.308**	-.287**
	N	706	706	706	704	706	672	672	672	672
Problem behaviors-hyperactive	Pearson Correlation	-.635**	-.584**	.664**	1	.530**	-.291**	-.300**	-.298**	-.267**
	N	704	704	704	704	704	670	670	670	670
Problem behaviors-withdrawn	Pearson Correlation	-.453**	-.447**	.406**	.530**	1	-.228**	-.221**	-.212**	-.208**
	N	706	706	706	704	706	672	672	672	672
Assessor report from child assessment										
Leiter-Attention	Pearson Correlation	.289**	.309**	-.323**	-.291**	-.228**	1	.919**	.841**	.794**
	N	672	673	672	670	672	686	686	686	686
Leiter-Organization/ impulse control	Pearson Correlation	.320**	.337**	-.334**	-.300**	-.221**	.919**	1	.834**	.815**
	N	672	673	672	670	672	686	686	686	686
Leiter-Activity	Pearson Correlation	.316**	.302**	-.308**	-.298**	-.212**	.841**	.834**	1	.731**
	N	672	673	672	670	672	686	686	686	686
Leiter-Sociability	Pearson Correlation	.277**	.316**	-.287**	-.267**	-.208**	.794**	.815**	.731**	1
	N	672	673	672	670	672	686	686	686	686

Source: Spring 2016 AI/AN FACES Direct Child Assessment and Teacher Child Report.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table A.3. Correlations of AI/AN FACES 2015 direct child assessment and Teacher Child Report measures: Spring 2016

		PPVT- 4 W (GSV) Score	EOWPVT-4 Total Raw Score	WJ Letter Word W Score	WJ Applied Problems W Score	WJ Spelling W Score	ECLS-B math ability estimate	ECLS-B Letter ability estimate[a]	Pencil Tapping[b]
Approaches to learning	Pearson Correlation	.385**	.332**	.372**	.392**	.380**	.414**	.251**	.307**
	N	669	673	666	669	670	668	248	541
Social skills	Pearson Correlation	.348**	.324**	.297**	.376**	.317**	.360**	.143*	.217**
	N	670	674	667	670	671	669	248	542
Problem behaviors-aggressive	Pearson Correlation	-.259**	-.193**	-.241**	-.265**	-.253**	-.239**	-.155*	-.209**
	N	669	673	666	669	670	668	248	541
Problem behaviors-hyperactive	Pearson Correlation	-.368**	-.304**	-.292**	-.354**	-.348**	-.346**	-.099	-.263**
	N	668	671	664	667	668	667	248	541
Problem behaviors-withdrawn	Pearson Correlation	-.253**	-.266**	-.201**	-.214**	-.258**	-.228**	-.110	-.164**
	N	669	673	666	669	670	668	248	541
Problem behaviors-total score	Pearson Correlation	-.374**	-.328**	-.304**	-.353**	-.355**	-.341**	-.155*	-.274**
	N	669	673	666	669	670	668	248	541

Source: Spring 2016 AI/AN FACES Direct Child Assessment and Teacher Child Report.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

[a]This task is administered only to children who meet a certain threshold on the WJ III Letter-Word Identification subtest.

[b]This task is administered only to children age 4 and older at the time of assessment.

This page has been left blank for double-sided copying.

Adaptation of the Diamond and Taylor (1996) *Peg-Tapping Executive Functioning Task.* Copyright © 1996; Blair 2002; Smith-Donald, Raver, Hayes, and Richardson, 2007.

Classroom Assessment Scoring System™ (CLASS™) by Robert C. Pianta, Karen M La Paro, and Bridget K. Hamre. Copyright © 2008 by Paul H. Brooks Publishing Co. Classroom Assessment Scoring System, CLASS, and the CLASS logo are registered trademarks of Robert C. Pianta. Used with permission of publisher.

Early Childhood Environment Rating Scale, Revised Edition. Reprinted by permission of the Publisher. From Thelma Harms, Richard M. Clifford, and Debby Cryer, Early Childhood Environmental Rating Scale--(ECERS-R) Revised Edition, New York: Teachers College Press. Copyright ©2005 by Thelma Harms, Richard M. Clifford and Debby Cryer. All Rights Reserved.

Expressive One-Word Picture Vocabulary Test-4 (EOWPVT-4). Copyright © 2011, Academic Therapy Publications, 20 Commercial Boulevard, Novato, CA, 94949-6191. All rights reserved. Reproduced by permission of Academic Therapy Publications.

Leiter International Performance Scale-Revised (Leiter-R), Examiner Rating Scale. Copyright © 1997, 2002 Stoelting Co., 620 Wheat Lane, Wood Dale, IL 60191. All rights reserved. The material may not be distributed to countries who have no copyright relations with the United States of America.

PreLAS 2000, by Sharon E. Duncan, Ph.D., and Edward A. DeAvila, Ph.D. Copyright © 1998 CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced by permission of CTB/McGraw-Hill LLC. Trademark: LAS is a registered trademark of The McGraw-Hill Companies, Inc.

Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4). Copyright © 2007, Wascana Limited Partnership. Adapted and reproduced with permission of the publisher NCS Pearson, Inc. All rights reserved.

Selected items from the *Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)*, National Center for Education Statistics. Used with permission.

Selected items from the *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K)*, National Center for Education Statistics. To include items reproduced from the *Test of Early Mathematics Ability, 3rd Ed. (TEMA-3)*, by H.P. Ginsburg, and A.J. Baroody. Copyright © 2003 by Pro Ed, Inc. Used with permission.

Social Skills Rating System (SSRS). Copyright © 1990, NCS Pearson, Inc. adapted and reproduced with permission. All rights reserved.

Woodcock-Johnson® III (WJ III®), WJ III® Tests of Achievement. Copyright © 2001, 2007 by the Riverside Publishing Company. *Woodcock-Johnson® III (WJ III®), WJ III® Tests of Achievement* reproduced with permission of the publisher. All rights reserved.