Using Item Response Theory to Describe the Non-Verbal Literacy Assessment (NVLA)

Danielle Fleming, M.A., University of California, Berkeley and San Francisco State University

Mark Wilson, Ph.D., University of California, Berkeley

Lynn Ahlgrim-Delzell, Ph.D., University of North Carolina, Charlotte

Corresponding author: Danielle Fleming

Cell: (510) 329-0812

Email: myteachermsfleming@gmail.com or

myteachermsfleming@berkeley.edu

Abstract

The Non-Verbal Literacy Assessment (NVLA; Baker, Ahlgrim-Delzell, Flowers, & Browder, 2010) is a literacy assessment designed for students with significant intellectual disabilities. The 218-item test was initially examined using confirmatory factor analysis. This method showed that the test worked as expected, but the items loaded onto a single factor. The current paper uses item response theory (IRT) to investigate the NVLA using Rasch models. First, we reduced the number of items using a unidimensional model, which resulted in high levels of test reliability despite decreasing the number of questions, providing the same information about

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/pits.22110](#).

ITEM RESPONSE THEORY FOR THE NVLA

student abilities in less time. Second, the multidimensional analysis indicated that it is possible to view the NVLA as a test with four dimensions, resulting in more detailed information about student abilities. Finally, we combined these approaches to obtain both specificity and brevity, with a 4-dimensional model using 133 items from the original NVLA.

Keywords: Rasch model, non-verbal, assessment

Using Item Response Theory to Describe the Non-Verbal Literacy Assessment

Federal legislation in the United States mandates assessment for all children. The Individuals with Disabilities Act (IDEA, 1997) and its reauthorization (IDEA, 2004) require students with disabilities to participate in state assessments. Prior to these requirements, many students with developmental disabilities were provided with limited exposure to reading instruction. There has also been a lack of appropriate measures of reading skills for this population. Many students with significant developmental disabilities or autism are not able to access pencil and paper tests, or to provide verbal responses to questions. To address these issues, many states created and implemented alternate assessments for this population of students.

However, these assessments were often of questionable validity and were not correlated with other external measures of literacy. While there are available measures of early literacy such as the Test of Early Reading Ability, 3rd ed. (TERA – 3, Reid, Hresko, & Hammill, 2001), Gates-MacGinitie Reading Test, 4th ed. (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000), and the Woodcock-Johnson III Diagnostic Reading Battery (WJ III DRB, Schrank, Mather, & Woodcock, 2004), these tests require students to respond verbally to test items. The student samples used to provide evidence of validity and

ITEM RESPONSE THEORY FOR THE NVLA

reliability for these tests did not include students with significant disabilities.

Previous researchers looking for ways to teach reading skills to students with severe developmental disabilities had difficulty finding appropriate measures of reading skills. Browder, Allor, Sevick, & Ahlgrim-Dezell (2008) found that only 20% of this population could successfully participate in published measures of reading due to the verbal demands of such tests. For many students researchers could not establish a basal level. This indicated that these students needed an assessment that could capture skills that *preceded* the skills included in these assessments. Researchers and practitioners currently lack adequate reading measures for students who may not have acquired the needed test-taking skills or who may need the support of augmentative communication systems in order to communicate.

The Non-Verbal Literacy Assessment (NVLA; Ahlgrim-Dezell, Browder, Flowers, & Baker, 2008) was developed to assess the literacy skills of students from K-5th grade (although it may also be considered for students in middle and high school if early reading instruction is designated as an IEP goal) who use alternative forms of communication including eye gaze boards, alternative augmentative communication systems, or other modes of expression. The test was designed to measure early literacy skills reflecting six constructs in line with the National Reading Panel (2000) recommendations: phonemic awareness, phonics, reading comprehension, vocabulary, listening comprehension, and text awareness. There are 218 items presented using a receptive response format with answers provided in two- to four-choice arrays. Responses are provided in a standardized format using finger pointing, eye gazing, pulling Velcro cards, or pulling a response from a fanned array held by the tester. Administration time is about 90 minutes (broken into smaller test sessions) depending upon the processing time needed for the student to respond to the items and the amount

ITEM RESPONSE THEORY FOR THE NVLA

of manipulation of the different response options. The test requires one on one attention from a test administrator (usually a teacher) with a student. Due to the need for teachers to spend time educating all students in their classrooms, the time requirements for administering a 90-minute test can be prohibitive. Ideally such a test should be as short as possible (20-30 minutes for one session to an hour broken into two sessions would be more reasonable) while still providing useful and detailed information about student abilities in order to inform lesson planning by teachers.

