JSLHR

Research Article

Hierarchy and Reliability of the Preschool Language Scales–Fifth Edition: Mokken Scale Analysis

Yu-Yu Hsiao,[a] Cathy Huaqing Qi,[b] Robert Hoy,[a] Philip S. Dale,[c] Glenda S. Stump,[d] Megan Dunn Davison,[e] and Yinglin Xia[f]

Purpose: This study examined the psychometric properties of the Preschool Language Scales–Fifth Edition (PLS-5 English) among preschool children from low–socioeconomic status (SES) families.
Method: The PLS-5 was administered individually to 169 3- to 4-year-old children enrolled in Head Start programs. We carried out a Mokken scale analysis (MSA), which is a nonparametric item response theory analysis, to examine the hierarchy among items and the reliability of test scores of the PLS-5 Auditory Comprehension (AC) and Expressive Communication (EC) scales.
Results: The PLS-5 EC items retained a moderate Mokken scale with the inclusion of all the items. On the other hand, the PLS-5 AC items formed a moderate Mokken scale only with the exclusion of five unscalable items. The latent class reliability coefficients for the AC and the EC scale scores were both above .90. Several items that violated the invariant item ordering assumption were found for both scales.
Conclusions: MSA can be used to examine the relationship between the latent language ability and the probability of passing an item with ordinal responses. Results indicate that for preschool children from low-SES families, it is appropriate to use the PLS-5 EC scale scores for comparing individuals' expressive language abilities; however, researchers and speech-language pathologists should be cautious when using the PLS-5 AC scale scores to evaluate individuals' receptive language abilities. Other implications of the MSA results are further discussed.

Language skills, and literacy skills that build on them (Catts & Kamhi, 2012), are foundational for cognitive, educational, and psychological outcomes such as peer relationships, social skills, relationships, vocational attainment, and many aspects of life satisfaction (National Academies of Science, Engineering, and Medicine [NASEM], 2016; Oreopoulos & Salvanes, 2011). Thus, accurate evaluation of language knowledge and skills of preschool children is essential. Based on national population surveys, approximately 11% of 3- to 6-year-old children were estimated to have speech and language disorders (NASEM, 2016). Language delays or language disorders increase the risk for peer rejection, higher levels of problem behaviors, and lower academic achievements later (Qi et al., 2020; Janus et al., 2019; NASEM, 2016; Norbury et al., 2016; Rantalainen et al., 2021; Slot et al., 2021). Poverty exacerbates the risk for language delays, as research shows that children from low-socioeconomic status (SES) families perform poorly on standardized language tests (Qi et al., 2003; Hammer et al., 2010; Letts et al., 2013; Levine et al., 2020; Nelson et al., 2011). Young children who appear to have a delay in language skills present speech-language pathologists (SLPs) with the challenging task of determining whether a language disorder exists.

Standardized language tests are widely used by SLPs and researchers as one of the primary methods for identifying language disorders and determining eligibility for related services (Kaderavek, 2011). Standardized language tests that

[a]Department of Individual, Family, & Community Education, The University of New Mexico, Albuquerque
[b]Department of Special Education, The University of New Mexico, Albuquerque
[c]Department of Speech and Hearing Sciences, The University of New Mexico, Albuquerque
[d]Learning Sciences Institute, Arizona State University, Tempe
[e]Department of Linguistics and Communication Disorders, Queens College, City University of New York, NY
[f]Department of Medicine, University of Illinois at Chicago

Correspondence to Cathy Huaqing Qi: hqi@unm.edu

All study procedures were approved by the institutional review board at The University of New Mexico.

Editor-in-Chief: Stephen M. Camarata
Editor: Megan York Roberts

Received January 25, 2021
Revision received April 8, 2021
Accepted June 18, 2021
https://doi.org/10.1044/2021_JSLHR-21-00043

__Disclosure:__ The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

demonstrate psychometric rigor in terms of what they measure and how well they measure are crucial for early identification of language disorders and effective intervention. Appropriate and accurate assessment of young children's language skills can help SLPs identify children's strengths and limitations in language learning, inform child-centered interventions, and monitor language progress. On the other hand, using a standardized language test that is inappropriate for children who are at risk for development of language delays or disorders would put them at a disadvantage due to potential misdiagnosis, failure to inform intervention planning, or inaccurate detection of change in language skills over time.

Preschool Language Scales–Fifth Edition

The Preschool Language Scales–Fifth Edition (PLS-5; Zimmerman et al., 2011) is a standardized language test designed to identify children with a language delay or disorder from birth to 7 years, 11 months (Zimmerman et al., 2011). It consists of two standardized scales: Auditory Comprehension (AC; 65 items) and Expressive Communication (EC; 67 items). The PLS-5 has been the most frequently used standardized test by SLPs to assess the language skills of children in early education settings (0–3 years or 3–5 years) to monitor progress of grammatical language (Finestack & Satterlund, 2018). In clinical practice, PLS-5 scores are used to determine whether a language delay or disorder (receptive or expressive or both) exists and whether a child will benefit from a particular speech and language therapy.

The PLS-5 has also been widely used in research where individual differences in language abilities are of interest. For example, PLS-5 scores have been used to measure existing levels or gains in language abilities among children with autism spectrum disorder, those who are deaf or hard-of-hearing (Bourque & Goldstein, 2020; Curiel et al., 2018; D'Agostino et al., 2020; Jones & Lord, 2013; Meinzen-Derr et al., 2014; Nevill et al., 2019; Piper et al., 2020; Vernon et al., 2019), and children with typical development (e.g., Bichay et al., 2020; Julien et al., 2019; Sanchez et al., 2020). PLS-5 scores are often used as predictors of later outcomes, such as language and behavioral development as well as academic performance (Bichay et al., 2020; Janus et al., 2019; Riley et al., 2019; Volpe et al., 2019).

The PLS-5 provides a Total Language composite score, an AC scale score representing a receptive language construct, and an EC scale score indicating an expressive language construct. The two scales are assumed to be distinct dimensions of measurement, though possibly correlated. The Total Language composite score can be interpreted as part of a comprehensive evaluation for determining whether a child has a language delay or disorder and the individual scale standard scores can be further used to identify a receptive or expressive language delay or disorder, or both. However, psychometric properties of the PLS-5, which can provide evidence to support the use of raw scores (i.e., the sum of the passed items) derived from dichotomous responses (i.e., 0 and 1), have been rarely examined. Hence, in this study, we focused on two particularly important psychometric properties: the hierarchy among test items and the reliability of the test scores.

