Article

Journal of Psychoeducational Assessment
1–20
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0734282919859804
journals.sagepub.com/home/jpa
SAGE

Psychometric Validation and Reorganization of the Desired Results Developmental Profile

Tutrang Nguyen[1], Stephanie M. Reich[2], Jade Marcus Jenkins[2], and Jamal Abedi[3]

Abstract

This study reports an independent investigation of the psychometric properties of Desired Results Developmental Profile (DRDP), a teacher-rated measure of school readiness for preschool-aged children. In a sample of 2,031 low-income, 3- to 5-year-old children attending Head Start, we tested three measurement models: a higher order one-factor model, a seven-factor model, and a five-factor model. To explore the appropriateness of the DRDP for use with diverse populations of young children, we used multiple group and differential item functioning (DIF) analyses to determine whether the DRDP works differently for dual language learners (DLL) and non-DLLs. The proposed five-factor structure fits the data best, with greater face and statistical validity. Using this conceptually driven factor structure, the multiple group analyses were robust for DLL and non-DLL preschool students. More than half of the items on the DRDP displayed little DIF. Items measuring emergent language and literacy exhibited DIF favoring non-DLL children.

Keywords

early childhood, school readiness, teacher-reported measure, validation

As early as when schooling begins, low-income children lag behind their higher income peers in critical early academic and socioemotional skills (Duncan & Murnane, 2011). Dual language learners (DLLs) also lag behind their native English-speaking peers on these same skills (Quirk, Nylund-Gibson, & Furlong, 2012). Comprehensive early education programs, such as Head Start, are one effort to support these at-risk children, based on a robust body of research showing that preschool programs can address disparities in the learning opportunities for young children prior to school entry (Yoshikawa et al., 2013). Public preschool programs consistently improve children's readiness for school in terms of early literacy, mathematics, and social-emotional development (Phillips et al., 2017), with low-income and DLL children benefiting the most from these programs (Duncan & Magnuson, 2013; Gormley, 2008; Magnuson, Lahaie, & Waldfogel, 2006). Over the past two decades, states and municipalities across the

[1]University of Virginia, Charlottesville, VA, USA
[2]University of California, Irvine, CA, USA
[3]University of California, Davis, CA, USA

Corresponding Author:
Tutrang Nguyen, University of Virginia, 405 Emmet Street South, Charlottesville, VA 22904, USA.
Email: tcn3rt@virginia.edu

country have developed more public educational opportunities, such as voluntary prekindergarten programs, to better serve young children from diverse backgrounds in their preparation for kindergarten.

As preschool education continues to expand nationwide, state and local efforts to monitor and improve programs are rapidly expanding as well. A key component of such efforts is documenting children's learning outcomes and assessing their readiness for school. For example, the recent Race to the Top–Early Learning Challenge initiative gave priority to applicants who focused on strengthening the use of assessments to understand individual children's progress and improve program quality (Ackerman & Coley, 2012; Congressional Research Service, 2016; Connors-Tadros, 2014). In turn, there now exists an increased demand for psychometrically sound measures of children's development and learning across multiple domains for diverse populations of 3- to 5-year-old children. Such assessment measures could yield information that not only informs ongoing decisions about teaching and children's learning but is also predictive of longitudinal academic achievement after school entry. Assessments that are valid and reliable, meet high psychometric standards, and are appropriate for their intended purpose can also inform a continuous cycle of program improvement.

Desired Results Developmental Profile (DRDP)

One broadly used assessment for promoting and assessing school readiness is the DRDP–Preschool, which is implemented statewide in California and Missouri (California Department of Education [CDE], Early Education and Support Division, 2010; Missouri Department of Elementary and Secondary Education, 2013). Preschools with state funding are required to complete this assessment 3 times a year for all children served, with checkpoints in the fall, winter, and spring. Federally funded programs, such as Head Start, have followed, adopting the DRDP throughout these states. As such, the DRDP is administered to about half a million children each year attending publicly funded preschool programs (Friedman-Krauss et al., 2018).

The DRDP was developed by the Center for Child & Family Studies at WestEd and the Berkeley Evaluation and Assessment Research Center at the University of California, Berkeley, to measure the learning and development of children aged 3 to 5 years. The DRDP comprised 43 items, which fall within seven developmental subscales. According to the developers, it is designed as a process measure for assisting early childhood educators with curriculum planning for individual children and guiding continuous program improvement (CDE, Early Education and Support Division, 2010). However, with federal assessment requirements and the widespread use of the DRDP in California and Missouri, it is frequently used as a summative assessment at several points during the school year. That means rather than utilizing the scores of individual children to tailor programming to the target child, results are often aggregated across children, centers, and agencies; compared for change over time; and reported to others such as the Office of Head Start (e.g., Improving Head Start for School Readiness Act of, 2007).

Despite the widespread use of this measure across Head Start and state-funded preschool centers in these two states, there is surprisingly limited research using this assessment and no research confirming its validity in measuring children's development, appropriateness for non-native English speakers, or its reliability. A comprehensive search for prior empirical work on the DRDP yielded one published study on the unidimensionality of the assessment using a select number of items and domains (Sutter et al., 2017) and one published study on its cross-age validity (Karelitz, Parrish, Yamada, & Wilson, 2010). Sutter et al. (2017) found that a unidimensional factor of the DRDP provided the best fit to their data from a convenience sample of 34 children. Their analysis only looked at portions of the DRDP for its factor structure with a very small sample. Specifically, they examined three (cognitive, language, and social development) of the seven domains and tested whether they would fit together as one factor of school readiness.

Karelitz et al. (2010), using a larger cross-sectional sample ($n = 751$), showed that the DRDP has valid properties as a screener for identifying relatively low- and high-achieving children from preschool to elementary school. However, they did not test the factor structure of the measure. One other published study has used the DRDP as an outcome but did not report any information on its reliability or validity (Mohler, Yun, Carter, & Kasak, 2009). In addition, we found no detailed technical documentation of the DRDP's content validity.

Another key element missing from the DRDP's psychometric evidence is validity for children who are considered DLL. This is particularly troubling for states currently implementing the DRDP, such as California and Missouri, where 45% and 8% of 3- and 4-year-old children, respectively, are DLLs (National Institute for Early Education Research, 2016). The number of young DLLs is also growing rapidly across the United States, with Spanish-speaking DLLs now representing 40% of all Head Start participants, and is the fastest growing subpopulation of students in the United States. Incorporating high-quality assessments into the education of DLL children is essential because DLL students trail their monolingual English-speaking peers in important English language skills at kindergarten entry (Hoff, 2013; Paez, Tabors, & Lopez, 2007), and these gaps in achievement persist through elementary school (Mancilla-Martinez, & Lesaux, 2011). Abedi and Gándara (2006) argue that the achievement gaps often observed between DLL and non-DLL children are in part because measurement tools are often ill-equipped to assess their skills and abilities. Indeed, most assessments are not invariant across DLL groups (Immekus & McGee, 2016; Quirk, Mayworm, Edyburn, & Furlong, 2016). And although the CDE (2018) states on their website that the DRDP "includes specific measures for assessing the English language development of children who are learning English as a second language," there exists no psychometric validation of this assessment for non-English-speaking children.

In summary, evidence of the DRDP is far too sparse relative to its widespread use among large populations of young children, children most in need of early childhood educational intervention, and the frequency with which teachers are required to use the DRDP. This is concerning given that reliable and valid preschool screening and assessment tools are key for assessing children's learning and development and providing the appropriate classroom experiences and supplemental services necessary to ensure that all children are successful at school entry (Kagan & Garcia, 2007; Shadish, Cook, & Campbell, 2001; Snow & Van Hemel, 2008). Therefore, a detailed psychometric analysis of the DRDP is urgently needed.