In the original (exploratory) factor analysis of the 218-item NVLA, estimates of reliability were found to be “good” with a test-retest reliability coefficient of .97 and Alpha internal consistency coefficients ranging from .80 to .98 (Browder, Ahlgrim-DeLzell, Courtade, Gibbs, & Flowers, 2008). A confirmatory factor analysis of the NVLA was conducted on three models of literacy: a six-factor model, a two-factor model, and a one-factor global model (Baker, et al., 2010). The multiple factor models showed correlation coefficients among the factors of between .82 and .99. The best fitting model was the one-factor model of literacy. As an area for future research, the authors suggested the use of item response models to provide additional evidence of validity for the NVLA.

The current study will explore item response theory (IRT) models to describe the psychometric properties of the NVLA. Two research questions guide this study. First, can the number of NVLA items be reduced to gain efficiency and avoid student and administrator fatigue while still preserving the high reliability of the original? This question is best answered using a unidimensional dichotomous Rasch model. Second, what is the diagnostic value of the NVLA on each dimension? More detailed information about a student's reading ability can be valuable to teachers hoping to target

interventions to promote student achievement. To answer this question a 4-dimensional item response model was used. In answering these two questions, we hoped to find a model that fit a version of the NVLA that both provides detailed information about student performance and also is optimally efficient to administer.

Method

Participants

There were 146 students' NVLA scores included in this secondary analysis with 0.46% of the total data missing. Participants in the original NVLA study, from which data were obtained for this secondary data analysis, were elementary-aged students in cross-categorical classrooms for students with specialized academic needs who had participated in a study on literacy instruction. Students were in the following grades: K (25.9%), 1st (14.3%), 2nd (17.7%), 3rd (15.6%), 4th (8.2%), and 5th (8.2%) with data on grade level missing for 10.2% of the participants. IQ scores (from regular IQ tests) were obtained for 45 of the 146 students (30.8%) and ranged from 20 to 78 with an average score of 44. Many students were not able to participate in regular IQ testing due to behavioral or language challenges. Student diagnoses included autism (27.2%), developmental delay (26.5%), intellectual disability (25.2%), multiple disabilities (5.4%), other disabilities such as health impairment or speech language impairment (9.5%), and missing (6.2%). Female participants accounted for 36.3% of the total and 54.8% were male (with 9.6% of gender identification scores missing). Ethnic and racial backgrounds included 44.9% African American, 29.9% White, 11.6% Hispanic, 8.2% other or multi-racial, and 1.4% Asian. Four percent of the data on this item were missing. Students who were able to speak (i.e., verbal) comprised 47.3% of the participants. Students who were unable to speak (i.e., non-verbal) made up 37.7% of the participants. Data were missing on this item for 14.33% of the participants. Although many students were identified as being verbal, their verbal skills

were not sufficient to respond to other published reading or literacy assessments. Missing data for demographics was due to incompleteness of forms submitted by the teachers.

Rasch's Simple Logistic Model

For the first analysis, Rasch's Simple Logistic Model was used. The Rasch model (1960, 1980) for dichotomous items is described by the following formula:

$$\text{logit}(X_{ij} = 1 | \delta_i, \theta_j) = \theta_j - \delta_i \quad (1)$$

where *logit* is the log of the odds (of success), and X_{ij} is the respondent's success (1) or failure (0) on the item i , θ_j is the student's estimated ability level, and δ_i is the estimated item difficulty. Following standard estimation practice for item response models, we constrained the person ability mean to zero. Once estimates were obtained for the item difficulty levels using this constraint, and these estimates were then used to estimate the person ability levels or θ .