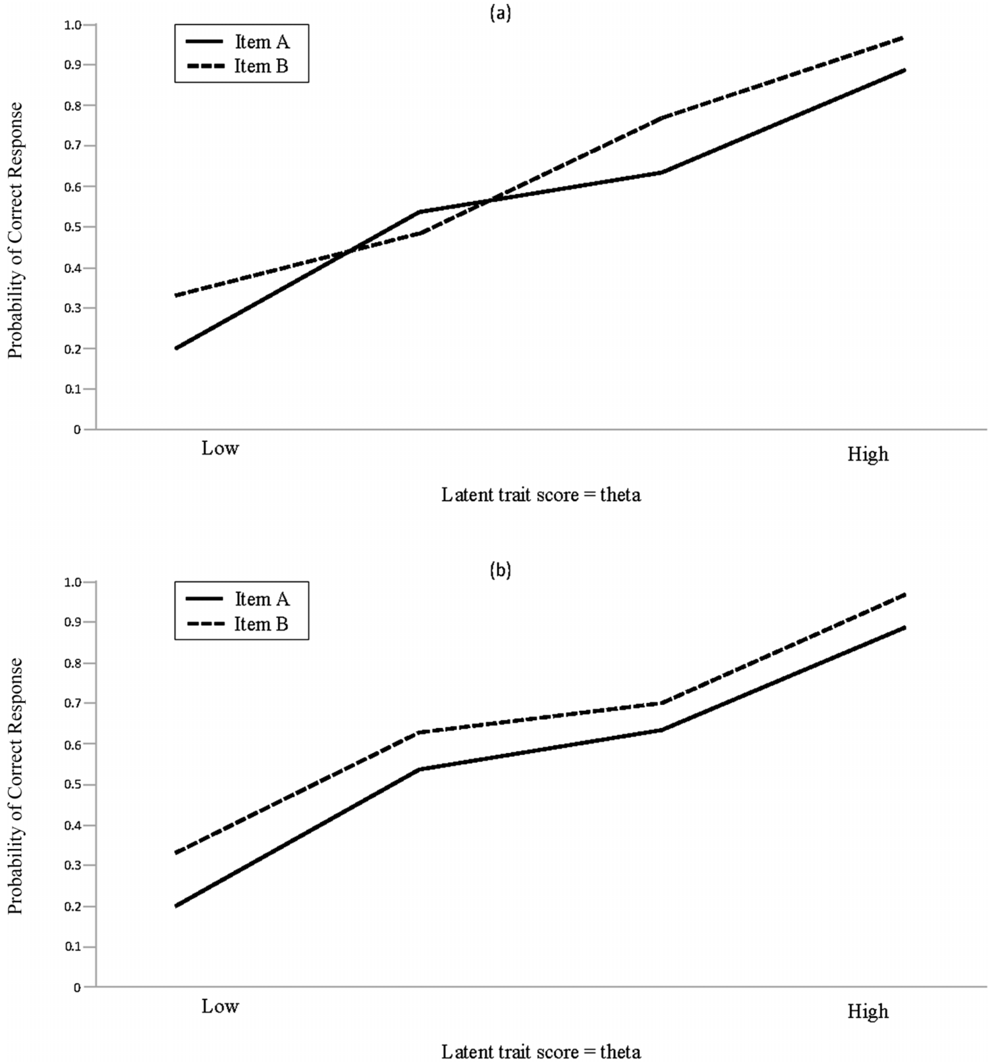
Hierarchy means the items in a test are correctly ordered based on their difficulty along the latent trait (e.g., the language abilities in this study) being measured (Mishra et al., 2011; Watson et al., 2008). This is crucial for the PLS-5; like many other early development measures that use a basal-ceiling administration protocol, test administration stops if a child answers six consecutive items incorrectly. Examination of the hierarchy among test items can shed light on whether the current item order of the PLS-5 is aligned with the difficulty level of each item. Additionally, raw scores from items that satisfy hierarchy have the potential to be used as indicators of children's ordering on the latent trait (Wind, 2017). In other words, if both the AC and EC raw scores are composed of items that satisfy hierarchy, researchers could be more confident in using each child's scale scores at their "face values" to represent the latent construct that the scales claim to measure.

Reliability is defined as the amount of measurement error of the test scores. High reliability values indicate that test scores from the same individuals would be consistent across different occasions (e.g., time points, equivalent test banks, etc.). Reliability estimates are sample dependent, and different estimates will arise from administrations with other populations.

Mokken Scale Analysis

Mokken scale analysis (MSA; Mokken, 1971) is a nonparametric item response theory (IRT) method for measures with dichotomous or polytomous responses, which can assist in examining both the reliability of the test scores and the hierarchy property among test items to ensure the legitimate use of the raw scores of a scale or a test (Stochl et al., 2012). MSA provides a series of tools for examining the relationship between a latent trait (e.g., language ability) and the probability for a particular response on an item. Such a relationship can be depicted via the item characteristic curve or the item response function (IRF; Wind, 2017). Figure 1 shows two pairs of IRFs for two items with binary response (e.g., pass or fail). The *y*-axis shows the probability for a correct response (pass) and the *x*-axis represents the level of the latent variable an individual possesses (theta). Taking the PLS-5 as an example, theta is a measure of a child's language ability at the individual level. Compared to parametric IRT models such as the Rasch models, MSA can test many of the psychometric properties without imposing the strong statistical assumptions of parametric IRT models (Stochl et al., 2012; Watson et al., 2012; Wind, 2016). Given such robustness to statistical assumptions, MSA does not require the substantially large sample sizes typically necessary for parametric IRT models (Straat et al., 2014; Wind, 2017). In addition, MSA can be used to estimate reliability coefficients other than Cronbach's alpha, which often underestimates the true reliability (Agbo, 2010; Stochl et al., 2012).

Figure 1. Item response functions (IRFs) showing a pair of items satisfy monotonicity (a) and monotonicity and invariant item ordering assumptions (b) for item A (solid curve) and item B (dashed curves).



Researchers in many different areas have increasingly used MSA over the past 2 decades in empirical practice such as psychiatry, psychology, social sciences, and health (Meijer et al., 2011; Myszkowski, 2020; Palmgren et al., 2018; Sijtsma & van der Ark, 2016; Tillema et al., 2021). However, MSA is less known in speech and language research than its parametric IRT counterparts or methods of classical testing theories such as exploratory factor analysis and confirmatory factor analysis. Despite the PLS-5 being widely used in research and clinical practice, to our knowledge, this is the first study to use MSA for examining the hierarchy and reliability of the PLS-5 AC and EC scale scores to provide empirical evidence on the interpretations of test scores for its proposed uses with preschool children from low-SES families.

Four assumptions are tested in the framework of MSA: unidimensionality, local independence, monotonicity, and invariant item ordering (Mokken, 1971). Unidimensionality means that there is a single latent trait (e.g., language ability) underlying the response to the test items (e.g., PLS-5 scale items; Price, 2017). For example, if the PLS-5 AC and EC scales meet the unidimensionality assumption, they only measure the child's receptive language ability and expressive language ability, respectively. A scale that violates the unidimensionality assumption might indicate that two or more latent traits influence the response of the scale items.

The second assumption of local independence is defined as an individual's responses to an item being influenced only by the latent trait being measured (e.g., language ability), rather than other test items (Watson et al., 2012). Hence, once the effect of the latent trait is controlled, there would be no observed associations between a response to any specific item and the responses to other items on the latent trait in the same test. If data violate the local independence assumption, the estimate of the IRF is likely to be biased (Wainer & Wang, 2005).