Current Study

Our study is an independent investigation of the psychometric properties of the DRDP. We tested the reliability and validity of the DRDP across the preschool year using two cohorts of 3- and 4-year-old children attending an urban Head Start program. Using data from 2,031 children collected in the fall, winter, and spring of 2014-2015, we (a) tested the fit of the seven developer-defined DRDP subscales as a higher order one-factor model and a seven-factor model, (b) conducted a face validity assessment of the 43 individual items and conceptually derived subscales into which the items belong as a five-factor model, (c) conducted a confirmatory factor analysis (CFA) of these new subscales, (d) tested whether the dimensionality of this conceptually driven model differed for DLL and non-DLL students with multiple group analysis, and (e) conducted differential item functioning (DIF) analysis for whether the DRDP works differently at the item level between DLL and non-DLL children. Based on the limited prior research available, we hypothesized that there will be poor fit of the DRDP on the seven developer-defined subscales and that a conceptually driven model will better fit the data. We also hypothesized that the dimensionality of the DRDP will differ between DLL and non-DLL children based on prior research on assessments of DLL students (Abedi, 2002).

Method

Study Context and Data

Our study drew from administrative data from a large, urban Head Start agency in California in 2014-2015. Our study sample includes 2,031 children, 157 teachers, and 25 centers. In the larger California context, Head Start enrolled more than 100,000 children using federal and state funding during the 2014-2015 program year (Barnett & Friedman-Krauss, 2016). The CDE requires every preschool program receiving state funding to complete the DRDP for each child enrolled (CDE, Early Education and Support Division, 2010). With federal assessment requirements and support from the CDE, this measure, in California, is typically aggregated as an outcome assessment throughout the school year within the Head Start agency and thousands of other preschool programs throughout the state. The organization of DRDP items into separate domains is based on the California Preschool Learning Foundations, which outline the key skills and knowledge a child can gain through a high-quality preschool program (CDE, 2018).

Using the DRDP, children were evaluated by their primary classroom teacher 3 times during the academic year. Once ratings were completed by the teachers, center personnel entered these data into a central information database housed at the agency, where DRDP data were linked with other child-level demographic variables (e.g., gender, race/ethnicity, age) and teacher-level demographic variables (e.g., education, experience). These data were then stripped of unique identifying information before being shared with the primary investigators for research purposes, per the requirements of the University Institutional Review Board.

Approximately half of the children in the sample were male, 8% were Asian, 2% were African American, 78% were Hispanic, and 7% were another race/ethnicity. About 50% of the sample was DLL. The average child age was 4.73 years (*SD* = 0.69). Almost the entire sample of teachers was female, 12% were Asian, 4% were African American, 64% were Hispanic, and 4% were another race/ethnicity. The demographic characteristics of our sample closely mirror those of the state, with the exception of language spoken by the teacher. The percentage of Head Start teachers in our sample who spoke another language was greater than the state average—64% compared with 76% in our sample (Barnett & Friedman-Krauss, 2016).

Measures

DRDP.

The DRDP is a teacher-reported 5-point school readiness rating scale consisting of 43 items organized into seven categories of development and school readiness: (a) Self and Social Development (12 items; for example, cooperative play with others), (b) Language and Literacy Development (10 items; for example, comprehension of age-appropriate text presented by adults), (c) English Language Development (four items; for example, understanding and responding to English literacy activities), (d) Cognitive Development (five items; for example, problem solving), (e) Mathematical Development (six items; for example, number sense of quantity and counting), (f) Physical Development (three items; for example, fine motor skills), and (g) Health (three items; for example, personal safety). The English Language Development items are only completed for children with DLL status. We provide a description of the items in each of the DRDP-derived domains in the first column of Table 1.

Each item is presented as a continuum for teachers to rate children's level of skill development, ranging from (1) *Not Yet Exploring*, (2) *Exploring*, (3) *Developing*, (4) *Building*, to (5) *Integrating*. After rating an initial level for the item on the developmental continuum, teachers can also rate the child as "emerging" if the child is beginning to show some skills from the next level. The "emerging" level is considered a half point on the measure in that children may show behaviors or skills associated with the next developmental level, but does not demonstrate those

Table 1. DRDP Domain Items From the Original Seven-Factor Structure and Proposed Domains and Items for the Five-Factor Structure.

Original DRDP with seven-factor structure		Proposed domains and items with five-factor structure	
Domain	Items	Domain	Items
Self and social development	1. Identity of self	Self-awareness, identity	1. Identity of self
	2. Recognition of own skills and accomplishments		2. Recognition of own skills and accomplishments
	3. Expressions of empathy		6. Awareness of diversity in self and others
	4. Impulse control		15. Expression of self through language
	5. Taking turns	Social skills	3. Expressions of empathy
	6. Awareness of diversity in self and others		5. Taking turns
	7. Relationships with adults		7. Relationships with adults
	8. Cooperative play with peers		8. Cooperative play with peers
	9. Sociodramatic play		9. Sociodramatic play
	10. Friendships with peers		10. Friendships with peers
	11. Conflict negotiation		11. Conflict negotiation
	12. Shared use of space and materials		12. Shared use of space and materials
Language and literacy development	13. Comprehension of meaning	Language and literacy	16. Language in conversation
	14. Following increasingly complex instructions		19. Concepts about print
	15. Expression of self through language		20. Phonological awareness
	16. Language in conversation		21. Letter and word knowledge
	17. Interest in literacy		22. Emergent writing
	18. Comprehension of age-appropriate text presented by adults	Domain-general cognitive skills	4. Impulse control
	19. Concepts about print		14. Following increasingly complex instructions
	20. Phonological awareness		28. Problem solving
	21. Letter and word knowledge		29. Memory and knowledge
	22. Emergent writing		30. Curiosity and initiative
English language development	23. Comprehension of English (receptive English)		31. Engagement and persistence
	24. Self-expression in English (expressive English)	Math	27. Cause and effect
	25. Understanding and response to English literacy activities		32. Number sense of quantity and counting
	26. Symbol, letter, and print knowledge in English		33. Number sense of mathematical operations
Cognitive development	27. Cause and effect		34. Classification
	28. Problem solving		35. Measurement
	29. Memory and knowledge		36. Shapes
	30. Curiosity and initiative		37. Patterning
	31. Engagement and persistence		

(continued)

Table 1. (continued)

Original DRDP with seven-factor structure		Proposed domains and items with five-factor structure	
Domain	Items	Domain	Items
Mathematical development	32. Number sense and quantity and counting		
	33. Number sense of mathematical operations		
	34. Classification		
	35. Measurement		
	36. Shapes		
	37. Patterning		
Physical development	38. Gross motor movement		
	39. Balance		
	40. Fine motor skills		
Health	41. Personal care routines		
	42. Healthy lifestyle		
	43. Personal safety		

Note. Items dropped: 13, 17, 18, 23, 24, 25, 26, 38, 39, 40, 41, 42, 43. Items 23, 24, 25, and 26 in the English language development domain were only completed for children with DLL status. DRDP = Desired Results Developmental Profile; DLL = dual language learners.

behaviors or skills typically or consistently. Thus, the possible scores are 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5. However, the sample of teachers in our study did not utilize this half point score for any of their ratings of children. Skill points are then averaged across all items within each category to create domain (i.e., subscale) scores and across all domains to create a total score. Teachers use the domains, items, and skill points to classify and rate their observations of children's school readiness.

Child and teacher characteristics. Demographic information about children and teachers was collected from the administratively linked data. Child-level variables are age, gender, race/ethnicity, and their DLL status. Teacher-level variables include gender, race/ethnicity, highest degree earned, and whether they spoke another language in addition to English. Descriptive statistics for the sample are presented in Table 2.