A multidimensional version of Rasch's simple logistic model can also be estimated for multidimensional contexts using the Multinomial Random Coefficients Multinomial Model (MRCML, Adams, Wilson, & Wang, 1997). The estimates obtained using this model provide expected a-posteriori (EAP), weighted maximum likelihood (WLE), and maximum likelihood estimates (MLE) for each dimension. A discussion of results obtained for the multidimensional model will follow the discussion for the single dimensional model.

Multidimensional Model

To examine the possibility that the NVLA can provide useful information about specific areas of early literacy, we used a between-item multidimensional item response model (Adams, Wilson, & Wang, 1997), which is an extension of the simple unidimensional Rasch model. This model assumes that each item is identified with one specific construct. The *ConQuest* software

(Adams, Wu & Wilson, 2015) uses the MRCML model (Adams, Wilson, & Wang, 1997) to analyze the data given the following formula:

$$\text{Logit}(X_{ij} = 1 | \delta_i, \theta_{jd(i)}) = \theta_{d(i)} - \delta_i \quad (2)$$

where $d(i)$ indicates which dimension item i is on, and the person ability (θ) is now a vector,

$$\theta_j = (\theta_{1j}, \theta_{2j}, \theta_{3j}, \theta_{4j})$$

As the MRCML model is a confirmatory model, the expected dimensional assignment of each item must be specified prior to the analysis.

The multidimensional model we used was a four dimensional model using: (1) conventions of reading, (2) comprehension, (3) word study/vocabulary and (4) a combination of phonological awareness items and phonics items called PhonSK. This model was based on the original design of the NVLA as a six-factor assessment including: (a) phonemic awareness, (b) phonics, (c) reading comprehension (d) vocabulary, (e) listening comprehension, and (f) text awareness (Baker et al., 2010). We collapsed the two comprehension sections envisioned by the original authors (listening and reading comprehension) in order to appropriately identify them as belonging to a specific construct. When reviewing the original NVLA questions, the authors found that in fact there was no way to distinguish between the listening and reading comprehension items (as all items are read aloud to the students, who respond by pointing to an answer). Therefore, the comprehension items were grouped together in the model specification before running the analysis. We also collapsed the phonics and phonological awareness items in order to obtain a more generalized and reliable score for items on the phonics skills construct. This was due to the unclear distinction between these two categories in the presentation of these items.

The original factor analysis also attempted a two-factor model, but it was unclear whether only two dimensions would provide optimally useful data for teacher planning purposes. When we considered splitting the items into two dimensions, we could have divided the items into (1) concepts about print (such as book orientation and following lines of print) and (2) comprehension, vocabulary, and reading skills. But this did not seem like a model that would reflect the nature of the items nor provide useful information about the range of skills teachers would be interested in, because too many skills would fall into the second dimension, preventing differentiated information from being communicated.

Results

Unedited NVLA

Wright maps for each dimension are used to show the distribution of items and students on a logit scale. A Wright map showing the item difficulty levels for the unedited NVLA with one dimension can be found in Figure 1. The left-hand side shows the estimated person ability, with higher ability individuals towards the top of the scale. The right-hand side shows the item difficulties with items that have similar or equal difficulty parameters stacked next to each other.

Figure 1 about here

Nearly all items exhibited appropriate levels of fit. In the initial unidimensional analysis of all items included on the NVLA (218 items), only 3 items exhibited misfit (all other mean-squared values were between $-.75$ and 1.33). There were many clusters of items on the original test that were found to have similar difficulty levels (see Figure 1). This was especially true for items in the upper ability levels of the test.

Can the number of NVLA items be reduced?

