Journal of Educational and Developmental Psychology; Vol. 3, No. 1; 2013
ISSN 1927-0526 E-ISSN 1927-0534
Published by Canadian Center of Science and Education

The Factor Structure and Invariance of an Observational Checklist to Measure Children's Emergent Literacy Skill Development across Male and Female Samples

Jason C. Immekus[1]

[1] Department of Educational Research and Administration, California State University, Fresno, USA

Correspondence: Jason C. Immekus, Department of Educational Research and Administration, California State University, Fresno, 93740, USA. Tel: 1-559-278-0265. E-mail: jimmekus@csufresno.edu

Received: December 11, 2012 Accepted: January 17, 2013 Online Published: March 21, 2013

doi:10.5539/jedp.v3n1p101 URL: http://dx.doi.org/10.5539/jedp.v3n1p101

Abstract

Study purpose was to test the factor structure of the Jumpstart School Success Checklist (JSSC) and tests its measurement invariance (factor structure similarity) across male and female samples, based on national Jumpstart data (N = 5,545). Factor analytic results supported conceptualizing the JSSC item-level data in terms of a bifactor model (Gibbons & Hedeker, 1992), where each scale item related to a primary factor (Literacy) in addition to one sub-domain: Language Arts or Social Relationships. A comparison of the equivalence of the JSSC factor structure across sex groups indicated that the scale's factor structure met *partial measurement invariance* (Bryne, Shavelson, & Muthén, 1989). A follow-up latent means structure analysis reported that females had slightly higher latent means across the factors than males. Study implications pertain to (a) the degree to which the JSSC scores function across sex groups, and (b) how factorial invariance research can be used to examine raters' assessment of students' literacy skill development.

Keywords: test score validity, factor analysis, measurement invariance, observational checklist

1. Introduction

Children's emergent literacy skill development is a continual process that occurs well before exposure to formal schooling in kindergarten (Fields, Groth, & Spangler, 2008; Teale & Sulzby, 1986; Whitehurst & Lonigan, 1998). Acquisition these foundational literacy skills (e.g., awareness of print) is recognized as a prerequisite of successful attainment of future reading and writing outcomes (Rouse, Brooks-Gunn, & McLanahan, 2005; Snow, Burns, & Griffin, 1998). Recently released National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 2011) report a sobering picture of the low reading attainments of minority (e.g., African American) and low income 4th and 8th grade students. Empirical findings suggest that gaps in students' school readiness may explain these findings in terms of economic, social, and health indicators (Rouse, Brooks-Gunn, & McLanahan, 2005). In response, early literacy preschool programs have been advanced as strategic investments to promote children's early literacy skill development to improve subsequent reading and writing outcomes (e.g., Whitehurst & Lonigan, 1998). A primary consideration with the screening and identification of children's early literacy skills within the context of preschool-aged programs is the extent to which the psychometric properties (e.g., reliability) of literacy assessment scores support their use for decision-making purposes (e.g., progress monitoring).

Accurate measurement of preschool-aged children's emergent literacy skills is critical for early identification and prevention in addition to program evaluation purposes. Within the context of early literacy programs, test scores serve a wide array of functions, including: progress monitoring; assessing the responsiveness of instruction to students' needs; identification of children who may need additional support; and, communicating information on children's literacy development to parents/guardians (Neuman, Copple, & Bredekamp, 2000). Two approaches to the measurement of young children's emergent literacy skills include direct and indirect assessments. Direct assessments embody individually- or group- administered, standardized instruments designed to yield scores for screening and diagnostic purposes. However, empirical findings have raised important questions regarding the use of such measures with young children (Konold & Pianta, 2005; La Pora & Pianta, 2000). Furthermore, individually-administered measurements may not be feasible to administer and score in early literacy programs

for a number of reasons, including: limited funds to cover the costs of the assessments and materials, qualified individuals to administer the tests, and feasibility to assess a large number of students in an efficient, timely manner. While direct assessments may serve useful in the context of controlled, empirical studies to investigate instructional or program effectiveness to promote children's literacy skill acquisition, in many cases they may not provide program providers a quick, efficient method of assessment in a naturalistic environment.

On the other hand, indirect assessments can be broadly characterized as instruments that involve informants (e.g., teachers) evaluating a child's emergent literacy skills. Both observational rating forms and checklists represent indirect measures that have promoted to assess young children's behavior and literacy outcomes (e.g., Neuman & Roskos, 2007). Cabell, Justice, Zucker, and Kilday (2010) identify several attractive features of these measures, including: time and cost efficient; convenient completion; elimination of child characteristics in testing (e.g., mood); and, lastly, may offer more specific developmental information than provided by direct instruments. Notwithstanding these benefits, there are important factors to consider related to their use for assessing literacy skills. For instance, these measures are not designed to measure strengths/weaknesses in certain dimensions of literacy skills (Lonigan, 2006). Popham (2000) identifies the scoring scale, raters (e.g., teachers), and scoring procedure as key sources of error associated with the practice of evaluating student outcomes. Nonetheless, such instruments enjoy widespread use as one source of information used by early childhood programs to evaluate childhood outcomes, as well as within empirical research (e.g., Lapointe, Ford, & Zumbo, 2007). Consequently, a body of research is emerging regarding the psychometric properties of these types of measures (e.g., predictive validity; Cabell et al., 2011).