Analytic Plan

We conduct analyses for each of the three measurement time points separately. Descriptive statistics for the analysis sample, including means, standard deviations, and correlations, and missing data were examined using Stata 14 (StataCorp, 2015). Missingness occurred on the DRDP measure (0%-15% across all three time points) and on the child and teacher demographic variables (0%-5%). Most of the missingness occurred on teacher education level (18%). Logistic regressions indicated that children and teachers' observable baseline characteristics were not predictive of missingness on the DRDP measure. All models were run using a maximum likelihood estimator, which estimates parameters by maximizing the likelihood of obtaining the observed values (Brown, 2006) and also addresses missing data. Specifically, with this method, all available data were allowed to be included in the analyses, and the parameters with the highest possibility of generating the sample data are identified (Baraldi & Enders, 2010).

Table 2. Descriptive Statistics of Child and Teacher Characteristics.

	M (*SD*)
Child characteristics	
Dual language learner	0.52
Age (years)	4.73 (0.75)
Female	0.50
Race/ethnicity	
Asian	0.08
Black/African American	0.01
Hispanic	0.82
Other	0.07
Observations (child)	2,031
Teacher characteristics	
Female	0.99
Speaks another language	0.76
Asian	0.12
Black/African American	0.04
Hispanic/Latino	0.64
Other	0.04
Highest degree earned	
Associate	0.35
Bachelor's	0.59
High school diploma	0.03
Master's	0.02
Observations (teacher)	157
Observations (centers)	25

Testing the fit of the developer-defined factor structure. First, we performed a CFA of the DRDP domains specified by the measure's authors using cross-sectional data from all three checkpoints (fall, winter, and spring). We tested two factor structures—a unidimensional model and the original seven domains (hereafter referred to as a seven-factor structure model)—using all of the items from the DRDP. Our study sample well exceeded the minimum sample size guideline of 400 to ensure stable correlations and a probable factor structure (Gorsuch, 2003). Model fit was evaluated based on several global goodness-of-fit indices: the comparative fit index (CFI), the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), and the Tucker–Lewis Index (TLI). Although we report the chi-square test, we did not include this in our analytic decisions because of its sensitivity to large sample sizes (Brown, 2006). We follow the goodness-of-fit recommendations made by Hu and Bentler (1999), with good fit characterized by CFI $>.95$, RMSEA $<.06$, SRMR $<.08$, and TLI $>.95$.

Face validity assessment. To conceptually group all 43 items independently from the DRDP structure presented by the publishers, we then conducted a Q-sort (Brown, 1993). Twenty-two raters, all doctoral students and faculty in Education and Developmental Psychology, were asked to sort the 43 items according to what they believed belong together. They also provided a label for the groupings of items. We then used the constructs and items from the Q-sort exercise to assist us in conceptually deriving our own categories to test the factor structure with CFA. The first three authors examined all of the Q-sort responses and met periodically to conceptually derive these five categories until consensus was reached. All conflicts were discussed and resolved among the authors.

CFA of conceptually derived subscales. To confirm the conceptually driven structure derived from the Q-sort, we conducted a CFA with the items and their newly assigned constructs. We investigated the goodness of fit of the different combinations of categories and constructs from the Q-sort and assessed the models with the same fit indices mentioned above. These factor descriptions are presented in the second column of Table 1.

Multiple group analysis for DLL and non-DLL students. Informed by these results, multiple group analyses were then used to investigate measurement invariance of the preferred model for non-DLL and DLL children. A set of steps—from least restrictive to most restrictive—was considered in determining the best model fit (Vandenberg & Lance, 2000): (a) same form, (b) equal loadings, (c) equal loadings and errors, and (d) equal loadings, errors, and variances. We examined the change in CFI values of .01 or greater to indicate a significant difference in model fit for testing measurement invariance (Cheung & Rensvold, 2002) because the chi-square difference test is extremely sensitive to large sample sizes such as ours.

DIF. Finally, to determine whether the DRDP worked differently at the item level between DLL (focal group) and non-DLL children (reference group), we conducted DIF analysis with the data separately for each time point and for each subscale. The total score for each domain was used as the estimate of ability. We employed the commonly used method of ordinal logistic regression to examine both uniform and nonuniform DIF. Uniform DIF is detected when the item favors one group over another across all levels of development being measured. For example, non-DLL children may be systematically rated as higher on an item than DLL children, regardless of the overall score. Nonuniform DIF is detected when there is a significant group-by-ability interaction, suggesting that the probability of being rated higher on an item is not the same across ability levels for the two groups (Zumbo, 1999). The logistic regression method involves a series of nested models, where each item is regressed first onto the ability variable alone (Model 1), then onto the grouping variable in addition to the ability variable (Model 2), and then onto the interaction term of the ability variable by grouping variable in addition to their main effects (Model 3). DIF is detected when there is a significant difference in fit between Model 1 and Model 3, suggesting that group membership influences item-level ratings in addition to ability level. The type of DIF, if present, is determined by testing the difference in fit between Models 1 and 2 for uniform DIF and Models 1 and 3 for nonuniform DIF.

We determined differences in model fit using the chi-square difference test. Because of the tendency for the chi-square significance test to overidentify DIF items in large samples even if the effects are negligible, we attempted to reduce type I error by also examining the magnitude of the effect size quantified with the pseudo R^2 statistic (Gelin & Zumbo, 2003; Zumbo, 1999). Thus, an item was classified as exhibiting nontrivial DIF if there was a significant chi-square difference test between Models 1 and 3 and if there was a change in R^2 from Models 1 to 3 of .035 or greater, which represents at least a moderate effect (Jodoin & Gierl, 2001).

Results

Descriptive Analyses

Prior to conducting substantive analyses, descriptive statistics on the DRDP were computed for the analytic sample. We report the descriptive statistics for our preferred five-factor model, which we discuss in greater detail below. On average, children were rated as either “2: exploring” or “3: developing” on the items across all three assessment time points. Specifically, the means of the items ranged from 1.66 to 2.55 points (SD = 0.66-1.11) for fall, 2.19 to 3.11 points (SD = 0.73-1.00) for

Table 3. Descriptive Statistics of Items in the Proposed Five-Factor Model.

	M	*SD*	Minimum	Maximum
Self-awareness, identity				
1. Identity of self	2.89	0.84	1	5
2. Recognition of own skills and accomplishments	2.83	0.80	1	5
6. Awareness of diversity in self and others	2.69	0.78	1	5
15. Expression of self through language	2.87	0.88	1	5
Social skills				
3. Expressions of empathy	2.87	0.89	1	4
5. Taking turns	2.82	0.79	1	5
7. Relationships with adults	2.84	0.81	1	5
8. Cooperative play with peers	2.94	0.75	1	5
9. Sociodramatic play	2.99	0.80	1	5
10. Friendships with peers	2.99	0.85	1	5
11. Conflict negotiation	2.60	0.82	1	5
12. Shared use of space and materials	3.11	0.83	1	5
Language and literacy				
16. Language in conversation	2.87	0.90	1	5
19. Concepts about print	2.62	0.90	1	5
20. Phonological awareness	2.19	0.79	1	5
21. Letter and word knowledge	2.37	0.91	1	5
22. Emergent writing	2.71	0.95	1	5
Domain-general cognitive skills				
4. Impulse control	2.81	0.81	1	5
14. Following increasingly complex instructions	2.90	0.84	1	5
28. Problem solving	2.85	0.82	1	5
29. Memory and knowledge	2.87	0.85	1	5
30. Curiosity and initiative	2.88	0.83	1	5
31. Engagement and persistence	2.83	0.75	1	5
Math				
27. Cause and effect	2.90	0.82	1	5
32. Number sense of quantity and counting	3.05	1.00	1	5
33. Number sense of mathematical operations	2.60	0.94	1	5
34. Classification	2.88	0.73	1	5
35. Measurement	2.67	0.89	1	5
36. Shapes	2.63	0.86	1	5
37. Patterning	2.56	0.87	1	5
Observations (child)	2,031			

Note. Descriptive statistics of items for the proposed five-factor model for the fall and spring time points are presented in the supplemental materials.

winter, and 2.64 to 3.43 points (*SD* = 0.74-0.96) for spring. The pattern of scores indicates that teachers rated children higher on the items as the school year progressed. Within each of the proposed domains, pairwise correlations between items were moderate to high, ranging from .35 to .75 for fall, .37 to .76 for winter, and .39 to .76 for spring. Table 3 displays the means and standard deviations of the DRDP items grouped by our reorganization of the domains for the winter time point. Correlations of all the items for the winter time point are presented in Table 4. Complete descriptive statistics and correlations for the fall and spring assessment time points are presented in the supplemental materials.