The third assumption of monotonicity states that the probability of a correct response is nondecreasing across increasing locations on the latent trait—as a child's language ability increases, the probability of answering the item correctly should also increase or remain at the same level but cannot decrease over the range of the latent trait (Sijtsma & van der Ark, 2016). Figure 1a shows a pair of IRFs (Item A and Item B) that satisfies the monotonicity assumption. Although an intersect can be observed between the two IRFs for Item A and Item B, both IRFs increase or stay level on the latent trait (i.e., language scores). Violating this assumption would render the scores unusable to infer the ordering of children.

For a scale that does not violate the above three MSA assumptions (i.e., unidimensionality, local independence, and monotonicity), data are considered adequately fit by the monotone homogeneity (MH) model (Mokken, 1997). For a scale fitting the MH model, the raw score of the scale can be used to order individuals on their latent trait (Grayson, 1988; van der Heijden et al., 2003). Taking the PLS-5 as an example, if the AC or EC scale meets all the three assumptions mentioned above, a child's AC or EC raw scores can be used to represent their standing in terms of receptive and expressive language abilities, respectively.

The fourth assumption is invariant item ordering (IIO). IIO is satisfied if the IRF for each item does not intersect with any other items in the scale. In other words, a test taker's probability of answering one difficult item correctly is lower than the probability of answering an easier item correctly, regardless of the level of the latent trait that the test taker possesses. As mentioned earlier, IIO is particularly important for a scale like the PLS-5, where answering six consecutive items incorrectly results in stopping the test administration. Hence, test items confirmed to be ordered by their difficulty levels can assure that test takers at any level of the latent trait would not be expected to get any future, that is, questions later in the sequence, correctly if they were actually given after the stopping point is achieved.

A scale that meets all four MSA assumptions can be described as fitting the double monotone (DM) model. While the MH model is a model for ordering persons on their latent ability, the DM model is a model for ordering items "by means of mean item scores" in addition to person ordering (Sijtsma & van der Ark, 2016, p. 142). The DM model implies that the item ordering is equal for test takers of different abilities (Sijtsma & van der Ark, 2016). Violating the DM assumption would make test scores less meaningful as "the difficulty order of the items is different for different people" (van der Heijden et al., 2003, p. 197). Figure 1b shows a pair of IRFs that satisfies the IIO assumption: the two IRFs do not intersect, regardless of the ability of each test taker. If the PLS-5 AC or EC items fit the DM model, the scale can be used both to compare individuals in reference to their ability levels and to compare the difficulty levels of items on the scale.

The purpose of this study was to apply MSA to examine the characteristics of hierarchy of the PLS-5 AC and EC scales among test items and to estimate reliability of the scores of the AC and EC scales. We examined the PLS-5 for use with preschool children from low-SES families as they are at risk of language delay due to poverty (Qi et al., 2003; Hart & Risley, 1995). Specific research questions are:

1. How do the PLS-5 AC and EC scale items perform in terms of MSA assumptions (i.e., unidimensionality, local independence, monotonicity, and IIO)?
2. To what extent is the scale score (i.e., AC or EC) consistent based on the reliability analysis using MSA?
3. Can results from (1) and (2) warrant the use of raw scores to indicate individuals' language abilities on both scales? Can these results provide support of items being ordered based on their difficulty levels?

Method

Participants

Participants for this study were 169 preschool children (M_{age} = 44 months, *SD* = 3.6 months, range: 37–54 months; 47% females) enrolled in Head Start programs in a mid-sized city in the southwestern United States. Head Start programs serve preschool children from low-SES families by providing education, health, nutrition, social services, and other services (U.S. Department of Health & Human Services, 2019). Data for this study were collected as part of a longitudinal project examining the relationships among language, behavioral problems, and social skills of preschool children. A total of 44 (26.0%) children were between 36 and 41 months old, 89 (52.7%) were between 42 and 47 months old, 35 (20.7%) were between 48 and 53 months old, and one (.6%) was between 54 and 59 months old. To be eligible for the larger study, children (a) were expected to continue in Head Start programs the following year, (b) spoke primarily English, and (c) were not receiving special education services except for speech or language impairments. The ethnic composition was 65% Hispanic, 7% White Non-Hispanic, 4% Native American, 4% African American, 2% Asian, and 18% in a category designated as "Other."

Measure

The PLS-5 (Zimmerman et al., 2011) assesses receptive and expressive language abilities of children from birth to the age of 7 years and 11 months. As stated in the PLS-5 Examiner's Manual, the test–retest reliability for the age group of 3:0 to 4:11 ([years;months] 3 to 4 years and 11 months old) was .90 for AC, .91 for EC, and .93 for Total Language. The split-half reliability was reported based on the Spearman-Brown formula to estimate internal consistency, which ranged from .91 to .93 for AC, and .94 to .95 for EC for the same age group of the total normative sample. The correlation between the AC and EC scales was .75

(Zimmerman et al., 2011). Many items in both scales contain subitems. For example, if an EC item tests the use of plurals, the tester may have asked the child "What are these?" and there may have been up to three subitems. A child who correctly answered at least two subitems would receive a score of 1. Each child's responses were scored in accordance with the PLS-5 Examiner's Manual. The format of the PLS-5 is performance based with a dichotomous scoring method (score 0 for incorrect responses and 1 for correct responses). Higher scores suggest a greater level of language ability; lower scores, based on suggested thresholds, can be used to indicate a language disorder.

Procedure

All test assessors completed extensive training in the administration of the PLS-5. Test administration took approximately 40 to 60 min. Trained graduate students enrolled in either a speech-language pathology program or a special education program administered the PLS-5 individually to each participant at their Head Start centers. To avoid fatigue, many breaks were given to children. The AC and EC scales were administered in counterbalanced order. Some children took the AC scale first, while others took the EC scale first. The recommended administration protocol was followed. The Examiner's Manual specifies a starting point (basal) after three consecutive items are answered correctly and a stopping point (ceiling) after six consecutive items are answered incorrectly. Each item was administered sequentially until a ceiling was reached.

Data Analysis

We used R 4.0.2 (R Core Team, 2020) and the Mokken package in R (van der Ark, 2007, 2012) to conduct MSA for the items on the PLS-5 AC and EC scales separately. We followed the procedures recommended by Sijtsma & van der Ark (2016) and Wind (2017). Four MSA methods: scalability, local independence, monotonicity, and IIO were used to test the MSA assumptions: unidimensionality, local independence, monotonicity, and IIO, respectively. Previous MSA literature indicated that current MSA methods are limited in testing the unidimensionality assumption (Smits et al., 2012). Hence, we followed the suggestion of Wind (2017) and used both the scalability and monotonicity analyses available in MSA to inform us about the unidimensionality of the AC and EC scales. For the other three MSA assumptions (i.e., monotonicity, local independence, and IIO), we conducted the corresponding MSA analyses to examine them.