Based on the use of indirect assessments across practical and research settings, imperative questions related to their use includes (a) whether the resultant scores represent the scale's theoretical factor structure, and (b) the extent to which obtained scores have similar measurement properties across diverse student sub-groups (e.g., sex, race/ethnicity). Factor analysis represents a broad class of statistical procedures to investigate empirical questions related to the structure of scale data. Thompson (2004) identifies three purposes of factor analysis: (a) gather empirical evidence on test score validity, (b) develop theory on hypothetical constructs (e.g., literacy), and (c) summarize relationships among variables using factor scores. The two major classes of factor analysis include exploratory and confirmatory factor analysis. Whereas exploratory factor analysis (EFA) is a data-driven approach to identifying the number of factors underlying scale data, confirmatory factor analysis (CFA) is based on the use of *a priori* information (or theory) to test the number of factors explaining the relationship among a set of observed variables (e.g., rating scale items).

Measurement invariance is a desired property of test scores that indicates that the psychometric properties (e.g., discrimination) of the scores are the same across compared groups (e.g., experimental vs. control). Within the factor analytic framework, measurement invariance is tested using multi-sample CFA by testing the statistical fit of competing models that differ in terms of the model parameters set equal across compared groups (Millsap & Yun-Tein, 2004). A finding of measurement invariance indicates that scores can be interpreted similarly, whereas a lack of invariance indicates scores cannot be interpreted the same. Thus, a lack of invariance indicates that across group score disparities may be due to trait (e.g., literacy) differences in addition to measurement error (Raju, Laffitte, & Bryne, 2002). Based on these considerations, measurement invariance research provides literacy researchers vital information on the meaning of obtained scores, as well as an avenue to pursue research into students' literacy development.

Structural equation modeling (SEM) offers a valuable model-based approach to investigate the relationships among observed (e.g., items) and latent (e.g., literacy) variables (Bollen, 1989). SEM can also be used to formally test the measurement invariance of scale items to judge whether the psychometric properties of scores are similar across diverse groups (e.g., sex, race/ethnicity). This entails fitting and comparing a series of increasingly restrictive models that differ by the particular item parameter(s) constrained equal across groups. Measurement model parameters of interest include: factor loadings (i.e., discrimination), thresholds, and residuals (error terms). Factor loadings characterize the relationship between the observed and latent variables and are directly related to item discrimination (Bock & Gibbons, 2010). Thresholds indicate the point on the trait continuum where there is a given probability of selecting a particular response option over the next lowest category (e.g., selecting Agree over Neutral). Residuals indicate the amount of item variance unexplained by the underlying latent trait, or unexplained error. The degree to which these model parameters are similar across groups corresponds to the level of invariance of an instrument's factor structure. A finding of *partial measurement invariance* (Byrne, Shavelson, & Muthén, 1989) provides a basis for comparing groups on the underlying latent means.

The purpose of this study was twofold. First, it sought to employ factor analytic procedures to test the factor

structure of the Jumpstart School Success Checklist (JSSC), a 15-item rating scale completed by informants (e.g., mentors) regarding the literacy skills of preschool aged children enrolled in Jumpstart, a national supplemental pre-kindergarten program designed to promote young children's language and literacy skills (see www.jstart.org for program description and background). Second, the study tested the extent to which the JSSC factor structure was invariant (or similar) across male and female samples? Study findings are designed to contribute to the literature base regarding the psychometric properties of indirect or observational measures to assess young children's emergent language and literacy skills.

2. Method

2.1 Participants

Study data included the item-level pretest data of males (n = 2,760) and females (n = 2,739) with complete data comprising the 2007-2008 JSSC dataset (N = 5,545). As reported in Table 1, females comprised 50.30% of the sample, and slightly half of the sample (47.75%) was enrolled in the Jumpstart program. The majority of the sample was African American (40.19%), followed by Hispanic (31.40%), White (16.50%), Asian (6.55), and other (5.37%). The majority of the children sample spoke English only (62.90%), and the average age was 48.78 months (SD = 6.14; range = 36 to 59 months).