Table 4. Correlations of Items Included in the Proposed Five-Factor Model for the Winter Time Point.

	1	2	6	15	3	5	7	8	9	10	11	12	16	19	20	21	22	4	14	28	29	30	31	27	32	33	34	35	36	37
Self-awareness, identity																														
1. Identity of self	—																													
2. Recognition of own skills and accomplishments	.65	—																												
6. Awareness of diversity in self and others	.59	.60	—																											
15. Expression of self through language	.60	.60	.55	—																										
Social skills																														
3. Expressions of empathy	.56	.56	.55	.52	—																									
5. Taking turns	.50	.55	.53	.49	.52	—																								
7. Relationships with adults	.58	.59	.56	.58	.53	.53	—																							
8. Cooperative play with peers	.54	.56	.54	.56	.50	.59	.57	—																						
9. Sociodramatic play	.54	.55	.53	.56	.52	.53	.53	.65	—																					
10. Friendships with peers	.55	.55	.52	.56	.51	.56	.59	.68	.64	—																				
11. Conflict negotiation	.54	.54	.56	.53	.50	.59	.54	.53	.51	.54	—																			
12. Shared use of space and materials	.43	.46	.44	.46	.47	.57	.47	.54	.51	.56	.55	—																		
Language and literacy																														
16. Language in conversation	.58	.61	.58	.76	.52	.50	.56	.57	.56	.56	.53	.46	—																	
19. Concepts about print	.53	.55	.54	.57	.53	.53	.53	.53	.52	.53	.51	.47	.58	—																
20. Phonological awareness	.45	.46	.47	.49	.42	.45	.41	.45	.41	.43	.46	.37	.50	.55	—															
21. Letter and word knowledge	.45	.46	.46	.46	.40	.45	.44	.41	.39	.41	.46	.40	.47	.55	.54	—														
22. Emergent writing	.48	.50	.49	.50	.42	.51	.47	.46	.46	.50	.49	.45	.51	.56	.48	.60	—													
Domain-general cognitive skills																														
4. Impulse control	.48	.49	.50	.45	.49	.66	.50	.51	.50	.52	.61	.55	.48	.48	.41	.43	.49	—												
14. Following increasingly complex instructions	.53	.53	.53	.53	.48	.53	.50	.50	.49	.52	.54	.45	.56	.56	.50	.45	.52	.53	—											
28. Problem solving	.50	.52	.53	.52	.48	.52	.51	.52	.48	.52	.53	.50	.54	.56	.45	.47	.51	.50	.48	—										
29. Memory and knowledge	.55	.55	.54	.64	.49	.51	.56	.53	.55	.55	.51	.50	.63	.56	.45	.45	.49	.46	.53	.53	—									
30. Curiosity and initiative	.54	.53	.53	.53	.48	.47	.53	.53	.50	.52	.49	.45	.53	.54	.45	.44	.48	.45	.47	.58	.56	—								
31. Engagement and persistence	.51	.54	.53	.52	.47	.54	.53	.53	.50	.54	.52	.50	.51	.54	.49	.48	.52	.50	.52	.57	.54	.59	—							
Math																														
27. Cause and effect	.55	.54	.56	.58	.50	.54	.55	.56	.52	.56	.54	.52	.59	.58	.50	.47	.54	.48	.53	.61	.57	.56	.54	—						
32. Number sense of quantity and counting	.44	.44	.43	.48	.42	.44	.41	.43	.41	.43	.40	.41	.48	.49	.41	.51	.51	.39	.44	.45	.47	.47	.47	.48	—					
33. Number sense of mathematical operations	.48	.49	.52	.53	.47	.50	.46	.46	.43	.46	.50	.44	.53	.54	.49	.57	.55	.47	.50	.53	.52	.51	.56	.55	.60	—				
34. Classification	.51	.51	.54	.52	.48	.47	.50	.49	.47	.48	.46	.46	.50	.52	.47	.49	.50	.43	.54	.50	.52	.52	.57	.53	.51	.56	—			
35. Measurement	.49	.49	.51	.51	.47	.46	.48	.48	.48	.49	.46	.42	.51	.53	.47	.47	.49	.45	.52	.51	.53	.52	.52	.52	.47	.57	.59	—		
36. Shapes	.53	.53	.52	.54	.48	.49	.50	.47	.46	.48	.51	.41	.54	.54	.52	.61	.55	.45	.51	.52	.54	.49	.51	.52	.54	.59	.55	.53	—	
37. Patterning	.46	.47	.52	.47	.46	.46	.44	.43	.42	.43	.46	.40	.48	.52	.51	.52	.55	.45	.51	.50	.47	.46	.52	.51	.49	.60	.58	.54	.59	—

Table 5. Confirmatory Factor Analysis Model Fit Statistics.

Time point	Structure	CFI	RMSEA	SRMR	TLI	Standardized factor loadings	Cronbach's α
Fall	Unidimensional	.89	.10	.09	.90	0.69-0.82	.83
	Seven-factor	.94	.09	.08	.91	0.63-0.76	.85
	Five-factor	.99	.05	.01	.99	0.74-0.81	.88
Winter	Unidimensional	.90	.10	.09	.89	0.66-0.78	.81
	Seven-factor	.92	.09	.09	.93	0.64-0.75	.86
	Five-factor	.98	.07	.02	.97	0.66-0.75	.91
Spring	Unidimensional	.90	.09	.09	.90	0.65-0.79	.82
	Seven-factor	.92	.08	.09	.94	0.60-0.79	.81
	Five-factor	.99	.07	.01	.98	0.68-0.77	.91

Note. The seven-factor model is the developer-defined model. We follow the goodness-of-fit recommendations made by Hu and Bentler (1999), with good fit characterized by CFI >.95, RMSEA <.06, SRMR <.08, and TLI >.95. All factor loadings are statistically significant, $p < .001$. Cells for the standardized factor loadings are ranges (minimum and maximum values). CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; TLI = Tucker–Lewis index.

Substantive Analyses

Factor structure of the DRDP. We first examined the unidimensional and seven-factor model that corresponds to the designed structure of the DRDP. The unidimensional model fit the data poorly in the fall (CFI = .89, RMSEA = .10, SRMR = .09, and TLI = .90), winter (CFI = .90, RMSEA = .10, SRMR = .09, and TLI = .89), and spring (CFI = .90, RMSEA = .09, SRMR = .09, and TLI = .90). In addition, correlations between the seven domains were moderate (0.39-0.65). The seven-factor model fit the data poorly, as evidenced by CFI = .94, RMSEA = .09, SRMR = .08, and TLI = .91 in the fall; CFI = .92, RMSEA = .09, SRMR = .09, and TLI = .93 in the winter; and CFI = .92, RMSEA = .08, SRMR = .09, and TLI = .94 in the spring.