Due to the length of the NVLA and issues of fatigue for both students and test administrators, we hoped to reduce the number of items on the test while preserving the reliability of the instrument. To do so, we performed several iterations of unidimensional IRT analyses after discarding items with similar difficulty parameters. We attempted to edit items for discarding in a way that preserved the balance of the test with respect to the original blueprint by using a stratified random sampling approach. This involved randomly selecting items at similar difficulty levels at random and dropping them from the analysis, then reanalyzing the data until we started to lose reliability, at which point, we stopped dropping items. Use of the Wright map allowed ease of selection for items for dropping at the same difficulty levels. For example, by looking at Figure 1, it is possible to see that many items (numbers listed to the right of the map) are at about the same difficulty level as other items. Many items are clustered at difficulty levels between 1 and 2.5 logits. Items in these clusters were randomly dropped when rows were highly redundant (for example, by dropping the last column or every other column of redundant items). Eventually, we arrived at a reasonable reduced version of the NVLA with just 50 items, given the one-dimensional model intended by the test-makers. No items in this version of the instrument exhibited misfit. The NVLA in its unedited original form has high reliability both in terms of person separation reliability and in EAP/PV reliability. For the 218-item NVLA the *expected a posteriori plausible values reliability* (*EAP/PV reliability*) is 0.97. For the 50-item, version of the NVLA, the *EAP/PV reliability* is 0.93. For the 218-item NVLA the *weighted likelihood estimate (WLE) person separation reliability* is 0.94. The *WLE person separation reliability* for the 50-item NVLA is 0.93. These results indicate that the unidimensional, 50-item NVLA is, in practical terms, comparable in reliability to the original 218-item NVLA, implying an affirmative answer to the first research question. The NVLA can be made substantially shorter without losing the amount of the reliability of the original version.

ITEM RESPONSE THEORY FOR THE NVLA

A Wright map for the items and person abilities for the 50 –item NVLA (not illustrated) was also obtained showing a roughly normal distribution for both versions of the test. The distribution of abilities measured by the reduced version of the test is consistent with the distribution obtained with all 218 items. These results indicate that the original NVLA may include more items than are necessary to obtain reliable estimates of student abilities in early literacy. While such redundancy may be useful for a situation when one needs to use multiple forms of the same test, using items with the same difficulty level on different versions of the test, for say, a pre-test/post-test design, it does not appear that all 218 items are needed for a single administration. Note that having item banks with assessment items measuring the same or similar levels of ability may be a useful tool for computerized adaptive testing (CAT), however.

The *ConQuest* analysis showed that for the 4-dimensional model the dimensions are highly correlated with each other using the original 218 items (between .88 and .98, see Table 1). However, we believe that each dimension provides unique educational information about the respondents' abilities. For example, it may be interesting to know how a given student performs on tasks requiring phonics skills and how those abilities differ from the student's abilities in reading conventions or comprehension. By having a more fine-grained analysis of the student response data, teachers may be able to make informed decisions about individualized education program goals or areas to focus intensively on in lesson planning for each student. For students with unique educational needs who may have difficulties with specific areas of reading, information regarding the student's performance in the four dimensions in this model could inform teaching interventions. A student may have difficulty with conventions of reading (book orientation, left to right text reading) yet be able to identify letter sounds or blend words (phonics skills) indicating a need for more hands-on experience with using books (rather than computer phonics programs, for instance). Alternatively, a

student may perform well in conventions of reading and phonics, but score lower in comprehension, indicating that the child may benefit from interventions targeting meaning making from reading and listening experiences (for example, by providing picture answer cards, for students who cannot answer verbally in storybook reading groups).

Table 1 about here

As for the one-dimensional model, we attempted to reduce the number of items in the 4-dimensional model such that the representation of the dimensions and the level of reliability was similar to that provided by all 218 items. After several iterations, we found that it was possible to obtain very similar results using 133 items, rather than 218 items. To fit the final, reduced 4-dimensional model, we categorized items to be analyzed under each of the dimensions. The dimension, conventions of reading, included 14 items. Comprehension included 17 items. Word study included 33 items. Phonics skills included 69 items

Even though we know the variables are highly correlated, the differentiation provides interesting information about participants' abilities in different areas of reading. The correlation of dimensions for the edited 133-item version of the NVLA can be found in the upper triangle of Table 1. Reliability estimates are as follows: *WLE person separation reliability* for conventions of reading is *0.80*, for comprehension: *0.64*, for word study: *0.83*, and for phonics skills: *0.85*. *EAP/PV reliability* for conventions of reading is *0.88*, for comprehension: *0.91*, for word study: *0.96*, and for phonics skills: *0.96*. Since it is possible to use a shortened version of the NVLA (i.e., with 133 items) and obtain reliable information regarding respondents' abilities on different areas of the test, for those seeking such detailed information, this version of the test will be more useful than a global measure of early literacy skills.