Table 1. 2007-2008 jumpstart population demographics

Variable	*n*	%
Program Status		
Comparison	2,897	52.25
Jumpstart	2,648	47.75
Sex		
Males	2,789	50.30
Females	2,756	49.70
Race/Ethnicity		
African American	2,044	40.19
Asian	333	6.55
Hispanic	1,597	31.40
White	839	16.50
Other	273	5.37
Missing	459	-
Primary Spoken Language		
English	3,382	62.90
Spanish	635	11.81
Chinese	76	1.41
Haitian-Creole	13	0.24
Other single language	100	1.86
English-Spanish	798	14.84
English-Chinese	106	1.97
English-Haitian-Creole	51	0.95
English-Other	216	4.02
Child Language		
English Only	3,382	62.90
Spanish Only	635	11.81
Other Single	189	3.51
English Other	1,171	21.78

Note. N = 5,545

2.2 Instrumentation

The JSSC is a 15-item observational rating form designed to assess preschool students' literacy skills. Each scale item relates to a specific area of literacy (e.g., using vocabulary, relating to adults) and corresponds to either the Language Arts or Social Relationships subscale. As shown in Table 2, the Language Arts sub-domain consists of 8 items and Social Relationships consists of 7 items. The instrument is administered as a pre- and post-test and completed by program providers (e.g., mentors) who rate each child's literacy skills across the items using a 5-point scale based on the child's demonstration of specific levels of literacy proficiency. The scale also collects student demographic information, such as: date of birth, sex, and language spoke, among others.

Table 2. Jumpstart school success checklist

Domain/Item	Question
Language Arts	
1	List to and understanding speech
2	Using vocabulary
3	Using complex patterns of speech
4	Showing awareness of sounds in words
5	Demonstrating knowledge about books
6	Using letter names and sounds
7	Reading
8	Writing
Social Relationships	
9	Making choices and plans
10	Solving problems with materials
11	Initiating play
12	Resolving interpersonal conflict
13	Understanding and expressing feelings
14	Relating to adults
15	Relating to other children

Note. Ratings provided on 5-point scale.

2.3 Data Analysis

Due to the limited availability of information on the underlying JSSC factor structure, the sample was randomly divided in half to investigate scale dimensionality. CFA was used to test the JSSC two-factor model of the item-level data, based on the first random group data (n = 2,749). CFA was deemed appropriate since the JSSC is a theoretically-based instrument designed to measure preschool students' literacy skills across the domains of language arts and social behavior (Kline, 2005). Model specification entailed first fitting a correlated two-factor model to the data, with the first eight items specified to the Language Arts factor and the remaining seven items specified to the Social Relationships factor (see Table 2). Model-data fit was based on inspection of the statistical fit of the theoretical model to the data.

Due to the ordinal nature of the item-level data (e.g., Likert scale), robust weighted least squares (WLSMV; Muthén, du Toit, & Spisic, 1997) was used for parameter estimation using MPLUS 5.0 (Muthén & Muthén, 1998-2006). Model fit was evaluated in terms of the following fit statistics: chi-square statistic (WLSMV), root mean square error of approximation (*RMSEA*), and comparative fit index (*CFI*). The *RMSEA* provides a measure of the discrepancy between the actual and estimated variance-covariance matrix per degree of freedom. *RMSEA* values less than .06 were used to indicate good model fit and those less than .08 suggested reasonable fit (Hu & Bentler, 1999). The *CFI* provides a measure of the discrepancy between a restricted and null model in relation to the fit of the null model (Bentler, 1990), with values equal to or above .95 used to indicate adequate fit (Hu &

Bentler, 1999).

Provided the two-factor model of the JSSC did not reported acceptable model-data fit (as based on above fit statistics), an EFA based on a principle axis with promax rotation was used to further investigate the scale's underlying factor structure based on the second random group data. The use of EFA to investigate scale dimensionality following a poorly fit confirmatory-based model was intended to address the issue of finding an acceptable model due to chance when testing a series of modified models (MacCallum, Roznowski, & Necowitz, 1992). As the eigenvalue greater than 1.00 rule has been found to result in the overextraction of factors (Zwick & Velicer, 1986), factor retention was based on comparing the eigenvalues of the EFA to those obtained from a parallel analysis (Henson & Roberts, 2006; O'Connor, 2000). To obtain a parsimonious model that demonstrated simple structure (Thurstone, 1947), items reporting cross-loadings (>.30) on multiple factors were considered for removal from subsequent analyses (Hinkin, 1998). Subsequently, CFA was used to test the fit of the exploratory-based model.