Face validity assessment and CFA. We operationalized the 39 rating scale items into five domains based on our conceptually driven Q-sort exercise: self-awareness and identity (four items), mathematics (seven items), social skills (eight items), language and literacy (five items), and domain-general cognitive skills (six items). We confirmed this five-factor model using CFA. This model fit the data reasonably well for fall (CFI = .99, RMSEA = .05, SRMR = .01, and TLI = .99), winter (CFI = .98, RMSEA = .07, SRMR = .02, and TLI = .97), and spring (CFI = .99, RMSEA = .07, SRMR = .01, and TLI = .98) time points. All factor loadings were generally large (βs = 0.66-0.81) and statistically significant at $p < .001$. We considered this five-factor model to be our preferred model. The fit statistics for the unidimensional, five-factor, and seven-factor models are presented in Table 5.

With regard to the specific changes made to the five-factor model, we dropped items related to children's physical development (three items), which are listed at the bottom of the first column of Table 1. Modification indices were also consulted in our analyses, but they did not improve the model fit. Although the items in this domain were grouped together in our Q-sort exercise, the likelihood ratio test on the difference between the five-factor model and the six-factor model in our series of CFAs suggested that the five-factor model was better and more parsimonious. In addition, we dropped all items in the English language development domain because teachers were instructed to only complete these items for children with DLL status. That is, all children considered non-DLL were missing these four items, preventing us from conducting tests of measurement invariance and making group comparisons between DLL and non-DLL children. Finally, we dropped three low loading items (<0.40; Items 13, 15, and 18)

Table 6. Tests of Measurement Invariance for the Proposed Five-Factor Model.

Time point	Model	CFI	RMSEA	SRMR	TLI	Change in CFI
Fall	Model 1: Same form	.98	.07	.03	.97	—
	Model 2: Equal loadings	.98	.07	.03	.97	.00
	Model 3: Equal loadings and errors	.97	.06	.04	.98	.01
	Model 4: Equal loadings, errors, and variances	.96	.06	.03	.99	.02
Winter	Model 1: Same form	.99	.04	.02	.96	—
	Model 2: Equal loadings	.99	.05	.04	.96	.00
	Model 3: Equal loadings and errors	.98	.05	.03	.97	.01
	Model 4: Equal loadings, errors, and variances	.97	.06	.04	.98	.02
Spring	Model 1: Same form	.97	.08	.03	.95	—
	Model 2: Equal loadings	.97	.07	.03	.96	.01
	Model 3: Equal loadings and errors	.97	.07	.03	.96	.01
	Model 4: Equal loadings, errors, and variances	.96	.08	.03	.98	.03

Note. We follow the goodness-of-fit recommendations made by Hu and Bentler (1999), with good fit characterized by CFI >.95, RMSEA <.06, SRMR <.08, and TLI >.95. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; TLI = Tucker–Lewis index.

in the language and literacy domain to improve model fit, following the standard recommendation (Floyd & Widaman, 1995; Kline, 2005), even though these items were originally in our proposed five-factor model.

Multiple group analysis. Based on the data fit of the proposed five-factor model, we proceeded to assess the measurement invariance for DLL and non-DLL students. We first fit an unrestricted baseline CFA model to allow factor loadings, factor variances, covariances, and means to be freely estimated across groups. As shown in Table 6, the fit statistics for the measurement invariance CFA indicate that the CFI, RMSEA, SRMR, and TLI values showed excellent fit of these data, suggesting that the factorial pattern of the DRDP was similar across groups. After establishing the baseline model, we constrained the factor loadings to be equal across time. The change in CFI between Model 1 and Model 2 was negligible, and the fit values remained acceptable, suggesting that the factor loadings for the items were invariant across time. The next step was to constrain the intercepts to be equal across groups to evaluate scalar invariance. The CFI difference between Model 1 and Model 3 was the same as the cutoff criterion of 0.01, and the fit statistics were still acceptable, indicating that the intercepts for the items were invariant across groups. Finally, we constrained the residual variances among the items to be equal across groups. Although the change in the CFI value between Model 1 and Model 4 did not attain the desired cutoff value of 0.01, this most restrictive model did not indicate a significant reduction in fit compared with a less restricted model where the data fit the model reasonably well.

DIF results for DLLs and non-DLLs. Table 7 displays the results for the DIF analyses between DLL and non-DLL children. Most items indicated negligible DIF across the three time points: 77% in the fall and 87% in both the winter and the spring. The majority of items that exhibited DIF came from the language and literacy domain. For the fall assessment data, four items displayed intermediate DIF (Items 14 and 19 favored non-DLLs; Items 5 and 29 favored DLLs) and two items displayed large DIF favoring non-DLLs (Items 20 and 22). In the winter assessment data, three items displayed intermediate DIF (Items 14 and 20 favored non-DLLs; Item 29 favored DLLs) and one item displayed large DIF in favor of non-DLLs (Item 22). By the spring assessment, one item exhibited intermediate DIF (Items 19) and two items exhibited large DIF (Items 20 and 22), both in favor of non-DLLs. Looking across all three assessment time points, Items 19, 20, and 22

Table 7. Results of DIF Analyses Between DLL and Non-DLL Students.

		Number of items exhibiting intermediate DIF		Number of items exhibiting large DIF		Uniform DIF		Nonuniform DIF	
Time point	Number of items exhibiting negligible DIF	Favor DLL	Favor non-DLL	Favor DLL	Favor non-DLL	Chi-square difference	Change in R^2	Chi-square difference	Change in R^2
Fall	24	2	2	0	2	1.28-12.36	.001-.037	0.00-3.58	.000-.033
Winter	26	1	2	0	1	0.71-10.83	.001-.038	0.00-2.81	.000-.029
Spring	27	0	1	0	2	0.54-9.48	.001-.037	0.01-2.38	.000-.016

Note.. We follow the goodness-of-fit recommendations made by Hu and Bentler (1999), with good fit characterized by CFI >.95, RMSEA <.06, SRMR <.08, and TLI >.95. Cells for the chi-square difference and change in R^2 are ranges (minimum and maximum values). DIF = differential item functioning; DLL = dual language learners; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; TLI = Tucker–Lewis Index.

measuring children's concepts about print, emergent writing, and phonological awareness exhibited either intermediate or large DIF. All the DIF items showed uniform DIF as the chi-square differences and R^2 effect size were statistically significant, displayed in the last four columns of Table 4. Therefore, not all of the items function equivalently for DLL and non-DLL children.

Discussion

Increasingly, programs, districts, and states are requiring assessments of children's school readiness prior to kindergarten entry. The DRDP is a measure of school readiness currently used in Head Start centers and state-funded preschools as an assessment tool for kindergarten readiness. California and Missouri have sought to standardize the assessment process by mandating the use of the DRDP—an instrument initially designed to help educators understand student progress and individualize instruction—for all preschool programs receiving state funding. However, until now, little has been known about the psychometric properties of this highly used instrument. Drawing on administrative data from one Head Start agency in California, our study focused on the reliability and validity of the DRDP and explored its appropriateness for use with diverse populations of young children. To the best of our knowledge, this study is the first to establish psychometric information on the DRDP. Below, we summarize our findings and discuss the implications and limitations of this study.