Scalability

Indicators of scalability are quantified by estimating the amount of Guttman errors—passing a difficult item but failing an easier item—among test takers. According to van der Ark (2012), an item scalability coefficient, a test scalability coefficient, and an item pair scalability coefficient all "play an important role in MSA" (p. 5–6).

We evaluated the scalability of the PLS-5 AC and EC scale items by estimating Loevinger's coefficient H (Meijer et al., 1995) using the *CoefH* function. Loevinger's coefficient H, a value between 0 and 1, was designed for testing the level of scalability for each item (H_i), the entire set of items to form the subscales (H_{total}), and any pair of items (H_{ij}) under the AC and the EC scales. A scale which is composed of items with high H values is considered highly scalable and has the potential to satisfy the unidimensionality assumption. Based on Mokken (1971) and Molenaar and Sijtsma's (2000) suggestions, $.50 \leq H < 1.00$ indicates a strong scale, $.40 \leq H < .50$ indicates a moderate scale, $.30 \leq H < .40$ indicates a weak scale, and $H < .30$ is considered unscalable. Additionally, items with negative H values are flagged as problematic items.

For a scale that contains problematic items and/or item pairs flagged by the H indices, the automated item selection procedure (AISP) from the Mokken package in R (Mokken, 1971; R Core Team, 2020; Sijtsma & Molenaar, 2002) can be used to check for any potential subtrait that indicates the need to form another scale. Additionally, the AISP can further identify deviating items that may not contribute enough to form a strong Mokken scale (Sijtsma & van der Ark, 2016). Those deviating items are considered unscalable. Through AISP, a constant, c, is set up as the cutoff point for excluding items or assigning items to a new subscale, based on individual items' H values. For both the AC and EC scale items, AISP was run 12 times from $c = 0$ to .55 with an increment of .05, as suggested by Hemker et al. (1995), using the function *aisp* with the genetic algorithm (Stratt et al., 2013).

Local Independence

Local independence was examined by using the *check.ca* function to estimate the conditional associations among items (Straat et al., 2016). Three indices, namely, W_1, W_2, and W_3, were computed to detect items violating this assumption. W_1 indicates the likelihood of an item pair being positively locally dependent. W_2 indicates the likelihood of an item being positively locally dependent with any other item. W_3 indicates the likelihood of an item pair being negatively locally dependent. The *check.ca* function reported both the W_i values and flagged the problematic items and item pairs violating the local independence assumption.

Monotonicity

To assess the monotonicity of items, we used the function *check.monotonicity* to plot the IRF (Wind, 2017) for each item and check for nondecreasing patterns over increasing values of the latent trait score. The function also provided Z tests (see Molenaar & Sijtsma, 2000) results for detecting significant violations of the monotonicity assumption for each item. The number of violations were reported through the function. Results from the monotonicity analyses were used to inform both the unidimensionality and the monotonicity assumptions (Wind, 2017).

IIO

We used the *check.restscore* function to test the assumption of IIO. The rest score method was implemented to test any intersect between any pair of IRFs (Hemker et al., 1997; Junker, 1993, Junker & Sijtsma, 2000; Sijitsma & Molenaar, 2002). A *t* test was then conducted to detect significant violations of IIO assumptions between any item pair. The number of significant violations were reported.

Reliability

We used the *check.reliability* function to calculate two types of reliability coefficients. The first reliability coefficients to be calculated were the internal consistency coefficients (e.g., coefficient alpha, α; Cronbach, 1951) of both the PLS-5 AC and EC test scores. The second type of reliability coefficients to be computed were the latent class reliability coefficients (LCRC) of the AC and the EC scales scores (see van der Ark et al., 2011, for a review of LCRC). The LCRC reliability is robust to violations of the assumptions underlying the DM models and has been shown to be superior to Cronbach's alpha, which is often an underestimation of the true reliability (van der Ark et al., 2011).

Results

Descriptive Analyses and Data Examination

The children's mean standard score was 94.00 (*SD* = 14.4) for the AC scale, 90.80 (*SD* = 12.58) for the EC scale, and 92.02 (*SD* = 12.08) for the Total Language scale. There were no missing observations among the 169 children. For both the AC and EC scales, items with zero variance (i.e., scored 0 or scored 1 for all children) were AC 62, AC 63, AC 64, AC 65, and EC 67. Such items along with those before the start point (i.e., AC 1 to AC 26 and EC 1 to EC 26) were excluded from the following analyses since they were not suitable for MSA. Overall, the remaining 35 AC items and 40 EC items were included in MSA (see Table 1).

Scalability

AC Scale

The PLS-5 AC scale (35 items) had medium scalability, with a scalability coefficient of H_{total} of .48 (SE = .04). Based on the 95% confidence interval (CI) for this estimate (H_{total} ± 1.96*SE(H_{total}) = [.40, .56]), the AC scale could be considered to be a moderate (.40 ≤ H < .50) to strong Mokken scale (H > .50).

Individual item scalability coefficients (*Hi*) and their corresponding standard errors (*SE*) for the AC scale are presented in Table 1. Examination of these coefficients revealed that the scalability of each item on the AC scale was above Mokken (1971) minimum value of H_i = .30, ranging from the least scalable item (AC 43), H_i = .33 (*SE* = .07), to the most scalable item (AC 30), H_i = .74 (*SE* = .08).

However, six item pair scalability coefficients (*Hij*) were negative for AC items, indicating that some of the items were problematic. Hence, we conducted a series of AISPs to identify which items deviated from the AC scale, given prespecified cutoff values, *c*, for the items' *H* coefficients. We ran AISP with *c* = 0, .05, .10, …, .30, and found only one scale was formed, which contained all 35 AC items. For *c* = .35, after two items (AC 44 and AC 49) that fell out of the scale were excluded, two scales were formed. For *c* = .40, five items (AC 40, AC 43, AC 44, AC 49, and AC 51) were excluded from the scale and were unscalable. For *c* > = .45, the AISP produced more than two scales and several unscalable items. Thus, we found the AC items formed a moderate strength Mokken scale (*H* > .40), after leaving out five items (i.e., AC 40, AC 43, AC 44, AC 49, AC 51).

Table 1. Scales, item scalability coefficients (*Hi*), standard errors (*SE*), and invariant item ordering (IIO) for Items on the PLS-5 AC and EC Scales.