Provided that a model was fit to the JSSC data that reported acceptable model-fit, multisample confirmatory factor analysis (MCFA) was used to formally test the measurement invariance of model parameters (e.g., factor loadings; Millsap & Yun-Tien, 2004). MCFA was used to test the invariance of the following matrices: (a) factor loadings (pattern coefficients), (b) thresholds, and (c) error variances. Factor loadings report the strength of association between items and the underlying scale factors (e.g., Language Arts) and, thus, provide critical test score validity information (Keith, 1997). Thresholds indicate the location on the underlying trait continuum (e.g., Literacy) where a student would be assigned to a particular item response category (Rating of 1, 2, 3, 4, or 5) (Bock & Gibbons, 2010). For rating scale data, there are *m*-1 categories (where *m* equals number of rating categories). For the JSSC, there are (5-1) 4 threshold parameters for each item. The error variances deal with the amount of unexplained error in items, and represent final parameters tested for invariance. Each model parameter provides relevant information on the functioning of JSSC scale scores.

As based on Millsap and Yun-Tien (2004), invariance testing of ordered-categorical data was based on testing the statistical difference between a series of increasingly restrictive nested measurement models. The models differed in terms of the matrices (e.g., factor loadings) or parameters (e.g., thresholds) constrained equal across groups. First, an acceptable CFA model was fit to each group's data (referred to as the *free model*). This baseline (or free) model provided the basis for the subsequent test of the invariance of the factor loading matrix. This was conducted by constraining the matrix of factor loadings equal across groups to obtain the likelihood chi-square value of the *constrained model*. The statistical significance of the likelihood chi-square difference value (obtained by comparing the likelihood chi-square statistics of the free and constrained models) provided an indication of whether the factor loading matrix was invariant.

A nonsignificant chi-square difference statistic (based on use of the DIFFTEST in MPLUS, as per Muthén & Muthén, 1998-2006) was used to judge the invariance of model parameters. Conversely, a statistically significant chi-square difference statistic indicated that the constrained model resulted in a decline in model-data fit and that at least one parameter in the matrix lacked invariance. Subsequently, each parameter within the matrix was individually tested for invariance (Reise, Widaman, & Pugh, 1993). Parameters found to lack invariance were specified to be freely estimated in subsequent invariance tests. Sequential tests of nested model comparisons were continued until all matrices (thresholds, residuals) and corresponding parameters were tested for invariance.

As the chi-square difference statistic is known to reject the null hypothesis of equivalent model parameters in invariance testing based on trivial differences in large sample sizes, the incremental changes of the *CFI* and *RMSEA* values were also used in the tests of measurement invariance (Cheung & Rensvold, 2002). Furthermore, the procedures for invariance testing using WLSMV, as well as the *theta parameterization* option in MPLUS to test error variance equality, were employed (Muthén & Muthén, 1998-2006).

If the JSSC factor structure exceeded *partial measurement invariance* (Byrne et al., 1989), across group differences on the latent means were conducted. Inspection of latent mean differences entailed constraining the latent mean of Group 1 to zero and freely estimating the latent mean of Group 2 (Muthén & Muthén, 1998-2006). An effect size estimate (as per Hancock, 2004) was used to indicate the magnitude of the difference between the latent means, with values interpreted as: small (0.2), medium (0.5), and large (0.8) (Cohen, 1988).

3. Results

A test of a two-factor model of the JSSC item-level data resulted in unacceptable model-data fit, X^2 (56) = 2,623.17, *CFI* = .90, *RMSEA* = .13. A subsequent EFA support a one-factor model, based on the retention of empirical factors by comparing the eigenvalues from the EFA to those obtained in a parallel analysis (Henson & Roberts, 2006; O'Connor, 2000). Eigenvalues for the first two factors, based on the EFA, were 9.61 and 0.53,

whereas those based on the parallel analysis were 0.13 and 0.10, respectively. A factor is retained if its eigenvalue based on the EFA is greater than that obtained from the parallel analysis. It can therefore be inferred that a general dominant factor (Literacy) underlies the JSSC scale data.

A subsequent test of a series of CFA models suggested that the JSSC scale data may be modeled in terms of a bifactor model (Gibbons et al., 2007; Gibbons & Hedeker, 1992; Immekus & Imbrie, 2008). Figure 1 illustrates the path diagram depicting the final bifactor model of the JSSC item-level data, which reported acceptable model-data fit, X^2 (52) = 2,107.15, *CFI* = .96, *RMSEA* = .08. Notably, to achieve a simple structure where each item reports a clear relationship to a designated factor, Item 10 was dropped from the analysis due to cross-loading on both sub-domains. As shown, all items were specified to load on a primary dimension, with specific items also allowed to load (or relate) on a secondary dimension (i.e., Language Arts, Social Relationships). The basis of the bifactor model is that each scale item is related to a primary dimension in addition to one sub-domain: Language Arts and Social Relationships.