We first demonstrated that the purported seven-factor structure of the measure was not supported by data. In addition, we tested a higher order unidimensional model and found that it was also not a better fitting model. Next, we theoretically derived a five-factor model through a Q-sort exercise and an understanding of the constructs gleaned from the literature. We conducted a CFA and showed that our proposed five-factor structure is a better fit to the data, having greater face and statistical validity (Shadish et al., 2001). We then conducted multiple group analysis and verified that this factor structure was robust when used with DLL and non-DLL preschool students. Finally, measurement equivalence then was also evaluated by examining DIF for the same children at the three time points (fall, winter, spring). Encouragingly, more than half of the items on the DRDP displayed little DIF, with the number of items exhibiting large DIF ranging from zero to two. Items measuring a child's language and literacy development tended to display DIF favoring non-DLL children. However, only a few items related to language were consistently identified as having DIF across all three time points, and some items displaying DIF at one time point did not indicate DIF at the other time points. No distinct pattern emerged to explain why particular items were only problematic at certain time points. In general, it is encouraging to see that DIF properties did not vary substantially over time, suggesting that the DRDP, when reorganized conceptually, works well throughout the course of the school year, even early in the year when teachers are less familiar with the children and their abilities.

The interpretation of the results of items being identified as having highly significant DIF ("C" DIF) is complex and multidimensional in nature. Across the three measurement time points of the DRDP, we found that Items 19 ("concepts about print"), 20 ("phonological awareness"), and 22 ("emergent writing") consistently displayed DIF between DLLs and non-DLLs. In terms of children's print concepts (e.g., functions of print, concept of letter and word, directionality of print), this DIF might be due to teachers' perceptions of children having little prior experience with print, leading to biased ratings as a source of DIF. A number of studies have documented that non-English-speaking families have different print-related practices in the home (Dixon, Zhao, Quiroz, & Shin, 2012; Schick & Melzi, 2016). For example, Reese and colleagues (Reese, Arauz, & Bazán, 2012; Reese & Gallimore, 2000; Reese & Goldenberg, 2008) showed that Latinx families often focus on environmental print, such as words and letters on food labels and signs on the street. This suggests that children might have to adjust to more academically based print-related activities once they are in the classroom setting. In addition, research in bilingualism indicates that the transfer of phonological skills and emergent writing occur when children have developed some proficiency in both languages (Cummins, 1991; Gillanders, Franco, Seidel, Castro, & Méndez, 2017; López, 2012; López & Greenfield, 2004; Quiroga, Lemos-Britton, Mostafapour, Abbott, & Berninger, 2002). It is possible that teachers might misinterpret this phenomenon as children having poor phonological and writing skills rather than as a common development of second language acquisition. We understand that it is imperative to examine DIF items very carefully by a group of experts including experts in the focal construct, assessment of English learners, and multicultural experts to identify the main causes of such differences between the focal and the reference groups. Although we were not able to explore this further in our study, this area should be a priority for future research.

There are a number of differences between the purported seven-factor model and our final five-factor model that should be noted. One notable difference between the two models is that we dropped items in the physical development and health domains. These domains are certainly important for children's development (Grissmer, Grimm, Aiyer, Murrah, & Steele, 2010), and this is reflected in the fact that a number of other teacher-reported performance-based assessments also include the physical development domain in their measure, such as the Child Observation Record (COR; High/Scope Educational Research Foundation, 1992), Teaching Strategies GOLD (TS GOLD; Heroman, Burts, Berke, & Bickart, 2010), and the Work Sampling System (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994). In our series of analyses, we originally had a six-factor model that included the physical domain but found that by dropping this domain, we had a more parsimonious model to describe the factor structure of the DRDP. It might also make sense to drop these items because there is a concern in the literature that teachers' perceptions of children's physical development, including their fine and gross motor skills, are also influenced by other factors salient to teacher, such as their ability to sit in their seat or pay attention (Cameron et al., 2012a, 2012b). In fact, some studies have found weak to moderate correlations between teacher reports of children's physical development and their directly assessed motor skills (Lalor, Brown, & Murdolo, 2016; Soderberg et al., 2013). We also dropped the DRDP health domain items because they were not indicators of discrete skill mastery. As the DRDP is scored to reflect the beginning stages of skill acquisition up to full mastery, skills such as personal care, healthy routines, and personal safety did not seem to conceptually fit those scoring procedures. Commonly used teacher-reported measures (e.g., COR, TS GOLD, WSS) also do not include items related to these health domains, so this raises our confidence in the conceptual underpinnings of the final five-factor model.

Another difference between the two models is that we dropped some of the language and literacy items in the final five-factor model. Empirically, Items 13 ("comprehension of meaning"), 15 ("expression of self through language"), and 18 ("comprehension of age-appropriate text presented by adults") were dropped because they had low loadings. As part of the CFA technique, we worked with the domains we proposed to adjust the five-factor structure and improve model fit. These literacy and language items were in our original conceptualization of the language and literacy domain, but removing them improved our model fit. Direct measures of these skills focusing on comprehension of text, word meaning, and self-expression have been shown to have cultural variation (Dixon et al., 2012; Harris & Schroeder, 2013), and measures that do not consider this variation perform differentially with different ethnic and racial groups (Argulewicz & Abel, 1984; Fernandez, Pearson, Umbel, Oller, & Molinet-Molina, 1992; Lee Webb, Cohen, & Schwanenflugel, 2008). Although the DRDP is based on teachers' perceptions of children's learning and development, it is worthwhile to consider whether there may be cultural bias in teacher ratings. These ratings might be influenced by subjective biases in the ways they observe children's skills (Engelhard, 2002). To be sure, the skills that these three items represent have been shown to be important for children's later literacy and language achievement (Lonigan, Allan, & Lerner, 2011; Sénéchal, Ouellette, & Rodney, 2006), and future studies on the DRDP should consider all of these items in their psychometric evaluation among diverse preschoolers. Although dropping these items improved the model fit and internal consistency of scores, it reduces the degree to which items might provide adequate coverage of children's language and literacy skills. Conceptually, we moved Item 14 ("following increasingly complex instructions") to the domain-general cognitive skills category because this is typically associated with working memory in the executive function literature (e.g., Best & Miller, 2010; Gioia & Isquith, 2004; Klingberg, 2010). Finally, we made the decision to not include Item 17 ("interest in literacy") in our final model because prior work has suggested that it is difficult to accurately capture children's interest in literacy activities with teacher reports (Baroody & Diamond, 2013). It might also be that this item represents a component of children's academic motivation (Oldfather & Wigfield, 1996; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006), which would be salient in a teacher's perceptions of a child's interest in literacy.

Implications and Limitations

Our results suggest that the DRDP has some promise as an assessment measure of school readiness for use with children of differing language backgrounds, but not in the structure proposed by WestEd and the CDE. We were especially interested in examining measurement invariance for DLL children because of their increasing presence in early childhood programs, the associated challenges of fair and accurate assessment, and the research documenting the achievement gaps between DLL and non-DLL children (National Center for Education Statistics, 2011; Reardon & Galindo, 2009). The children in our sample were all low-income, by definition of attending a Head Start program, and were also majority Hispanic and DLL. Although this represents an important demographic profile for early childhood educators, researchers, and policy makers, this feature greatly limits the extent to which our study is externally valid for other states and preschool programs using the DRDP. Future research should extend our analytic approach to other populations being assessed with the DRDP to develop a comprehensive evidence base for its validity and generalizability. Given the recent interest examining the validity and reliability of teacher-rated assessments in publicly funded preschool programs (e.g., Miller-Bains, Russo, Williford, DeCoster, & Cottone, 2017; Russo, Williford, Markowitz, Vitiello, & Bassok, 2019; Wakabayashi, Claxton, & Smith, 2019), it is important to emphasize that replicating our five-factor DRDP model with other diverse samples, settings, and policy contexts is a critical next step for this work. This replication would help us understand and improve the existing DRDP measure in terms of its psychometric properties so that it can be better utilized in large-scale implementation and as a tool for improving the quality of children's early learning experiences.