	Auditory Comprehension			Expressive Communication		
Item #	**Item H_i**	***SE***	**IIO**	**Item H_i**	***SE***	**IIO**
27	.56	.01	0	.73	.07	0
28	.63	.11	0	.67	.07	0
29	.59	.14	0	.74	.06	0
30	.74	.08	0	.84	.07	0
31	.52	.08	0	.82	.06	0
32	.50	.08	0	.79	.04	0
33	.42	.11	0	.76	.05	0
34	.47	.07	0	.60	.07	0
35	.57	.05	0	.60	.07	0
36	.57	.05	1	.64	.06	0
37	.48	.07	1	.61	.06	1
38	.51	.06	0	.71	.05	0
39	.52	.06	0	.56	.06	0
40	.37	.06	1	.53	.06	2
41	.50	.05	3	.63	.05	1
42	.45	.06	1	.61	.08	0
43	.33	.07	2	.64	.05	0
44	.36	.06	1	.60	.08	0
45	.50	.06	0	.60	.06	0
46	.46	.05	0	.65	.08	0
47	.40	.07	0	.61	.06	0
48	.40	.07	0	.77	.09	0
49	.41	.07	1	.63	.08	0
50	.40	.08	0	.62	.07	0
51	.40	.10	0	.75	.06	0
52	.49	.16	0	.81	.09	0
53	.60	.05	1	.80	.07	0
54	.49	.07	0	1.00	.00	0
55	.60	.09	0	.77	.07	0
56	.56	.05	0	.76	.07	0
57	.55	.06	0	.91	.07	0
58	.45	.12	0	1.00	.00	0
59	.55	.07	0	.97	.08	0
60	.49	.01	0	1.00	.00	0
61	.43	.03	0	1.00	.00	0
62	NA	NA	NA	.96	.03	0
63	NA	NA	NA	.91	.08	0
64	NA	NA	NA	.87	.08	0
65	NA	NA	NA	.91	.08	0
66	NA	NA	NA	1.00	.00	0

Note. PLS-5 = Preschool Language Scales–Fifth Edition; AC = Auditory Comprehension; EC = Expressive Communication; NA = not applicable.

After removing these five problematic AC items, no negative *H* coefficients were found. The PLS-5 AC scale

(30 items) had strong scalability, with a scalability coefficient of H_{total} of .54 (SE = .04). Based on the 95% CI for this estimate ($H_{total} \pm 1.96 * SE(H_{total})$ = [.46, .62]), the AC scale could be considered a moderate-to-strong Mokken scale.

EC Scale

The PLS-5 EC scale (40 items) had strong scalability, with a scalability coefficient of H_{total} of .70 (SE = .05). Based on the 95% CI for this estimate ($H_{total} \pm 1.96 * SE(H_{total})$ = [.60, .80]), the EC scale could be considered a strong Mokken scale ($H > .50$). The 95% CI [.60, .80] for this estimate indicated a strong Mokken scale ($H > .50$). The Hi for individual items ranged from .53 (SE = .06) for EC 40 to 1.00 (SE = 0.00) for EC 54, EC 58, EC 60, EC 61, and EC 66, all above the .30 criterion of scalability. No negative scalability coefficients were observed for any of the EC items pairs. Given the evidence of a strong Mokken scale, no AISP was conducted for EC items.

Local Independence

For the AC scale (35 items), the conditional association procedure flagged three (0.5%) item pairs among the 595 item pairs to be locally dependent, suggesting that after controlling for the effect of the receptive latent ability, there were other factors that may influence children's responses to these items. Specifically, the W_1 index identified one positive local dependency (AC 33 and AC 39); the W_3 index suggested negative local dependencies between AC 40 and AC 42 and between AC 40 and AC 43. No local dependence was detected by the W_2 index. Overall, less than 1% of the item pairs in the AC scale violated the local independence assumption. Hence, based on our evaluation, we concluded that the AC scale has satisfactory local independence.

For the EC scale (40 items), the conditional association procedure showed that among the 780 item pairs, nine (1.15%) item pairs were suspected to be locally dependent. Specifically, the W_1 index identified one positive local independency (EC 28 and EC 29); the W_3 index suggested negative local dependencies for eight item pairs (i.e., EC 36 and EC 39, EC 36 and EC 43, EC 37 and EC 45, EC 40 and EC 43, EC 40 and EC 50, EC 41 and EC 44, EC 44 and EC 45, and EC 45 and EC 47). No local dependence was detected by the W_2 index. Overall, less than 2% of the item pairs in the EC scale violated the local independence assumption. Consistent with the results of the AC scale, we concluded that the EC scale satisfied the local independence assumption.

Monotonicity

Results from the IRFs of the AC and EC items revealed no violations of monotonicity in reference to the Z test results. Inspection of the IRFs showed monotonically nondecreasing patterns in IRFs of all items on each scale (see Figure 1 for example of IRFs that satisfy the monotonicity assumption). Hence, the relative ordering of children in terms of their receptive and expressive language abilities as measured on the AC and EC scales in each item was consistent across all the AC and EC items.

IIO

Table 1 presents the results from the IIO analyses of the AC and EC items using the rest score method. For each item, the frequency of significant violations of IIO is presented. For AC items, 25.71% (9 out of 35) had at least one violation of IIO. Specifically, one significant violation was observed for seven items (AC 36, AC 37, AC 40, AC 42, AC 44, AC49, AC 53), two significant violations were observed for AC 43, and three violations of IIO were observed for AC 41. For EC items, only 7.5% (3 out of 40) items had at least one violation of IIO. Specifically, one significant violation was detected for EC 37 and EC 41, and two significant violations were observed for EC 40. Results of IIO indicated that for children with the same scale scores, their performance on these IIO-violated items might be different. This is especially the case for the AC scale.

Reliability

The Cronbach's reliability coefficient was .87 for the PLS-5 AC scale scores and .93 for the EC scale scores. The LCRC estimated from the AC scale scores for our sample was .94 and from the EC scale score for our sample was .96. The reliability estimates yielded from the LCRC indicated high consistency across conditions. The LCRC reliability coefficients suggested that both the AC and EC scales had satisfactory reliability for both basic research and clinical applications.

Discussion

The purpose of this study was to use MSA to evaluate the hierarchy among items and the reliability of the scale scores of the PLS-5 AC and EC scales among a racially-ethnically diverse sample of preschool children from low-SES backgrounds. These psychometric properties were examined to determine whether the use of raw scores of the AC and EC scales were justified. We also examined whether the items are ordered by increasing difficulty regardless of the test takers' language ability. Overall, except for a few problematic items, we found that the EC scale generally meets the four MSA assumptions in the present sample. Along with the high reliability estimates of the EC scale scores, we conclude that the EC scale is appropriate for the use of raw scores to compare children's expressive language skills with this population. On the other hand, large numbers of violations of several MSA assumptions were found for the AC scale. Researchers and SLPs should be cautious when using the AC scale scores to interpret receptive language abilities of children from low-SES backgrounds. Additionally, there are some AC and EC items that violated the IIO assumption, indicating that there is not an ordering that would satisfy the IIO assumption.

Both the AC and EC items revealed no violations of monotonicity. This finding suggests that the relative ordering of children in terms of their language abilities was consistent across the AC or EC items. Compared with a child with lower language ability, a child with higher language ability will have a higher probability of correctly answering any AC or EC item from the scale.