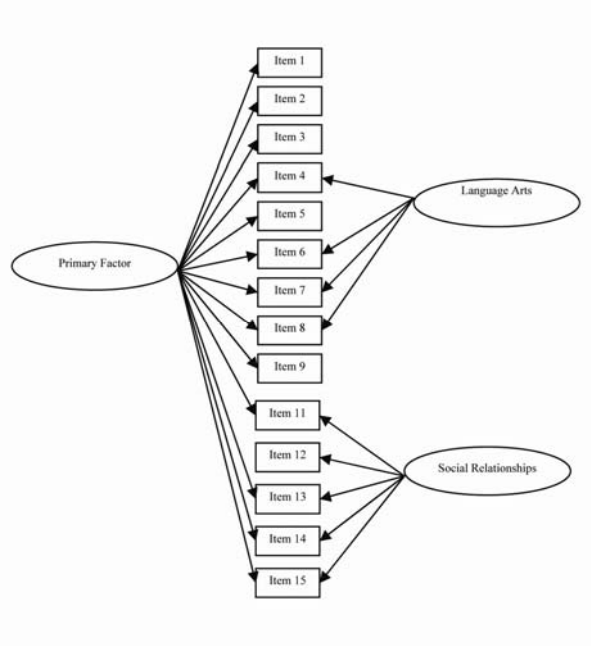

Figure 1. Bifactor model of the JSSC item-level data

Table 3 reports results based on nested model comparisons between the bifactor model and competing one – and two-factor structures of the JSSC (Rindskopf & Rose, 1988). Model comparisons supported conceptualizing the JSSC in terms of a bifactor model (Figure 1) with a primary (Literacy) dimension and two sub-domains: Language Arts and Social Relationship. The bifactor model is well suited for modeling survey data that is typically based on sampling items from sub-domains (e.g., Language, Social Relationships) situated within a broader domain (e.g., literacy).

Table 3. Fit statistics of competing JSSC factor structures

	X^2	*df*	*p*	$X^2_{Difference}$	$Df_{Difference}$	$p_{Difference}$	CFI	RMSEA
Single-factor	5,400.54	51	<.01	-	-	-	.89	.14
Two-factor	4,578.72	51	<.01	-	-	-	.90	.13
Bifactor	2,107.15	52	<.01				.96	.08
Bifactor vs. 1-factor	-	-	-	2,333.68	7	<.01	-	-
Bifactor vs. 2-factor	-	-	-	1,784.78	6	<.01	-	-

Note. Pre-test *N* = 5,499. Item 10 not included in analysis. *CFI* = Comparative fit index. *RMSEA* = Root mean square error of approximation.

The bifactor model (Figure 1) reported acceptable model-data fit for the data of males (X^2 = 1,088.18, $p < .01$) and females (X^2 = 1,039.39, $p < .01$), as well as for the entire sample, X^2 (df = 101) = 2,127.57, $p < .01$, CFI = .96, $RMSEA$ = .08. Although the chi-square statistic p-value was statistically significant ($p < .01$), the statistic is well known to be influenced by sample size and other fit statistics were acceptable (e.g., $CFI > .95$). Table 4 reports the factor loadings and residual errors of the items across groups.

Based on an acceptable fit of the bifactor model to the data, analyses proceeded to testing model parameters for invariance. The initial test that all factor loadings were equal across groups indicated that at least one or more lacked invariance, $X^2_{Difference}$ (df = 18) = 104.97, $p < .01$. Subsequently, a test of the invariance of the factor loadings on the primary (Literacy) factor was conducted, which indicated that one or more loadings differed across groups, $X^2_{Difference}$ (df = 12) = 51.91, $p < .01$. A sequential test of the invariance of each factor loading indicated that the factor loadings of the following items differed across groups on the primary Literacy factor: 4, 5, and 13.

Table 4. Factor loadings and residuals of jumpstart school success checklist for males and females[A]

Item	Factor			Residual
	Primary	Language Arts	Social Relationships	
1	.87 (.87)[B]			.24 (.25)
2	.87 (.88)			**.25 (.22)**
3	.89 (.89)			.21 (.21)
4	**.79** (**.74**)	.28 (.31)[B]		.29 (.36)
5	**.82** (**.81**)			**.32 (.34)**
6	.74 (.72)	.51 (.51)		.19 (.22)
7	.76 (.70)	**.25 (.37**)		**.36 (.38)**
8	.78 (.68)	.43 (.56)		.30 (.23)
9	.86 (.85)			**.25 (.28)**
11	.77 (.78)		**.22** (**.26**)	.35 (.32)
12	.72 (.72)		.37 (.23)	.35 (.44)
13	**.78** (**.80**)		.37 (.26)[B]	**.25 (.29)**
14	.80 (.79)		**.23** (**.26**)	.31 (.32)
15	.75 (.74)		**.27** (**.39**)	**.37 (.30)**

Note. Completely standardized parameter estimates reported.

