Brief/Psychometric Report

Hammill Institute on Disabilities

Assessment for Effective Intervention
37(4) 243–250

DOI: 10.1177/1534508411413760
http://aei.sagepub.com
SAGE

Psychometric Analysis of the Diagnostic Evaluation of Language Variation Assessment

Yaacov Petscher, PhD[1], Carol McDonald Connor, PhD[2], and Stephanie Al Otaiba, PhD[2]

Abstract

This study investigated the psychometrics of the *Diagnostic Evaluation of Language Variation–Screening Test* (DELV-S) test using confirmatory factor analysis, item response theory, and differential item functioning (DIF). Responses from 1,764 students in kindergarten through second grade were used in the study, with results indicating that the DELV-S is multidimensional and measures syntactic skills and nonword repetition ability. Item response theory suggested that most items were easy and that the measured skills were most reliable for students who had low language abilities. Standardized effect sizes for DIF suggested small differences existed on syntactic skills between white and minority students. Scores were vertically scaled to produce reference tables to assess performance at specific points in time, as well as growth over time.

Keywords

DELV-S, language ability, item response theory, DIF effect size, vertical scaling

The *Diagnostic Evaluation of English Variation–Screening Test* (DELV-S, Seymour, Roeper, & deVilliers, 2003) was developed to identify children with significant language delays without misidentifying children who speak dialects that differ from mainstream American English (Stockman, 2008). Although there are other widely used assessments of child language, such as the *Clinical Evaluation of Language Fundamentals* (Semel, Wiig, & Secord, 2003) and the *Peabody Picture Vocabulary Test* (Dunn & Dunn, 1981), these assessments have a history of cultural bias, particularly with regard to students who use dialects that vary from mainstream American English (Washington & Craig, 1992). This is because many features of, for example, African American English, are similar to clinical markers used by clinicians to diagnose specific language impairments (Rice & Wexler, 1996). As a result, African American and other students who use non-mainstream dialects are frequently overidentified for special education services (Oetting & McDonald, 2001).

Although the DELV-S is an important, widely used screener of language delays, the psychometric properties of the scores have been largely unexplored, and currently, the classification results provide only four categories of risk for language disorder (i.e., lowest, low to medium, medium to high, and highest risk). As a result, the DELV-S offers limited utility for many research and clinical purposes, especially for comparing student performance using a standard score based on a normative sample, or documenting gains in language skills as children mature or respond to therapy. In addition, more advanced measurement theory applications such as item response theory (IRT) have not been applied, which could be used to understand the difficulty and discrimination of items, as well as to determine whether or not scores may be scaled for progress monitoring across grades. In this study, we focused only on Part II of the DELV-S, which is designed to assess children's language skills.

Given the wide use of the DELV-S, particularly with African American and other minority students, but lack of psychometric evidence, this study has five specific aims. The first aim is to examine the factor structure of scores from DELV-S Part II using confirmatory factor analysis. In this way, we can assess the extent to which the 17 items represented a single dimension or multiple dimensions. Because IRT represents a stronger theory of measurement in item analysis, the second aim of the study is to estimate the difficulty and discrimination of each item. Hambleton and Jones (1993) noted that classical test theory methods are typically easy to apply, yet suffer from weak theoretical assumptions, which are easy to meet. Thus, using IRT provides a more rigorous approach to item estimation, which is also more generalizable than classical test methodology (Petscher & Schatschneider, 2011). Moreover,

[1]Florida Center for Reading Research, Tallahassee, FL, USA
[2]Florida State University, Tallahassee, FL, USA

Corresponding Author:
Yaacov Petscher, Florida Center for Reading Research, 2010 Levy Ave, Suite 100, Tallahassee, FL 32310, USA
Email: ypetscher@fcrr.org

because no item analysis has been performed for the DELV-S, it is plausible that items could be biased for different demographic groups—a concern for a test designed explicitly to be culture fair. Thus, a third aim was to test the extent to which subgroups in the sample might differentially respond to items.