Five AC items (AC 40, AC 43, AC 44, AC 49, AC 51) were flagged as unscalable in a moderate Mokken scale. Among them, we found two items (AC 40 and AC 43) that showed local dependency with another item (e.g., AC 40 and AC 42, and AC 40 and AC 43). Previous literature indicates that local dependence can be caused by item chains (Balazs & de Boeck, 2007), repeated test-taking practice, or test taker's fatigue (Stochl et al., 2012). However, these problematic items are not adjacent items, and practice or fatigue cannot explain why only a few items are influenced. We further examined the purpose of these items—AC 40 is designed to test pronouns, AC 42 is to identify shapes, and AC 43 is used for alphabetic letter recognition—and found no common trait other than receptive language ability that might influence the response to these items. Based on the MSA results, it is unclear to us whether there is another dominant trait occurring among AC items. Further examination is needed to determine whether other unknown extraneous variables may contribute to the responses to AC items.

On the other hand, scalability results suggested that all the PLS-5 EC items form a moderate Mokken scale, indicating that EC items are meaningful in ordering children based on their expressive language abilities as measured by the EC scale. Along with findings from monotonicity and local independence, we argue that EC scale raw scores are an appropriate measurement of children's expressive language ability.

Item invariant ordering allows test items to be ordered according to their difficulty level independent of children's ability level and helps researchers or clinicians interpret outcome measurement in terms of tasks administered at different levels of difficulty (Stochl et al., 2012). It is worth noting that the PLS-5 AC Item 41 had three significant violations of IIO. This item checked a child's understanding of the concepts of *more* and *most*, which is in the realm of early numeracy skills. These findings suggest that items such as AC 44 (testing a child's ability to identify body parts) might be harder than AC 41 for some children, but easier than AC 41 for other children, depending on the individual child's overall ability level. Overall, we found nine AC items and three EC items violated the IIO assumptions. Results from IIO indicate a discrepancy between performance on these problematic individual items and scale scores. Researchers have recommended that speech and language clinicians should not develop language goals and interventions based on children's responses to specific items on the test (Merrell & Plante, 1997; Plante & Vance, 1994). The PLS-5 is designed to diagnose language disorder, not to identify particularly difficult items. Hence, clinicians should not determine a child's language ability based on a few correct or incorrect responses (Epstein, 2018).

In terms of reliability, we found that the PLS-5 demonstrates good test score reliability evidence for preschool children from low-SES families. Our results showed that the LCRC yielded higher reliability estimates under MSA than the conventional Cronbach's alpha. Cronbach's alpha has been shown to be the lower bound of reliability (McNeish, 2018) and can be biased for ordinal data (Yang & Green, 2011). The LCRC reliability estimates for both the PLS-5 AC and EC scale scores were sufficient for basic research and group comparisons (> .80, Bland & Altman, 1997; Nunnally, 1994) as well as for clinical practice (e.g., identification of a language delay or disorder and determination of eligibility; > .90, Molenaar, 1997; Mokken & Lewis, 1982). The high LCRC reliability estimates for both the PLS-5 AC and EC scale scores in a sample of children from low-SES backgrounds in our study make the scales appropriate for comparing language abilities of this population for research purposes and for being used as part of assessment battery to identify a language delay or language disorder for clinical practices.

Some MSA application studies have used this technique to remove misfit items from the scale. However, removing items may alter the construct validity of the scale composed of the remaining items (Watson et al., 2012). We agree with this concern and argue that deleting items based on a few indicators from the whole test may change the dynamic of the test. Additionally, the problematic items account for only a small portion of the entire item banks. Whether such a small portion of misfit items would have a significant impact on the use of the raw scores remains unknown. Although several problematic items are identified through the MSA procedure, we do not advocate excluding some of these items from the scale scores calculation. Future research is needed to examine the reasons for the number of items on the AC scale failing to load onto a common factor and whether these items need to be removed.

Limitations of the Study

There are several limitations of our study that warrant comments. First, due to the relatively small sample size, MSA was used in this study, as that is one of its strengths. However, even in MSA, limited sample size may decrease the accuracy of estimated parameters. Additionally, the participants were all 3–4 years old, making the results less generalizable to younger or older children. The study would merit replication with a larger sample size with more diverse age groups and a parametric IRT model so that more information can be provided for evaluating the psychometric properties of the PLS-5.

The second limitation lies in the appropriateness of using MSA to test unidimensionality. Applying the scalability analysis and the AISP alone used to be considered a standard procedure in examining unidimensionality in MSA. However, a recent simulation study (Smits et al., 2012) has shown that using those procedures in MSA may result in inconsistent findings of dimensionality. Acknowledging such critiques, we used results from both scalability and monotonicity analyses to examine unidimensionality, as suggested by Wind (2017). However, it is worth noting that there are other nonparametric dimensionality assessment methods that are not in the realm of MSA that can be used (see Smits et al., 2012). Under large sample size conditions, researchers may want to apply alternative parametric dimensionality methods. For such application in language testing, readers can refer to Anthony et al. (2014), Tomblin and Zhang (2006), and Language and Reading Research Consortium (2015).

Finally, due to the design of the PLS-5 (Zimmerman et al., 2011), not every item was administered (i.e., basal or ceiling items) or analyzed (i.e., items with zero variance). Hence, we could not conduct a more thorough examination of the PLS-5 using a multidimensional IRT (MIRT) model (Hartig & Höhler, 2009; see Reckase, 2009, for an extensive review). An MIRT can simultaneously address multiple skills, for example syntax and semantics in the analysis. Future research may use a two-parameter logistic (2PL)-MIRT model, which is the extension of the 2PL unidimensional IRT model (Bandalos, 2018), with a larger sample size, to address this issue.

Conclusions

As a new application of MSA in the field of the language and speech research, we tested four underlying MSA assumptions on the PLS-5. Based on our results, we conclude that the PLS-5 EC scale follows the four MSA assumptions and the scale scores yielded high reliabilities. Hence, it is appropriate to use raw scores of the EC scale to assess individuals' expressive language abilities among predominantly Hispanic preschool children from low-SES backgrounds. However, the raw scores of the PLS-5 AC scale should be used with caution, given many violations in the MSA assumptions. Several IIO-violated items are identified in both the AC and the EC scales, which implies that researchers and clinicians should not evaluate the details of a child's language ability based on responses to particular items, as children with the same sum scale scores might have different patterns of mastery of test items.

Author Contributions

Yu-Yu Hsiao: Formal analysis (Lead), Investigation (Equal), Methodology (Equal), Writing – original draft (Equal), Writing – review & editing (Equal). **Cathy Huaqing Qi:** Conceptualization (Lead), Formal analysis (Supporting), Funding acquisition (Lead), Investigation (Lead), Methodology (Equal), Project administration (Lead), Resources (Lead), Supervision (Lead), Writing – original draft (Equal), Writing – review & editing (Equal). **Robert Hoy:** Conceptualization (Equal), Formal analysis (Supporting), Methodology (Supporting), Writing – original draft (Supporting). **Philip S. Dale:** Conceptualization (Equal), Investigation (Equal), Writing – original draft (Supporting), Writing – review & editing (Supporting). **Glenda S. Stump:** Methodology (Supporting), Writing – review & editing (Equal). **Megan Dunn Davison:** Writing – original draft (Supporting), Writing – review & editing (Equal). **Yinglin Xia:** Writing – review & editing (Supporting)

Acknowledgments

This research was supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development through Exploratory/Development Grant (1R21 HD069759-01A1, PI: second author).