Bolded values indicate factor loadings that differed across groups ($p < .05$).

[A] Parameter estimates in parenthesis.

[B] Parameter fixed to 1.0 to set factor scale.

An omnibus test of the factor loadings comprising the Language Arts sub-domain indicated a lack of invariance of at least one parameter, $X^2_{Difference}$ (df = 3) = 19.34, $p < .01$. A follow-up test of the invariance of each factor loading on the Language Arts factor indicated that the loading for Item 7 lacked invariance.

Last, a test of the factor loadings comprising the Social Relationships sub-domain indicated a lack of invariance among the parameters, $X^2_{Difference}$ (df = 4) = 62.27, $p < .01$. A follow-up test of the invariance of each factor loading on the Social Relationships factor indicated that the following three parameters lacked invariance: Item 11, Item 14, and Item 15.

Table 5 reports the threshold parameters for each item across sex groups. Threshold parameters indicate the location on the underlying trait continuum (Literacy) where a student would be assigned by a rater (e.g., mentor, teacher) in a particular item response category (e.g., 2, 3, 4, or 5). Therefore, for Item 1, the point on the underlying trait continuum where males and females would be more likely to receive a rating of 2 instead of a 1

would be approximately -.40 (almost a half a standard deviation below the mean [0], based on z-score scale [mean = 0, standard deviation = 1]), and .10 (above mean) for a score of 3 instead of a 2. For Item 5, which lacked invariance, the threshold of 2-3 (being assigned to category 3 over category 2) indicated that males and females had a different point on the underlying continuum for receiving a rating of 3. In this case, females have a higher likelihood of being rated in this category than males, based on lower trait level. That is, compared to males, females were more likely to be assigned to a rating of 3 with lower levels of Literacy.

Results indicated that one or more threshold parameters differed across groups, $X^2_{Difference}$ (df = 27) = 117.04, p < .01. Subsequently, the four thresholds (one less than number of response categories) for each item were tested for equality across groups. Results indicated that specific thresholds (see Table 12) for the following items lacked invariance: 5, 6, 7, 11, 12, 13, and 15.

Of less importance was the degree to which each item's residual error was similar across gender groups. Residual errors are reported in Table 11. A lack of invariance among error terms indicated that there were different amounts of error in each group's item score. A test of the similarity of the JSSC item error terms indicated that one or more lacked invariance, $X^2_{Difference}$ (df = 11) = 43.36, p < .01. In particular, the following items' error terms lacked invariance: 2, 5, 7, 9, 13, and 15. Notably, a lack of invariance among error terms is often not considered in applied research, and a finding of error term invariance would represent a very strict level of invariance.

Table 5. Item threshold parameters for males and females[A]

	Category			
Item	1-2	2-3	3-4	4-5
1	-0.40 (-0.41)	0.10 (0.10)	0.61 (.61)	1.29 (1.31)
2	-0.74 (-0.75)	-0.05 (-0.03)	0.63 (0.67)	1.60 (1.71)
3	-0.88 (-0.86)	-0.10 (-0.08)	0.71 (0.74)	1.48 (1.51)
4	-0.29 (-0.32)	0.79 (0.86)	1.37 (1.48)	2.01 (2.14)
5	-0.94 (-1.07)	**-0.13** (**-0.23**)	**0.77** (**0.65**)	**1.91** (**1.92**)
6	0.01 (0.23)	**0.74** (**0.81**)	**1.13** (**1.23**)	**1.79** (**2.01**)
7	-0.97 (-0.88)	**0.46** (**0.50**)	**1.26** (**1.36**)	**2.31** (**2.58**)
8	0.01 (0.02)	0.44 (0.42)	1.47 (1.41)	2.38 (2.42)
9	-1.22 (-1.23)	-0.35 (-0.31)	0.73 (0.77)	1.69 (1.70)
11	-1.03 (-1.04)	**-0.31** (**-0.43**)	0.69 (0.65)	1.40 (1.41)
12	-0.68 (-0.84)	**0.45** (**0.39**)	**1.15** (**1.16**)	**1.89** (**1.96**)
13	-0.36 (-0.50)	0.23 (0.24)	**1.10** (**1.00**)	**1.47** (**1.39**)
14	-0.48 (-0.45)	-0.03 (-0.03)	0.70 (0.65)	1.35 (1.30)
15	-0.97 (-0.99)	**-0.10** (**-0.18**)	**0.46** (**0.34**)	1.21 (1.21)

Note. Completely standardized parameter estimates reported. Bolded values indicate thresholds that differed across groups (p < .05).

[A] Threshold parameter estimates in parenthesis.