By looking at the Wright maps in Figure 2, we are able to see that the conventions of reading tasks are fairly evenly spread out among person ability levels. Nearly every item is

providing a unique level of information. This is less true for the phonics skills dimension, which has quite a few items matched to respondents with higher ability levels, but does not give so much information about students on the lower ability levels. In a sense, there is a sharp cutoff between students without phonics abilities and those who possess them, as represented by this Wright map. The comprehension and word study dimensions had a more even distribution than the phonics dimension, indicating that it was more likely for students in the sample to have a range of skills in these areas. For each of the maps, the lowest ability levels were represented by a group of 12 students who answered all questions incorrectly using one of the several answering formats available to them.

Figure 2 about here

Further support for the use of this edited version of the NVLA using a 4-dimensional analysis can be seen in Table 2. To provide a comparison between the 4-dimensional model and the unidimensional model, unidimensional analysis results were obtained for the 133 items on the proposed reduced version of the NVLA. The 4-dimensional model was shown to provide more information than the unidimensional model of the same data using a chi-square test of model difference: $\chi^2 = 114.5$ ($df = 9$, $p < 0.001$). Rabe-Hesketh and Skrondal (2012, 88-89) argued that this model difference is conservative, since there is a lower boundary of zero. The test between the models could therefore be seen as one-sided, and the probability value should be divided by 2 (in which case, $p < 0.0005$). The size of this difference in deviance value indicates that the 4-dimensional model is a better fit than the unidimensional model, in terms of statistical significance. This is confirmed by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, the lower values of which indicate that the 4-dimensional model represents the data better than the unidimensional model.

Table 2 about here

Discussion

The NVLA is currently the only test that reliably measures student literacy abilities for those with severe developmental disabilities who require access to non-verbal forms of communication. The original test with all 218 items has excellent reliability, yet can be difficult to administer due to its length and the abilities and attention span of the student population. A drastically shortened version of the test would provide essentially the same information that it currently offers using only 50 items and taking about 20 minutes to administer. Additionally, if test administrators were interested in learning more about student skills in specific areas of literacy, the NVLA can be designed as a four dimensional instrument using 133 of the original 218 items, taking about an hour to administer. Although somewhat longer, this model provides considerably more information about student skills useful for identifying strengths and weakness. The 133-item, four-dimensional model appears to both more efficient than the original model with 85 fewer items and more useful than the 50-item model by providing information that can be used for subsequent instructional planning.

Supplementary materials including secondary de-identified data (exempt from human subjects review) and *ConQuest* output may be obtained by emailing the authors.

Limitations and Recommendations for Future Research

Limitations include the amount of missing demographic information of the participants, the small sample size, and a restricted sample from a large, urban area in the southeastern United States. These issues may impact the generalizability of these findings to other individuals with developmental disabilities. Additional validation studies with other populations of individuals with developmental disabilities are needed.

Little information is known about the role of co-morbid diagnosis for this population of students. The test is designed for students with intellectual disabilities including autism, Down Syndrome, traumatic brain injury, or any other disability which includes a moderate to severe intellectual disability. Further studies might consider the unique impact of specific disabilities related to intellectual disability. However, this was beyond the scope of the current paper.

Future studies might look at construct validity for the four dimensions used in this analysis. With further analysis, Thurston step thresholds could be used to either verify the theorized levels within each of the four constructs, or allow for revision of these constructs to better portray the levels measured by the NVLA. However, this level of analysis was beyond the scope of this paper. Additionally, a closer analysis of the test might look for items that violate the local independence assumption of IRT. Items that rely on the same stimulus (story comprehension items, for example) might be better represented as bundled, dependent groups of test questions. Another potential concern with the formatting of the test is that no verbal responses are required of participants, indicating, that for participants who are verbal, an additional reading sub-test that includes the dimension of fluency (per the National Reading Panel's suggestion) may be necessary to supplement the information provided by the NVLA.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Adams, R. J., Wu, M., & Wilson, M. (2015). ACER ConQuest (Version 3.0) [Generalized item response modeling software]. Retrieved from Acer.edu.au.