References

Agbo, A. A. (2010). Cronbach's alpha: Review of limitations and associated recommendations. *Journal of Psychology in Africa, 20*(2), 233–239. https://doi.org/10.1080/14330237.2010.10820371

Anthony, J. L., Davis, C., Williams, J. M., & Anthony, T. I. (2014). Preschoolers' oral language abilities: A multilevel examination of dimensionality. *Learning and Individual Differences, 35,* 56–61. https://doi.org/10.1016/j.lindif.2014.07.004

Balazs, K., & de Boeck, P. (2007). *Detecting local item dependence stemming from minor dimensions*. Interuniversity Attraction Pole Statistics Network. http://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseV/publications_2006/TR/TR0684.pdf

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* Guilford.

Bichay, K., Qi, C. H., Bulotsky-Shearer, R., & Carta, J. (2020). Bidirectional relationship between language skills and behavior problems in preschool children from low-income families. *Journal of Emotional and Behavioral Disorders, 28*(2), 114–128. https://doi.org/10.1177/1063426619853535

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal, 314*(7080), 570–572. http://doi.org/10.1136/bmj.314.7080.572

Bourque, K. S., & Goldstein, H. (2020). Expanding communication modalities and functions for preschoolers with autism spectrum disorder: Secondary analysis of a peer partner speech-generating device intervention. *Journal of Speech, Language, and Hearing Research, 63*(1), 190–205. https://doi.org/10.1044/2019_JSLHR-19-00202

Catts, H. W., & Kamhi, A. G. (2012). *Language and reading disabilities* (3rd ed.). Allyn & Bacon.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Curiel, E. S. L., Sainato, D. M., & Goldstein, H. (2018). Matrix training for toddlers with autism spectrum disorder and other language delays. *Journal of Early Intervention, 40*(3), 268–284. https://doi.org/10.1177/1053815118788060

D'Agostino, S., Douglas, S. N., & Horton, E. (2020). Inclusive preschool practitioners' implementation of naturalistic developmental behavioral intervention using telehealth training. *Journal of Autism and Developmental Disorders, 50*(3), 864–880. https://doi.org/10.1007/s10803-019-04319-z

Epstein, B. (2018). Psychometrics for speech and language assessment: Principles and pitfalls. In S.-R. Cyndi & F. Renee (Eds.), *A guide to clinical assessment and professional report writing in speech-language pathology* (pp. 65–90). Slack.

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27*(4), 1329–1351. https://doi.org/10.1044/2018_AJSLP-17-0168

Grayson, D. A. (1988). Limitations on the use of scales in psychiatric research. *Australian and New Zealand Journal of Psychiatry, 22*(1), 99–108. https://doi.org/10.3109/00048678809158947

Hammer, C. S., Farkas, G., & Maczuga, S. (2010). The language and literacy development of Head Start children: A study using the family and child experiences survey database. *Language, Speech, and Hearing Services in Schools, 41*(1), 70–83. https://doi.org/10.1044/0161-1461

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Brookes.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2–3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*(4), 337–352. https://doi.org/10.1177/014662169501900404

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*(3), 331–347. https://doi.org/10.1007/BF02294555

Janus, M., Labonté, C., Kirkpatrick, R., Davies, S., & Duku, E. (2019). The impact of speech and language problems in kindergarten on academic learning and special education status in grade three. *International Journal of Speech-Language Pathology, 21*(1), 75–88. https://doi.org/10.1080/17549507.2017.1381164

Jones, R. M., & Lord, C. (2013). Diagnosing autism in neurobiological research studies. *Behavioural Brain Research, 251,* 113–124. https://doi.org/10.1016/j.bbr.2012.10.037

Julien, H. M., Finestack, L. H., & Reichle, J. (2019). Requests for communication repair produced by typically developing preschool-age children. *Journal of Speech, Language, and Hearing Research, 62*(6), 1823–1838. https://doi.org/10.1044/2019_JSLHR-L-18-0402

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annuals of Statistics, 21*(3), 1359–1378. https://doi.org/10.1214/aos/1176349262

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65–81. https://doi.org/10.1177/01466216000241004

Kaderavek, J. N. (2011). *Language disorders in children: Fundamental concepts of assessment and intervention*. Pearson.

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development, 86*(6), 1948–1965. https://doi.org/10.1111/cdev.12450

Letts, C., Edwards, S., Sinka, I., Schaefer, B., & Gibbons, W. (2013). Socio-economic status and language acquisition: Children's performance on the new Reynell Developmental Language Scales. *International Journal of Language & Communication Disorders, 48*(2), 131–143. https://doi.org/10.1111/1460-6984.12004

Levine, D., Pace, A., Luo, R., Hirsh-Pasek, K., Michnick Golinkoff, R., de Villiers, J., Iglesias, A., & Wilson, M. S. (2020). Evaluating socioeconomic gaps in preschoolers' vocabulary, syntax and language process skills with the Quick Interactive Language Screener (QUILS). *Early Childhood Research Quarterly, 50*(Pt. 1), 114–128. https://doi.org/10.1016/j.ecresq.2018.11.006

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. http://doi.org/10.1037/met0000144

Meijer, R. R., de Vries, R. M., & van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory-18 using item response theory: Which items are most strongly related to psychological distress? *Psychological Assessment, 23*(1), 193–202. https://doi.org/10.1037/a0021292

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT Model. *Applied Psychological Measurement, 19*(4), 323–335. https://doi.org/10.1177/014662169501900402

Meinzen-Derr, J., Wiley, S., Grether, S. M., Phillips, J. M., Choo, D., Hibner, J., & Barnard, H. (2014). Functional communication of children who are deaf or hard-of-hearing. *Journal of Developmental & Behavioral Pediatrics, 35*(3), 197–206. http://doi.org/10.1097/DBP.0000000000000048

Merrell, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*(1), 50–58. https://doi.org/10.1044/0161-1461.2801.50

Mishra, G. D., Gale, C. R., Sayer, A. A., Cooper, C., Dennison, E. M., Whalley, L. J., Craig, L., Kuh, D., Deary, I. J., & The HALCyon Study Team. (2011). How useful are the SF-36 subscales in older people? Mokken scaling of data from the HALCyon programme. *Quality of Life Research, 20*(7), 1005–1010. https://doi.org/10.1007/s11136-010-9838-7

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). Springer. https://doi.org/10.1007/978-1-4757-2691-6_21

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows* [Computer software]. IEC ProGAMMA.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter. https://doi.org/10.1515/9783110813203

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6_20

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(4), 417–430. https://doi.org/10.1177/014662168200600404

Myszkowski, N. (2020). A Mokken scale analysis of the last series of the standard progressive matrices (SPM-LS). *Journal of Intelligence, 8*(2), 22. http://doi.org/10.3390/jintelligence8020022