Based on the JSSC factor structure demonstrating partial measurement invariance (Byrne et al., 1989), a follow-up comparison of latent mean differences was conducted. Results indicated that females had higher standing on the underlying traits than males. Based on Cohen's (1988) interpretation of effect sizes, the magnitude of the difference in favor of females was small across the latent factors: Primary (Literacy) factor (.17), Language Arts (.16), and Social Relationships (.09).

4. Discussion

Results of this study found that the parameters of the JSSC bifactor model demonstrated partial measurement invariance. What this means is that the model parameters are roughly the same across groups, including: factor

loadings, thresholds, and error variances. Factor loadings are considered the most important model parameters since they indicate the strength of the relationship between the items (observed variable) and factors (unobserved variable) (Keith, 1997). More specifically, factor loadings indicate the degree to which an item measures a particular factor and deals directly with test score validity. The finding that the factor loadings of Items 4, 5, and 13 on the primary (Literacy) factor lacked invariance indicates that these items are not equally discriminating across male and female aged preschool aged children. That is, Items 4 and 5 were slightly more discriminating for males than females, whereas Item 13 was more discriminating for females than males. To recall, discrimination deals with an item's ability to tease apart differences between students with low and high literacy skills.

Specific factor loadings also lacked invariance on the secondary factors of Language Arts and Social Relationships. Again, the lack of invariance of the secondary loading of Item 7 on the Language Arts factor indicated that the item was more discriminating for females than males. Likewise, the lack of invariance for Item 11, Item 14, and Item 15 on the Social Relationships factor indicated that each item was more discriminating for females than males. The largest difference was for Item 15, whereas only slight differences were found for Items 11 and 14.

From a practical stand point, the findings of certain items being more discriminating for males than females, and vice-versa, indicate areas for further consideration not necessarily scale revision. For example, the factor loading of Item 4 on the primary dimension reported the largest discrepancy between males and females. The item deals with the awareness of sounds in words and was more discriminating for males. Practically speaking, this indicated that it was easier for raters (i.e., mentors) to identify males who could or could not show awareness of the sounds in words than females. In terms of the factor loading of Item 7 on the Language Arts factor, the item was more discriminating for females. Therefore, raters were more likely to identify females who were or were not beginning to read than males; or, it was more difficult to identify (or discriminate between) males who were beginning to read.

On the Social Relationships factor, Items 11, 14, and 15 were slightly more discriminating for females. Similarly, raters (e.g., teachers, mentors) were more readily able to discriminate between female students initiating play (Item 11), relating to adults (Item 14), and relating to other children (Item 15) than males. Overall, these findings point to areas of consideration to determine why those completing the JSSC may be able to discriminate between the literacy skills of males or females. This looks like a potential area for more research (e.g., factors associated with students' demonstration of literacy skills and subsequent observer ratings), not necessarily a finding that warrants revising the JSSC. Given that only a few factor loadings lacked invariance provides evidence that the items seem to measure the intended traits roughly the same across gender groups.

Several of the item thresholds were also found to lack invariance across gender groups. The finding of threshold differences deals with the ability level of the student and his/her likelihood of being assigned to a particular response category, such as: Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. For example, consider Item 5 (demonstrating knowledge about books), which reported a lack of invariance for thresholds 2 (point on scale of going from a rating of 1 to 2), 3 (point on scale of going from a rating of 2 to 3), and 4 (point on scale of going from a rating of 4 to 5). For this item, males had a reported threshold of -.13 and females had a value of -.23 for threshold 2, which indicated the point on underlying trait continuum for a probability of receiving a score of 2 instead of 1. Here, females with lower literacy skills were more likely to receive a score of 2 than males, who needed a higher trait level (-.13) to receive a rating of 2 on Item 5. Similarly, the trait level for females to receive a rating of 4 on the item was .65, compared to .77 for males. Thus, depending on the item, the trait level needed to be assigned to a particular categorical rating differed across gender groups for certain JSSC items. Similar to the finding of factor loading differences, the lack of invariance among certain thresholds provides a basis for more research into why a rater (e.g., mentor) may not assign the same score to a student based on gender.

Although not as critical as the aforementioned findings, specific item error terms differed across groups. Error terms that differed across groups included: Items 2, 5, 7, 9, 13, and 15. The lack of invariance among error terms indicated that there are different amounts of unexplained variance in item scores of males and females, after accounting for the variance explained by the latent traits (i.e., Literacy, Language Arts, & Social Relationships). This finding is not surprising given the types of error that may influence student ratings on the JSSC, including: experience level of rater, students' demonstration of literacy skills, and raters' familiarity of the student being rated, among many.