Another limitation of this study is that we were not able to examine the concurrent or discriminant validity of the DRDP by comparing it with other validated measures, such as the Woodcock-Johnson Tests of Achievement (Mather, McGrew, & Woodcock, 2001). Thus, we are unable to assess whether these results might reflect substantial between-teacher differences in the way that the DRDP is used in the classroom or whether the DRDP is a valid representation of certain skills. We were also not able to assess the fidelity with which the DRDP was used or how teachers apply the information collected from the instrument. However, our assessment of the DRDP's psychometric properties is within business-as-usual preschool practice, which is most relevant for applied research and state assessment policy guidance. It is important to note that more psychometrically rigorous measures, such as the Woodcock-Johnson Tests of Achievement, typically involve direct assessment of children individually, which is often not feasible in group-care settings with mandated ratios of teachers to children. For the DRDP, teachers make notes about children's performance on DRDP items during care time, but the scoring typically occurs when children are not present, making the DRDP more feasible to utilize for providers. However, despite the popularity of teacher-reported measures for assessing children's school readiness, the DRDP scores may be driven by assessor variance (Waterman, McDermott, Fantuzzo, & Gadsden, 2012). That is, the variability in children's DRDP scores is likely to be more attributable to the teachers who completed the measure rather than to the children themselves. Interestingly, Waterman et al. (2012) found about 28% of the variation from a teacher-reported measure was attributed to teachers, and not children. This issue of shared method variance is reflected in a number of other studies that rely on teacher-reported measures, and we encourage future investigations to make efforts to measure children's school readiness with a variety of methods if feasible.

Given the DRDP's widespread use, it is key that the evidence of its reliability and validity be developed with samples representative of children for whom it is administered and to understand whether it is appropriate for use in diverse populations. This study and our proposed reorganization of the subscale of the DRDP can help teachers and administrators better assess and forecast children's school readiness using these five domains. Evaluation of its psychometric properties, as well as a clear understanding of how teachers use the measure to support children's learning and development, should continue as long as the DRDP is in use. We hope that future research will, however, be able to make better use of this measure based on our proposed five-factor structure. With newer modifications of the DRDP to include eight domains, additional assessments of its factor structure will be needed, and this study can guide such an endeavor. Almost half a million children attending state-funded preschool in California and Missouri each year are being assessed with the DRDP (Friedman-Krauss et al., 2018), yet those data are not being used to inform educational research and there is little evidence of their use to shape educational practice. We hope this study provides insight into how to analyze and interpret large-scale samples of the DRDP more productively.

Acknowledgments

The authors are grateful to the many teachers, administrators, and staff who made this study and the larger data collection effort possible.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the Haynes Foundation.

Tutrang Nguyen is supported by the Institute of Education Sciences (IES) award #R305B170002 to the University of Virginia. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Haynes Foundation, IES, or the U.S. Department of Education.

ORCID iD

Tutrang Nguyen https://orcid.org/0000-0002-6741-6022

Supplemental Material

Supplemental material for this article is available online.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*, 231-257.

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36-46.

Ackerman, D. J., & Coley, R. J. (2012). State pre-K assessment policies: Issues and status. policy information report. *Educational Testing Service*. Retrieved from https://files.eric.ed.gov/fulltext/ED529449.pdf

Argulewicz, E. N., & Abel, R. R. (1984). Internal evidence of bias in the PPVT-R for Anglo-American and Mexican-American children. *Journal of School Psychology*, *22*, 299-303.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

Barnett, W. S., & Friedman-Krauss, A. H. (2016). *State(s) of head start*. Retrieved from http://nieer.org/wp-content/uploads/2016/12/HS_Full_Reduced.pdf

Baroody, A. E., & Diamond, K. E. (2013). Measures of preschool children's interest and engagement in literacy activities: Examining gender differences and construct dimensions. *Early Childhood Research Quarterly*, *28*, 291-301.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*, 1641-1660.

Brown, S. R. (1993). A primer on Q methodology. *Operant Subjectivity*, *16*, 91-138.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

California Department of Education. (2018). *Who are dual language learners?* Retrieved from https://www.desiredresults.us/dll/dual.html

California Department of Education, Early Education and Support Division. (2010). *Desired results regulations management bulletin 04-08*. Retrieved from https://www.cde.ca.gov/sp/cd/ci/mb040804mb.asp

Cameron, C. E., Brock, L. G., Murrah, W. R., Bell, L., Worzalla, S., Grissmer, D. W., & Morrison, F. J. (2012a). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development*, *83*, 1229-1244.

Cameron, C. E., Chen, W. B., Blodgett, J., Cottone, E. A., Mashburn, A. J., Brock, L. L., & Grissmer, D. (2012b). Preliminary validation of the motor skills rating scale. *Journal of Psychoeducational Assessment*, *30*, 555-566.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Congressional Research Service. (2016). *Preschool development grants (FY2014-FY2016) and race to the top—early learning challenge grants (FY2011-FY2013)*. Retrieved from https://www.everycrsreport.com/reports/R44008.html

Connors-Tadros, L. (2014). *Information and resources on developing state policy on Kindergarten Entry Assessment (KEA): CEELO FastFacts*. Washington, DC: Center on Enhancing Early Learning Outcomes.

Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 70-99). Cambridge, UK: Cambridge University Press.

Dixon, L. Q., Zhao, J., Quiroz, B. G., & Shin, J. Y. (2012). Home and community factors influencing bilingual children's ethnic language vocabulary development. *International Journal of Bilingualism*, *16*, 541-565.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *The Journal of Economic Perspectives*, *27*, 109-132.

Duncan, G. J., & Murnane, R. J. (2011). *Whither opportunity? Rising inequality, schools, and children's life chances*. New York, NY: Russell Sage Foundation.

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 261-288). Mahwah, NJ: Lawrence Erlbaum.

Fernandez, M. C., Pearson, B. Z., Umbel, V. M., Oiler, D. K., & Molinet-Molina, M. (1992). Bilingual receptive vocabulary in Hispanic preschool children. *Hispanic Journal of Behavioral Sciences*, *14*, 268-276.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286-299.

Friedman-Krauss, A. H., Barnett, W. S., Weisenfeld, G. G., Kasmin, R., DiCrecchio, N., & Horowitz, M. (2018). *The state of preschool: 2017 state preschool yearbook*. Retrieved from http://nieer.org/wp-content/uploads/2018/05/State-of-Preschool-2017-Full.5.15.pdf

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement*, *63*, 65-74.

Gillanders, C., Franco, X., Seidel, K., Castro, D. C., & Méndez, L. I. (2017). Young dual language learners' emergent writing development. *Early Child Development and Care*, *187*, 371-382.

Gioia, G. A., & Isquith, P. K. (2004). Ecological assessment of executive function in traumatic brain injury. *Developmental Neuropsychology*, *25*, 135-158.

Gormley, W. T. (2008). The Effects of Oklahoma's pre-K program on Hispanic children. *Social Science Quarterly*, *89*, 916-936.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 143-164). New York, NY: John Wiley.

Grissmer, D., Grimm, K. J., Aiyer, S. M., Murrah, W. M., & Steele, J. S. (2010). Fine motor skills and early comprehension of the world: Two new school readiness indicators. *Developmental Psychology*, *46*, 1008-1017.

Harris, Y. R., & Schroeder, V. M. (2013). Language deficits or differences: What we know about African American vernacular English in the 21st century. *International Education Studies*, *6*, 194-204.

Heroman, C., Burts, D. C., Berke, K., & Bickart, T. S. (2010). *Teaching strategies GOLD® objectives for development & learning: Birth through kindergarten*. Washington, DC: Teaching Strategies.