National Academies of Sciences, Engineering, and Medicine. (2016). *Speech and language disorders in children: Implications for the social security administration's supplemental security income program*. National Academies Press. https://doi.org/10.17226/21872

Nelson, K. E., Welsh, J. A., Trup, E. M. V., & Greenberg, M. T. (2011). Language delays of impoverished preschool children in relation to early academic and emotion recognition skills. *First Language, 31*(2), 164–194. https://doi.org/10.1177/0142723710391887

Nevill, R., Hedley, D., Uljarević, M., Sahin, E., Zadek, J., Butter, E., & Mulick, J. A. (2019). Language profiles in young children with autism spectrum disorder: A community sample using multiple assessment instruments. *Autism, 23*(1), 141–153. https://doi.org/10.1177/1362361317726245

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., & Pickles, A. (2016). Younger children experience lower levels of language competence and academic progress in the first year of school: Evidence from a population study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 57*(1), 65–73. https://doi.org/10.1111/jcpp.12431

Nunnally, J. C. (1994). *Psychometric theory* (3rd ed.). McGraw Hill.

Oreopoulos, P., & Salvanes, K. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives, 25*(1), 159–184. https://doi.org/10.1257/jep.25.1.159

Palmgren, P. J., Brodin, U., Nilsson, G. H., Watson, R., & Stenfors, T. (2018). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis—A pragmatic approach. *BMC Medical Education, 18*(1), 235. http://doi.org/10.1186/s12909-018-1334-8

Piper, A., Borrero, J. C., & Becraft, J. L. (2020). Differential reinforcement-of-low-rate procedures: A systematic replication with students with autism spectrum disorder. *Journal of Applied Behavior Analysis, 53*(2), 1058–1070. https://doi.org/10.1002/jaba.631

Plante, E., & Vance, R. (1994). Selection of preschool language tests. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24. https://doi.org/10.1044/0161-1461.2501.15

Price, L. R. (2017). *Psychometric methods: Theory into practice*. Guilford.

Qi, C. H., Kaiser, A. P., Milan, S., Yzquierdo, Z., & Hancock, T. (2003). The performance of low-income African American children on the Preschool Language Scales-3. *Journal of Speech, Language, and Hearing Research, 43*(3), 576–590. https://doi.org/10.1044/1092-4388(2003/046)

Qi, C. H., van Horn, M. L., Selig, J., & Kaiser, A. P. (2020). Relations between language skills and problem behaviour in preschool children. *Early Child Development and Care, 19*(6), 2493–2504. https://doi.org/10.1080/03004430.2019.1588264

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rantalainen, K., Paavola-Ruotsalainen, L., Alakortes, J., Carter, A. S., Ebeling, H. E., & Kunnari, S. (2021). Early vocabulary development: Relationships with prelinguistic skills and early social-emotional/behavioral problems and competencies. *Infant Behavior & Development, 62,* 101525. https://doi.org/10.1016/j.infbeh.2020.101525

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Riley, E., Paynter, J., & Gilmore, L. (2019). Comparing the Mullen Scales of Early Learning and the Preschool Language Scale-Fifth Edition for young children with autism spectrum disorder. *Advances in Neurodevelopmental Disorders, 3,* 29–37. https://doi.org/10.1007/s41252-018-0084-2

Sanchez, K., Spittle, A. J., Boyce, J. O., Leembruggen, L., Mantelos, A., Mills, S., Mitchell, N., Neil, E., John, M. S., Treloar, J., & Morgan, A. T. (2020). Conversational language in 3-year-old children born very preterm and at term. *Journal of Speech, Language, and Hearing Research, 63*(1), 206–215. https://doi.org/10.1044/2019_JSLHR-19-00153

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. https://doi.org/10.4135/9781412984676

Sijtsma, K., & van der Ark, L. A. (2016). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137–158. https://doi.org/10.1111/bmsp.12078

Slot, P. L., Bleses, D., & Jensen, P. (2021). Infants' and toddlers' language, math and socio-emotional development: Evidence for reciprocal relations and differential gender and age effects. *Frontiers in Psychology, 11,* Article 580297. https://doi.org/10.3389/fpsyg.2020.580297

Smits, I. A., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken scale analysis as a dimensionality assessment tool. *Applied Psychological Measurement, 36*(6), 516–539. https://doi.org/10.1177/0146621612451050

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology, 12,* 74. http://doi.org/10.1186/1471-2288-12-74

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30,* 75–99. https://doi.org/10.1007/s00357-013-9122-y

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement, 74*(5), 809–822. https://doi.org/10.1177/0013164414529793

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology, 12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115

Tillema, M., Bouwmeester, S., Verkoeijen, P., & Heijltjes, A. (2021). Psychometric properties of the short-form CART: Investigating its dimensionality through a Mokken Scale analysis. *Thinking Skills and Creativity, 39*. https://doi.org/10.1016/j.tsc.2021.100793

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*(6), 1193–1208. https://doi.org/10.1044/1092-4388(2006/086)

U.S. Department of Health & Human Services. (2019). *Office of head start*. Author. https://www.acf.hhs.gov/ohs

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1−19. https://doi.org/10.18637/jss.v020.i11

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1–27. https://doi.org/10.18637/jss.v048.i05

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*(5), 380–392. https://doi.org/10.1177/0146621610392911

van der Heijden, P. G. M., van Buuren, S., Fekkes, M., Radder, J., & Verrips, E. (2003). Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF-36. *Quality of Life Research, 12*(2), 189–198. https://doi.org/10.1023/A:1022269315437

Vernon, T. W., Holden, A. N., Barrett, A. C., Bradshaw, J., Ko, J. A., McGarry, E. S., Horowitz, E. J., Tagavi, D. M., & German, T. C. (2019). A pilot randomized clinical trial of an enhanced pivotal response treatment approach for young children with autism: The PRISM model. *Journal of Autism and Developmental Disorders, 49*(6), 2358–2373. https://doi.org/10.1007/s10803-019-03909-1

Volpe, V., Holochwost, S. J., Cole, V. T., & Propper, C. (2019). Early growth in expressive communication and behavior problems: Differential relations by ethnicity. *Early Childhood Research Quarterly, 47,* 89–98. https://doi.org/10.1016/j.ecresq.2018.10.002

Wainer, H., & Wang, X. (2005). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203–220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*(4), 575–579. https://doi.org/10.1017/S003329170800281X

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing, 21*(19), 2736–2746. https://doi.org/10.1111/j.1365-2702.2011.03893.x

Wind, S. A. (2016). Examining the psychometric quality of multiple-choice assessment items using Mokken Scale Analysis. *Journal of Applied Measurement, 17*(2), 142–165.

Wind, S. A. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice, 36*(2), 50–66. https://doi.org/10.1111/emip.12153

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century. *Journal of Psychoeducational Assessment, 29*(4), 377–392. https://doi.org/10.1177/0734282911406668

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales, Fifth Edition*. The Psychological Corporation. https://doi.org/10.1037/t15141-000