Fourth, we also wanted to study the precision (i.e., reliability) of scores for all ability levels and determine where the scores were more or less reliable. The final aim was to vertically scale scores from the DELV-S Part II in order to produce developmental scale scores on language variation across multiple grades that could be used for assessing gains over time as well as to provide standard scores for each grade. By assessing the dimensionality and relative difficulty of the DELV-S Part II items, we may be able to provide basic evidence for the reliability and validity of scores, which would support the utility of the DELV-S as a valid assessment of language. Moreover, estimating developmental scores would provide a meaningful score metric for evaluating changes in language skills over time while standard scores would provide more nuanced information about risk than the currently available categorical variables.

Method

Participants

Students in this study were participating in two large federally funded studies focusing on language and literacy intervention. All students who attended participating schools in the target grades were invited to join the studies and 82% of the students recruited participated. Students attended schools in a large district located in North Florida where several nonmainstream dialects are used, including African American English and Southern Vernacular English (Oetting & McDonald, 2001; Patton-Terry et al., in press). The DELV-S was selected specifically because we anticipated that a substantial proportion of the students would use nonmainstream dialects. Intentionally an economically diverse sample was recruited from schools where from 4% to 96% of students schoolwide qualified for free or reduced-price meals under the *National School Lunch Program*.

A total of 1,764 students were administered the DELV-S across five time points: 250 kindergarten students were assessed in the fall, 867 Grade 1 students were assessed in both the fall and spring, and 647 Grade 2 students were assessed in the both the fall and spring. As such, 3,277 unique data points existed across all time points. Student characteristics in the sample were as follows: 42% qualified for the U.S. Free and Reduced-Price Lunch Program; 51% male, 46% white, 38% Black, 4% Asian, 5% multiracial, 3% Latino, 3% were unidentified, and <1% Native American, Pacific Islander, or Other. Based on DELV-S Part II results, overall 47.1% of the children were at lowest risk for a language disorder, 15.6% at low to medium risk, 20.4% at medium to high risk, and 16.9% at highest risk for language disorders. Fifty-nine percent of kindergarteners, 34% of first-graders, and 22% of second-graders provided nonmainstream American English targets for at least 50% of their responses on Part I of the DELV-S. The sample has more students living in poverty, more African American students, and fewer White students than the general population. Thus, the sample is highly representative of the intended school population for which this test was designed.

Measures and Procedures

The DELV-S has two parts. Part I is used to assess students' use of nonmainstream American English (see Terry, Connor, Petscher, & Conlin, 2010). Part II, which is the focus of this study, screens children's risk for language disorder by testing their syntactic knowledge via *wh-* questions and use of verbs, as well as their nonword repetition ability (Seymour et al., 2003). There are 17 items of three types: morphosyntactic, *wh-* questions, and nonword repetition. In a typical item to assess syntax, in this case possessive pronouns, students are shown a colored picture of a boy with a kite and a girl with a ball. Pointing to the boy, the examiner says, "He has a kite." The examiner then points to girl and says, "She has a ball. The kite is his. The ball is . . ." and then waits for the student to supply "hers." In a *wh-* question item, children see a picture and are told, "This girl played different things in different ways. She played drums with her feet and the piano with her hands." And then they are asked, "How did the girl play with what?" A typical nonword repetition item would ask the child to repeat a nonsense word, such as "kighgeebowfoup." The DELV-S was administered to each student individually by trained research assistants in a quiet place at the student's school in the fall only for kindergarteners and in the fall and spring for first- and second-graders.

Items are scored as correct if they match the test form examples. They are scored incorrect if they match the test form incorrect examples, do not match either the correct or incorrect examples, or the child makes no response. The Diagnostic Error Score is calculated by adding together the number of incorrect responses. This score is analogous to a raw score except that a higher score represents lower performance. This error score is indexed with the students' age (4–9 years) to identify four categories of risk–lowest, low to medium, medium to high, and highest risk. Using classical test theory, the DELV-S reported interexaminer reliability of .80 (Seymour et al., 2003), which is a minimum acceptable value when making research decisions, but falls below the clinical decision threshold of .90 (Nunnally & Bernstein, 1994). Although a strength of the interexaminer analysis was that the assessment was administered by examiners who are of different ethnic backgrounds, it was conducted with a small sample of children (n = 25) and, for Part II, showed that none of the children were incorrectly classified by either

examiner although only 36% of the children were classified exactly the same. The raw score used in this study is the number of items from Column A, which represent correct responses, and not the Diagnostic Error Score. Clinicians should total Column A in Part II to use the tables provided.