High/Scope Educational Research Foundation. (1992). *High/Scope Child Observation Record (COR) for ages 2½-6*. Ypsilanti, MI: High/Scope Press.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*, 4-14.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Immekus, J. C., & McGee, D. (2016). The measurement invariance of the student opinion scale across English and non-English language learner students within the context of low-and high-stakes assessments. *Frontiers in Psychology, 7*, Article 1352.

Improving Head Start for School Readiness Act of 2007. Pub. L. No. 110-134, § 121 Stat.1363. (2007). Retrieved from http://www.gpo.gov/fdsys/pkg/PLAW-110publ134/html/PLAW-110publ134.htm

Jodoin, M. G., & Gierl, M. J. (2001.). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349.

Kagan, S. L., & Garcia, E. (2007). *Taking stock: Assessing and improving early childhood learning and program quality*. Washington, DC: Pew Charitable Trusts.

Karelitz, T. M., Parrish, D. M., Yamada, H., & Wilson, M. (2010). Articulating assessments across childhood: The cross-age validity of the desired results developmental profile–revised. *Educational Assessment*, *15*, 1-26.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, *14*, 317-324.

Lalor, A., Brown, T., & Murdolo, Y. (2016). Relationship between children's performance-based motor skills and child, parent, and teacher perceptions of children's motor abilities using self/informant-report questionnaires. *Australian Occupational Therapy Journal*, *63*, 105-116.

Lee Webb, M. Y., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody picture vocabulary test–III. *Educational and Psychological Measurement*, *68*, 335-351.

Lonigan, C. J., Allan, N. P., & Lerner, M. D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. *Psychology in the Schools*, *48*, 488-501.

López, L. M. (2012). Assessing the phonological skills of bilingual children from preschool through kindergarten: Developmental progression and cross-language transfer. *Journal of Research in Childhood Education*, *26*, 371-391.

López, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of Hispanic head start children. *Bilingual Research Journal*, *28*, 1-18.

Magnuson, K., Lahaie, C., & Waldfogel, J. (2006). Preschool and school readiness of children of immigrants. *Social Science Quarterly*, *87*, 1241-1262.

Mancilla-Martinez, J., & Lesaux, N. K. (2011). The gap between Spanish speakers word reading and word knowledge: A longitudinal study. *Child Development*, *82*, 1544-1560.

Mather, N., McGrew, N., & Woodcock, R. W. (2001). *Woodcock-Johnson tests of achievement* (3rd ed.). Itasca, IL: Riverside Publishing.

Meisels, S. J., Jablon, J., Marsden, D. B., Dichtelmiller, M. L., & Dorfman, A. B. (1994). *The work sampling system: An overview* (3rd ed.). Ann Arbor, MI: Rebus Planning Associates.

Miller-Bains, K. L., Russo, J. M., Williford, A. P., DeCoster, J., & Cottone, E. A. (2017). Examining the validity of a multidimensional performance-based assessment at kindergarten entry. *AERA Open*, *3*, 1-16.

Missouri Department of Elementary and Secondary Education. (2013). *School readiness tool*. Retrieved from https://dese.mo.gov/quality-schools/early-learning/school-readiness-tool

Mohler, G. M., Yun, K. A., Carter, A., & Kasak, D. (2009). The effect of curriculum, coaching, and professional development on prekindergarten children's literacy achievement. *Journal of Early Childhood Teacher Education*, *30*, 49-68.

National Center for Education Statistics. (2011). *The condition of education 2011: Section 1 participation in education*. Retrieved from http://nces.ed.gov/pubs2011/2011033_2.pdf

National Institute for Early Education Research. (2016). *Special report: Dual language learners and preschool workforce*. Retrieved from http://nieer.org/wp-content/uploads/2016/05/2015_DLL_and_Workforce_rev1.pdf

Oldfather, P., & Wigfield, A. (1996). Children's motivation for literacy learning. In L. Baker, P. Afflerbach, & D. Reinking (Eds.), *Developing engaged readers in the school and home communities* (pp. 89-113). Hillsdale, NJ: Lawrence Erlbaum.

Paez, M. M., Tabors, P. O., & Lopez, L. M. (2007). Dual language and literacy development of Spanish-speaking preschool children. *Journal of Applied Developmental Psychology*, *28*, 85-102.

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., . . . Weiland, C. (2017). Puzzling It Out: The current state of scientific knowledge on pre-kindergarten effects. In D. A. Phillips & K. A. Dodge (Eds.), *The current state of scientific knowledge on pre-kindergarten effects* (pp. 19-30). Washington, D.C.: Brookings Institution and Duke University.

Quirk, M., Mayworm, A., Edyburn, K., & Furlong, M. J. (2016). Dimensionality and measurement invariance of a school readiness screener by ethnicity and home language. *Psychology in the Schools*, 53, 772-784.

Quirk, M., Nylund-Gibson, K., & Furlong, M. (2012). Exploring patterns of Latino/a children's school readiness at kindergarten entry and their relations with Grade 2 achievement. *Early Childhood Research Quarterly*, *28*, 437-449.

Quiroga, T., Lemos-Britton, Z., Mostafapour, E., Abbott, R. D., & Berninger, V. W. (2002). Phonological awareness and beginning reading in Spanish-speaking ESL first graders: Research into practice. *Journal of School Psychology*, *40*, 85-111.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*, 853-891.

Reese, L., Arauz, R. M., & Bazán, A. R. (2012). Mexican parents' and teachers' literacy perspectives and practices: Construction of cultural capital. *International Journal of Qualitative Studies in Education*, *25*, 983-1003.

Reese, L., & Gallimore, R. (2000). Immigrant Latinos' cultural model of literacy development: An evolving perspective on home-school discontinuities. *American Journal of Education*, *108*, 103-134.

Reese, L., & Goldenberg, C. (2008). Community literacy resources and home literacy practices among immigrant Latino families. *Marriage & Family Review*, *43*, 109-139.

Russo, J. M., Williford, A. P., Markowitz, A. M., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, *48*, 14-25.

Schick, A. R., & Melzi, G. (2016). Print-related practices in low-income Latino homes and preschoolers' school-readiness outcomes. *Journal of Early Childhood Literacy*, *16*, 171-198.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 173-182). New York, NY: Guilford Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Snow, C., & Van Hemel, S. (2008). *Early childhood assessment: Why what and how? Report of the committee on developmental outcomes and assessments for young children*. Washington, DC: National Academies Press.

Soderberg, J., Stull, S., Cummings, K., Nolen, E., McCutchen, D., & Joseph, G. (2013). *Inter-rater reliability and concurrent validity study of the Washington Kindergarten Inventory of Developing Skills (WaKIDS)* (Unpublished report prepared for the State of Washington Office of Superintendent of Public Instruction). Retrieved from http://www.k12.wa.us/WaKIDS/pubdocs/WaKIDS_Report072613.pdf

StataCorp. (2015). *Stata statistical software: Release 14*. College Station, TX: Author.

Sutter, C., Ontai, L. L., Nishina, A., Conger, K. J., Shilts, M. K., & Townsend, M. S. (2017). Utilizing the desired results developmental profile as a measure of school readiness: Evaluating factor structure and predictors of school readiness. *Early Child Development and Care*, *187*, 1433-1445.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Wakabayashi, T., Claxton, J., & Smith, E. V. (2019). Validation of a revised observation-based assessment tool for children birth through Kindergarten: The COR advantage. *Journal of Psychoeducational Assessment*, *37*, 69-90.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—or whose score is it anyway? *Early Childhood Research Quarterly*, *27*, 46-54.

Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon, R. M. Lerner, & N. Eisenberg (Eds.), *Handbook of child psychology, Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933-1002). New York, NY: John Wiley.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W., et al. (2013). *Investing in our future: The evidence base on preschool education*. New York, NY: Foundation for Child Development, Society for Research in Child Development.

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.