- Ahlgrim-Delzell, L., Browder, D. M., Flowers, C., & Baker, J. N. (2008, February). The nonverbal literacy assessment (NVLA). Paper presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA.
- Baker, J., Spooner, F., Ahlgrim-Delzell, L., Flowers, C., & Browder, D. (2010). A measure of emergent literacy for students with severe developmental disabilities. *Psychology in the Schools*, 47, 501-513. doi:10.1002/pits.20486
- Browder, D. M., Ahlgrim-Delzell, L., Courtade, G., Gibbs, S. L., & Flowers, C. (2008). Evaluation of the effectiveness of an early literacy program for students with significant developmental disabilities. *Exceptional Children*, 75, 33-52.
- Browder, D. M., Allor, J., Sevick, R., & Ahlgrim-Delzell, L. (2008, March). Reading research: Results from 2 years of research with students with intellectual disability. Paper presented at the annual meeting of the American Educational Research Association Conference, New York.
- Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. (2000). *Gates-MacGinitie Reading Test*, 4th ed. Itasca, IL: Houghton Mifflin HarCourt.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Department of Health and Human Services [NIH Pub. No. 00-4754].
- Rabe-Hesketh, S. & Skrondal, A. (2012) *Multilevel and longitudinal modeling using stata*, Volume 1: Continuous response. College Station, TX: Stata Press.

ITEM RESPONSE THEORY FOR THE NVLA

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability* (3rd ed.).

Austin, TX: Pro-Ed.

Schrank, F. A., Mather, N., & Woodcock, R. W. (2004). *Woodcock-Johnson III Diagnostic Reading*

Battery. Itasca, IL: Houghton Mifflin HarCourt.

Table 1

Correlation matrix for the 4-dimensional model of the unedited NVLA (lower triangle) and the 133-item NVLA (upper triangle)

Dimension	Conventions of Reading (CR)	Comprehension	Word Study/Vocabulary (WS)	Phonics Skills
CR		0.91	0.88	0.89
Comprehension	0.91		0.91	0.90
WS	0.88	0.92		0.97
Phonics Skills	0.89	0.93	0.98	

Table 2

Additional Information for the 4-Dimensional Version of the NVLA Compared to the Unidimensional Model

	<u>4-dimensional model of NVLA with 133 items</u>	<u>Unidimensional model of the NVLA with 133 items</u>
No. of Parameters	142	133
Deviance	16671.7	16786.2
AIC	16955.7	17052.2
BIC	16979.0	17074.1
Chi-Square of model difference	(114.5 with 9 degrees of freedom)	

Figure 1. Wright Map for Numbered Questions on the NVLA

Figure 1. This Wright map of the unidimensional analysis shows the item number from the original NVLA on the right, and the distribution of students achieving a correct answer on questions on the left. This map allows the reader to see the distribution of scores for the NVLA as a whole.

Figure 2. Wright Maps From the 4-Dimensional NVLA Analysis

Figure 2. These Wright maps for each of the four dimensions show the distribution of scores for each section of the final NVLA. From left to right, the dimensions are; conventions of reading, comprehension, word study, and phonics skills.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Fleming, D;Wilson, M;Ahlgrim-Delzell, L

Title:

Using item response theory to describe the Nonverbal Literacy Assessment (NVLA)

Date:

2018-04-01

Citation:

Fleming, D., Wilson, M. & Ahlgrim-Delzell, L. (2018). Using item response theory to describe the Nonverbal Literacy Assessment (NVLA). PSYCHOLOGY IN THE SCHOOLS, 55 (4), pp.341-349. <https://doi.org/10.1002/pits.22110>.

Persistent Link:

<http://hdl.handle.net/11343/283657>