Statistical Procedures

Several analytic strategies were used to evaluate the five psychometric elements of the scores. To test the factor structures, three confirmatory factor-analytic (CFA) models were used to explore parsimony and model fit: (a) a unidimensional model to represent ability across all items; (b) a two-factor model that used the first 11 items as one factor and the remaining 6 items as another factor; and (c) a bifactor model to explore item correlations due to a shared underlying trait. The bifactor model differs from often-specified CFA models, as it states that item correlations may be attributed to a shared factor, whereas multidimensional CFA models suggest that items may be correlated because of multiple correlated traits (Reise, Morizot, & Hays, 2007). The comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean residual (SRMR) were used to evaluate model fit. CFI and TLI values greater than or equal to 0.95 are considered to be minimally sufficient criteria for acceptable model fit, and RMSEA and SRMR estimates <0.05 are desirable. In addition, the ratio between the model chi-square and degrees of freedom was calculated to assist in evaluating model fit; values <3.0 were deemed as acceptable.

The factor-analytic models were followed by multiple-group IRT and differential item functioning (DIF) analyses using Mplus 6.1 software (Muthen & Muthen, 2010) to address the remaining research aim of estimating item difficulty and discrimination, testing subgroup differential response, evaluating score reliability, and scaling resulting ability scores. We tested both one-parameter (i.e., item difficulty, 1PL) and two-parameter (i.e., item difficulty and discrimination, 2PL) IRT models and compared the results using the –2 log likelihood. A three-parameter model was not considered because of the large sample sizes that are often needed to obtain accurate parameter estimates (Swaminathan & Gifford, 1982).

A benefit of the 1PL model is that it requires fewer examinees to produce stable estimates of item difficulty. Although the full sample size was large, the number of students in kindergarten was relatively small ($n = 250$) compared to the other groups. No consensus yet exists on minimal sample sizes needed for IRT models; however, Petscher and Schatschneider (2011) noted that a number of sources have indicated that at least 200 participants should be used. Second, the scoring model for a 1PL model has the benefit of mapping the estimated ability scores directly on to the raw score totals. The 2PL model requires more participants, but retains the advantage of providing more precision in the estimation of an individual's ability as the discrimination parameter is individually estimated for each item. In 1PL and 2PL models, the range of item difficulties is –3 to 3, with lower values representing easier items, and higher values indicating more difficult items. When choosing the appropriate IRT model, it was important to carefully weigh the trade-offs between the statistical parsimony from the log likelihood test, as well as the overall ease with which the DELV-S could be administered by a practitioner with a paper-and-pencil protocol. From the IRT model, the standard error of the resulting ability score was used to estimate for whom the resulting scores were reliable.

Differential item functioning was conducted for differences between males and females, as well as white and minority groups with the Mantel-Haenszel procedure (1959). Moreover, to control for the multiple tests used, a linear step-up procedure (Benjamini & Hochberg, 1995) was applied to control for the false-discovery rate. To appropriately contextualize DIF results, Meade's (2010) method of providing a measure of effect size for the differences between groups was used. Specifically, the expected score standardized difference (ESSD) value was calculated, which is an expected score version of Cohen's (1988) *d*, and may be interpreted with the commonly used thresholds for small (.20), moderate (.50), and large (.80) differences.

Results

Factor Structure

The model fit for the unidimensional model was poor, $\chi^2(84) = 697.49$, $p < 0.001$; $\chi^2/df = 8.30$; CFI = 0.75; TLI = 0.84; RMSEA = 0.07, but improved greatly in the two-factor solution, $\chi^2(84) = 359.16$, $p < 0.001$; $\chi^2/df = 4.28$; CFA = 0.95; TLI = 0.96; RMSEA = 0.05, with an estimated correlation between the two factors of .56. The model fit for the bifactor model was not as strong as that for the two-factor solution, $\chi^2(84) = 542.31$, $p < 0.001$; $\chi^2/df = 6.45$; CFI = 0.88; TLI = 0.88; RMSEA = 0.05. Based on the fit indices from these models, the items were separated to comprise subscales reflecting syntactic skills and nonword repetition ability. The syntactic skills section was composed of the first 11 items whereas nonword repetition was made up of the remaining 6 items.