The finding of partial measurement invariance of the measurement parameters of the JSSC provided justification

for a comparison of latent mean differences across gender groups. A comparison of male and female latent means is desirable because it accounts for differences in the measurement properties of the instrument itself. In the present study, females were found to have higher standing on the latent traits underlying the JSSC item-level data (i.e., Literacy, Language Arts, & Social Relationships) than males. For the primary dimension of Literacy, the difference between the latent means of males and females was small, as based on the effect size of .17. Similarly, the difference in the latent means of males and females on the Language Arts sub-domain was small (effect size equal to .16), whereas negligible differences were found on the sub-domain of Social Relationships (.09). Notably, these differences should be interpreted cautiously given the finding of partial measurement invariance.

The JSSC was found to display partial measurement invariance. Whereas specific factor loadings, threshold, and error variances were found to differ across groups, such differences are not likely to impact programmatic decisions based on test scores. Alternatively, these findings suggest some areas of future consideration, such as why certain literacy skills, such as showing awareness of sounds in words (Item 4), discriminate between high and low ability males more than females, and vice-versa. One way to further explore this issue is testing the generalizability of these findings by conducting the same analyses based on more recent Jumpstart data; alternatively, conducting focused interviews with raters (e.g., mentors) to gauge factors they consider when assigning a particular rating to a student may provide a valuable source of information on this topic. As such, current findings do not necessarily lend themselves to specific recommendations to modify the JSSC instrument itself. Instead, the results point to areas for further inquiry into the judgments of raters when rating the literacy skills of male and female students.

References

Bentler, P. M. (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bock, R. D., & Gibbons, R. D. (2010). Factor analysis of categorical item responses. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 155-184). New York: Routledge.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, New York: John Wiley & Sons.

Bryne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466. http://dx.doi.org/ 10.1037/0033-2909.105.3.456

Cabell, S. Q., Justice, L. M., Zucker, T. A., & Kilday, C. R. (2010). Validity of teacher report for assessing the emergent literacy skills of at-risk preschoolers. *Language, Speech, & Hearing Services in Schools, 40*, 161-173. http://dx.doi.org/10.1044/0161-1461

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academy Press.

Fields, M. V., Groth, L A., & Spangler, K. L. (2008). *Let's begin reading right: A developmental approach to emergent literacy* (6th ed.). Upper Saddle River, New Jersey: Pearson.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19. http://dx.doi.org/10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57*(3), 423-426. http://dx.doi.org/10.1007/BF02295430

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393-416.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104-121. http://dx.doi.org/10.1177/109442819800100106

Hu, L., & Benter, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. http://dx.doi.org 10.1080/10705519909540118

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement, 68*, 695-709.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 373-403). New York: Guilford.

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

Konold, T. R., & Pianta, R. C. (2005). Empirically-derived, person oriented patterns of school readiness in typically-developing children: Description and prediction to first-grade achievement. *Applied Developmental Science, 9*, 174-187. http://dx.doi.org/10.1207/s1532480xads0904_1

La Paro, K. M., & Pianta, R. C. (2000). Predicting children's competence in early school years: A meta-analytic review. *Review of Educational Research, 70*, 443-484. http://dx.doi.org/10.3102/00346543070004443

Lapointe, V. R., Ford, L., & Zumbo, B. D. (2007). Examining the relationship between neighborhood environment and school readiness for kindergarten children. *Early Childhood and Development, 18*, 473-495.

Lonigan, C. J. (2006). Development, assessment, and promotion of preliteracy skills. *Early Education and Development, 17*, 91-114.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490-504.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479-515. http://dx.doi.org/10.1207/S15327906MBR3903_4

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equation in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript. Retrieved from http://www.statmodel.com/papers.shtml

Muthén, L. K., & Muthén, B. O. (1998-2006). *MPLUS user's guide* (4th ed). Los Angelos, CA: Muthén & Muthén.

National Center for Education Statistics. (2011). *The Nation's Report Card: Reading 2011 (NCES 2012–457).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Neuman, S. B., & Roskos, K. (2007). *Nurturing knowledge: Building a foundation for school success by linking early literacy to math, science, art, and social studies*. New York: Scholastic.

Neuman, S., B., Copple, C., & Bredekamp, S. (2000). *Learning to read and write: Developmentally appropriate practice*. Washington, DC: NAEYC.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*, 396-402.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3rd ed.). Boston, MA: Allyn & Bacon.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517-529. http://dx.doi.org/ 10.1037//0021-9010.87.3.517

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*, 51-67.

Rouse, C., Brooks-Gunn, J., & McLanahan, S. (2005). Introducing the issue. *The Future of Children, 15,* 5-14. http://dx.doi.org/10.1353/foc.2005.0010

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

Teale, W. H., & Sulzby, E. (1986). *Emergent literacy: Writing and reading*. Norwood, NJ: Ablex Publishing.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10694-000

Thurston, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69,* 848-872. Retrieved from http://www.jstor.org/stable/1132208

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.