Item Difficulty and Discrimination

To appropriately estimate the item parameters and vertically scaled ability scores, multiple-group 1PL and 2PL models were used, using the kindergarten fall scores as the referent group. The log likelihood difference for syntactic skills was statistically significant in favor of the 2PL model ($\Delta G^2 = 55$, $\Delta df = 11$, $p < .001$); however, when viewing the G^2 in light of the relative improvement in fit the 2PL would provide over the 1PL model using Haberman's (1978) method, only a 0.75% improvement in fit would be observed in using the 2PL model. No significant advantage for the 2PL was estimated for nonword

Table 1. IRT Parameters, Expected Score Standardized Differences (ESSD) for Differential Item Functioning, and Reliability for DELV-S Part II

		Item Parameters		ESSD		Reliability	
Dimension	DELV-S Item	*a*	*b*	Boy	African American	Total Correct	α
Syntactic skills	1	1.55	–0.39	0.08	0.22	1	0.84
	2	1.55	–1.05	0.05	0.24	2	0.72
	3	1.55	–0.26	–0.04	0.06	3	0.78
	4	1.55	–0.17	0.09	0.43	4	0.80
	5	1.55	0.40	0.07	0.24	5	0.81
	6	1.55	0.43	0.18	0.07	6	0.81
	7	1.55	0.68	0.10	0.36	7	0.79
	8	1.55	–1.18	0.10	0.22	8	0.77
	9	1.55	–0.83	0.06	0.16	9	0.70
	10	1.55	–1.27	0.07	0.18	10	0.50
	11	1.55	–1.44	0.09	0.10	11	0.10
Nonword repetition	12	1.69	–1.53	0.04	0.16	1	0.52
	13	1.69	–1.00	0.07	0.07	2	0.67
	14	1.69	0.22	–0.02	–0.11	3	0.69
	15	1.69	–0.68	–0.06	0.06	4	0.64
	16	1.69	–0.85	0.03	0.07	5	0.49
	17	1.69	0.56	0.00	–0.01	6	0.19

Note. IRT = Item Response Theory; DELV-S = *Diagnostic Evaluation of English Variation–Screening Test.*

repetition ($\Delta G^2 = 5$, $\Delta df = 2$, $p = .082$); thus, the 1PL model was selected for item estimation for each factor. The estimated item difficulties for the two separate factors are reported in Table 1. For the syntactic skills, the items ranged in difficulty from –1.44 (Item 11) to 0.68 (Item 7) with a mean of –0.46, and for nonword repetition the easiest item was Item 12 (–1.53), with Item 17 being the most difficult (0.56) and an overall mean of –0.55. Because the mean item difficulties for each factor were negative, this was an indication of the relative easiness when vertically equated across the five groups.

Differential Item Functioning

With the exception of Items 3 and 15, the Mantel-Haenszel procedure indicated that all items demonstrated DIF when comparing males to females and White to minority students. The effect size values, where higher values suggest that items are more difficult for boys or African American students, reported in Table 1 demonstrated that very small standardized differences actually occurred in the sample. The ESSD values ranged from –0.02 to 0.18 for gender differences, with a mean absolute ESSD of 0.05 across all items. When considering race differences, the observed range was –0.11 to 0.43, with a mean absolute ESSD of 0.16.

Reliability

The reliability of the students' syntactic skill and nonword repetition ability scores are reported in Table 1. From an IRT framework, reliability is something that can be estimated at any given point of ability, rather than being one value that is assumed to be static across ability levels (Petscher & Schatschneider, 2011). By estimating the standard error of the participants' ability scores, a reliability score similar to Cronbach's alpha can be computed with $\alpha = 1 - SE^2$. As mentioned previously, a benefit of the 1PL model is that raw total correct can be directly linked to the IRT ability score. Subsequently, the reliability of the total score is an analog of the reliability for the IRT ability scores. For syntactic skills, acceptable reliability for research decisions (i.e., approximately .80; Nunnally & Bernstein, 1994) was estimated when the total score ranged from 1 to 8 (i.e., low to average syntactic skills). Reliability for nonword repetition was lower across the range of total scores compared to syntactic skills, and was most reliable for total scores of 1 to 4 (i.e., low to average nonword repetition skills). In addition to the IRT estimates of reliability, Cronbach's alpha was estimated for each scale, with values of 0.67 for the 11 items on syntactic skills, 0.57 for the 6 nonword repetition items, and an overall scale reliability of 0.71.

Scaling of Ability Scores

Using the latent ability scores for each dimension, a series of look-up tables (see Table 2) were created for three different score types for each of the five administrative periods: (a) a standard score with a mean of 100 and standard deviation of 15 was developed, (b) a percentile rank, and (c) a developmental

Table 2. Developmental and Administrative Scale Scores–Syntactic Skills and Nonword Repetition

		Syntactic Skills			Nonword Repetition		
Time Point	Raw Score	D-Score	Standard	Rank	D-Score	Standard	Rank
Fall–K	0	220	58	1	248	62	1
	1	283	67	2	325	74	4
	2	339	76	5	389	83	13
	3	377	82	11	442	91	28
	4	410	87	18	498	100	49
	5	440	91	27	568	110	75
	6	470	95	38	651	123	93
	7	501	100	50			
	8	535	105	64			
	9	575	111	77			
	10	633	120	91			
	11	702	130	98			
Fall–G1	0	220	54	1	248	49	1
	1	283	54	1	325	62	1
	2	339	54	1	389	73	4
	3	377	61	1	442	83	12
	4	410	68	2	498	92	30
	5	440	73	4	568	104	61
	6	470	79	8	651	119	89
	7	501	85	15			
	8	535	91	28			
	9	575	99	47			
	10	633	110	74			
	11	702	123	94			
Spring–G1	0	220	53	1	248	47	1
	1	283	53	1	325	60	1
	2	339	53	1	389	71	3
	3	377	53	1	442	80	9
	4	410	60	1	498	90	25
	5	440	66	1	568	102	55
	6	470	71	3	651	116	86
	7	501	78	7			
	8	535	84	15			
	9	575	93	31			
	10	633	104	61			
	11	702	118	88			
Fall–G2	0	220	50	1	248	39	1
	1	283	50	1	325	53	1
	2	339	50	1	389	65	1
	3	377	50	1	442	75	5
	4	410	57	1	498	86	17
	5	440	63	1	568	99	48
	6	470	69	2	651	115	84
	7	501	76	5			
	8	535	82	12			
	9	575	91	27			
	10	633	103	57			
	11	702	117	87			
Spring–G2	0	220	47	1	248	43	1
	1	283	47	1	325	56	1
	2	339	47	1	389	68	2
	3	377	47	1	442	77	6
	4	410	47	1	498	87	19
	5	440	54	1	568	99	48
	6	470	61	1	651	114	82
	7	501	68	2			
	8	535	76	5			
	9	575	85	16			
	10	633	98	46			
	11	702	114	83			

Note. K = Kindergarten; G1 = Grade 1; G2 = Grade 2; D-Score = Developmental Scale Score.

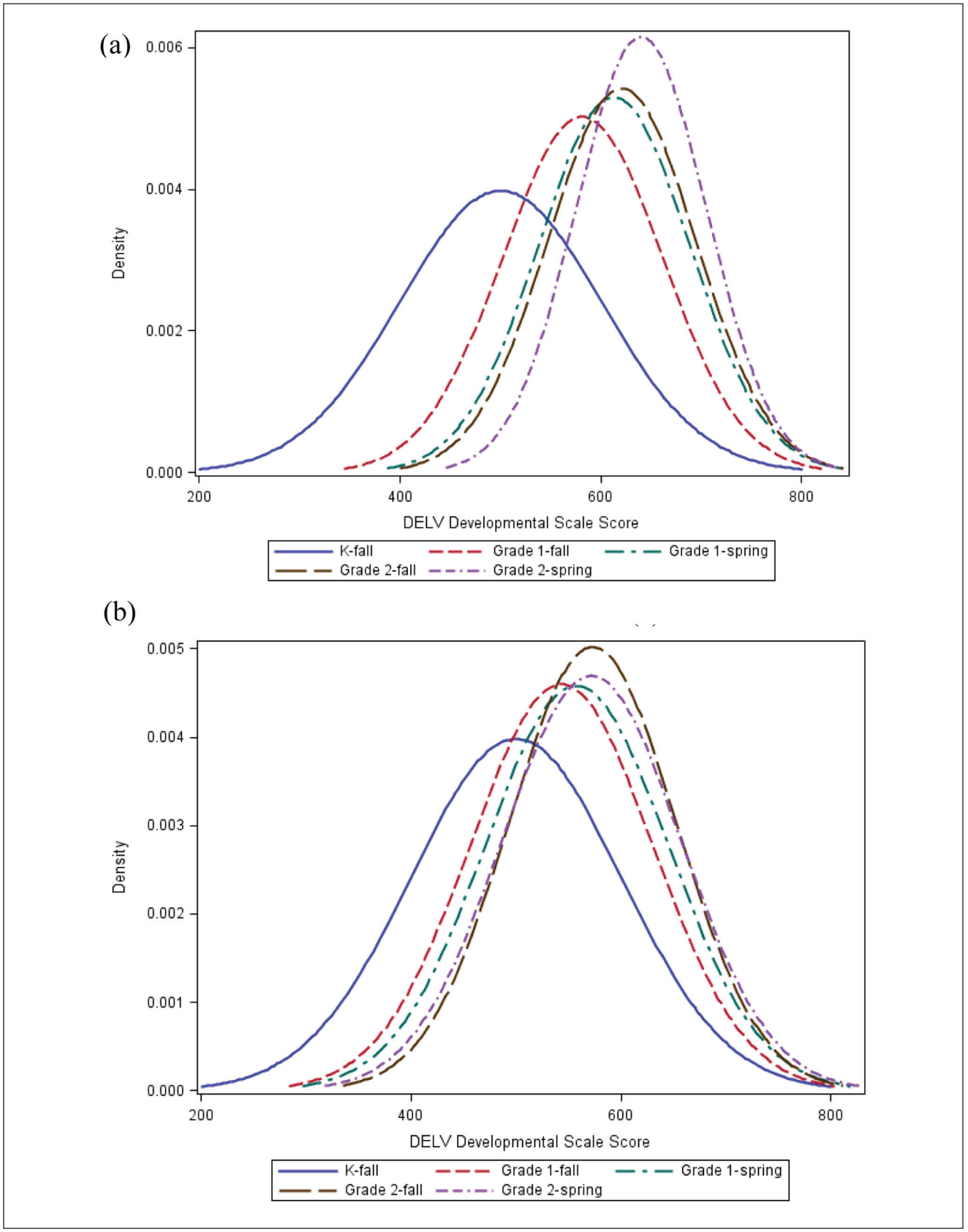

Figure 1. Ability score distributions of (a) syntactic skills and (b) nonword repetition by grade.

scale score (D-Score). The standard score and percentile ranks were provided so that clinicians and practitioners would be able to contextualize students' ability within a given point in time. The D-Score was created so that clinicians could assess the range of ability from kindergarten through second grade and to assess changes in scores from kindergarten through second grade. The D-Score has a mean of 500 and a standard deviation of 100 and operates in a similar fashion to a W-score in the Woodcock-Johnson battery of assessments. Notice that in Table 2, when examining the values close to the mean at each time point, the D-Score changes. For example, in the fall of kindergarten, a mean score of 100 corresponds to a raw score of 7 and a D-Score of 501, which is expected for the group used as the referent in the scaling analysis. At the fall–Grade 1 time point, the mean of 100 now approximately corresponds to a raw score of 9 and a D-Score of 575. Thus, the D-Score will reflect the actual growth in the latent ability. Figure 1a models the ability scores for syntactic skills, and the distributions indicate that a large shift in the distribution occurs between the fall of kindergarten and the fall of Grade 1. Although small changes occurred within first grade, another larger shift occurred between the fall of Grade 1 and the spring of Grade 2. When considering nonword repetition skills (Figure 1b), the largest ability shift occurs between fall of kindergarten and the fall of Grade 1, with minimal observed changes from first to second grade.

Discussion

The DELV-S is unique in that it was explicitly designed to provide a culture-fair screening of whether children, particularly children who speak nonmainstream dialects, might be at risk for language disabilities. In our highly diverse nation, such an assessment can be of great value. At the same time, it is imperative that scores from this assessment are reliable, valid, and not culturally biased. To date, only interrater reliability estimates have been reported (Seymour et al., 2003); hence the purpose of this study was to provide a more comprehensive psychometric analysis and to improve the utility of the DELV-S Part II for researchers and clinicians by explicating the factor structures, item parameter properties, subgroup differential response, score reliability, and scalability of scores.

Raw scores from DELV-S Part II were computed by totaling the correct answers, rather than incorrect as stated in the examiners' manual. Our findings extend the reported psychometrics by revealing that a multidimensional model provides the best description of the data and that the morphosyntactic items and the nonword repetition items should be scored separately because they represent two different constructs (see Table 2). Whereas the items were generally easy for students to correctly answer, this is appropriate for a screening assessment because it is designed to identify children with weak language skills. Furthermore, appropriately, this screening assessment is more reliable in assessing language skills for students who have relatively weaker skills, and is probably not appropriate for assessing students with above average language skills. Hence it appears to be an appropriate tool for practitioners and researchers, keeping the target students in mind.

Of concern in an assessment that is designed to be culture fair, some of the items are more difficult for African American students than they are for White students. Although the magnitude is small (average absolute Cohen's $d = .16$), this warrants further investigation. Inclusion of the standard scores in Table 2 allows a more nuanced assessment of the magnitude of the risk for language disorder when compared to the 4 categories currently available. This is important because students can just miss or just make a categorical determination of risk. Standard scores recognize that language skills fall on a normal distribution and allow practitioners to make better decisions. D-Scores allow clinicians and researchers to measure changes in students' language skills over time.

As a landmark culture-fair assessment, the DELV-S in a key tool in the clinician's battery. Our study provides important psychometric information for the DELV-S Part II that provides a note of caution with regard to cultural bias but, more importantly, improves its utility as a practical screening measure and research assessment.

Declaration of Conflicting Interests

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

Funding

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This research was funded by the Institute of Education Sciences (#R305F100005).

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*, 289–300.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: American Guidance Service.

Haberman, S. J. (1978). *Analysis of qualitative data: Vol. 1. Introductory topics*. New York, NY: Academic Press.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *NCME Instructional Module,* 253–262.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–48.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728–743.

Muthen, L. K., & Muthen, B. O. (2010). *Mplus software (Version 6)*. Los Angeles, CA: Author.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Oetting, J. B., & McDonald, J. L. (2001). Nonmainstream dialect use and specific language impairment. *Journal of Speech and Hearing Research, 44*, 207–223.

Petscher, Y., & Schatschneider, C. (2011). Validating scores from new assessments: A comparison of classical test theory and item response theory. In G. Tenenbaum, R. Eklund, & A. Kamata (Eds.), *Handbook of measurement in sport and exercise psychology* (pp. 75-101). Champaign, IL: Human Kinetics.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech & Hearing Research, 39*, 1239–1257.

Semel, E. M., Wiig, E. H., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals (CELF-4)*. San Antonio, TX: Psychological Corporation.

Seymour, H. N., Roeper, T. W., & deVilliers, J. (2003). *Diagnostic Evaluation of Language Variation, Screening Test examiner's manual*. San Antonio, TX: The Psychological Corporation.

Stockman, I. J. (2008). Toward validation of a minimal competence phonetic core for African American children. *Journal of Speech, Language, and Hearing Research, 51*, 1244–1262.

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss & R. D. Bock (Eds.), *New horizons in testing: latent trait test theory and computerized adaptive testing* (pp. 13–30). New York, NY: Academic Press.

Patton-Terry, N., Connor, C., Petscher, Y., & Conlin, C. R. (in press). Dialect variation and reading: Is change in Non-mainstream American English use related to reading achievement in first and second grade? *Journal of Speech, Language, and Hearing Research*.

Washington, J. A., & Craig, H. K. (1992). Performances of low-income, African American preschool and kindergarten children on the Peabody Picture Test–Revised. *Language, Speech and Hearing Services in Schools, 23*, 329–333.