Scaling Measures of Early Literacy

Anthony D. Albano, Michael C. Rodriguez, Scott McConnell,
Tracy Bradfield, & Alisha Wackerle-Hollman
University of Minnesota

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education, New Orleans, LA.
April, 2011

Scaling Measures of Early Literacy

Individual Growth and Development Indicators (IGDI) were originally designed as General Outcome Measures (GOM; Fuchs & Deno, 1991), brief assessments used to monitor student progress over the course of instruction. IGDI measures could be administered by a teacher to an individual student in as little as 1 minute. The simple design made it possible to collect IGDI scores on a regular basis, supporting decision making in a response to intervention (RTI) program and providing repeated measures across multiple time points, useful for plotting growth trajectories over time.

Like GOM and other classroom assessments, the first version of the IGDIs was based, implicitly, on a classical test theory model. Sample-specific item statistics, such as proportion correct (i.e., p-value) and item-total correlations (e.g., point-biserial), guided the item writing, piloting, and analyses stages, and raw summed scores were used to describe student ability. Though alternate forms of a given IGDI measure were created to be similar, no adjustments were made for form difficulty differences. The initial IGDI scaling also provided classical indices of measurement error. (For details on the design and implementation of version 1 of the IGDIs, see Wackerle-Hollman, Bradfield, McConnell, Albano & Rodriguez, 2011; Bradfield, Wackerle-Hollman, McConnell, Rodriguez, & Albano, 2011.)

To address these shortcomings, the second version of the IGDIs incorporated an item response theory (IRT) measurement model, one which describes the location of items on the measurement scale within each IGDI measure, in relation to the construct underlying the measure. Benefits of an IRT measurement model include, among other things, item parameters which are not sample-specific and person parameters which are not test-specific (Embretson & Reise, 2000), simplified procedures for linking alternate forms to the same measurement scale (Kolen & Brennan, 2004), and estimates of error for individual items and students.

Although IRT modeling is common in K-12 testing, the measurement literature has not addressed issues in creating an IRT measurement scale for early (e.g., pre-kindergarten) educational assessments. Conceptual and logistical concerns make early literacy a unique and challenging context for developing IRT-based measures. This paper details the application of the Rasch model (Rasch, 1960) in the scaling of the IGDIs, demonstrating the feasibility of IRT modeling in the context of early literacy assessment.

The purpose of this study was to identify specific measures of early literacy, and items within these measures, which functioned well as indicators of growth and development with preschool children. The effectiveness of each measure was assessed in terms of a) scale reliability, b) standard errors of item and person parameter estimates, c) correlations with criterion measures, and d) overlap between item and person distributions, in other words, performance of students on the measure, in terms of average ability and dispersion, in relation to the IGDI items.

Method

Measures of Early Literacy

The first phase of the study involved a large number of measures, each containing a large number of items. The study began with IGDIs designed to measure the following seven early literacy activities: Sound Identification (sident), Letter Orientation (letter), Definitional Vocabulary (vocab), Picture Naming (picture), Which One Doesn't Belong? (which), Rhyming (rhyme), and Alliteration (allit). These measures were each conceptualized as falling into one of three literacy domains: alphabetic knowledge and print awareness, oral language, and

phonological awareness. Table 1 contains domains and task descriptions for each measure. Table 2 contains sample sizes and final reliability estimates.

Table 1. *Domains and Task Descriptions.*

Construct	Measure	Description
Phonemic Awareness	Rhyming	Identify the word, among three possible choices, that rhymes with a target picture
	Alliteration	Identify the picture, among three options, that starts with the same sound as the target picture
Alphabet Knowledge	Sound Identification	Identify the letter, among three options, that makes the sound verbalized by the administrator
	Letter Orientation	Identify the symbol, among three options, that is a letter
	Sound Blending	Verbalize the correct pronunciation of a word after presented with word segments separated by pauses
Oral Language	Definitional Vocabulary	Answer a question regarding the definition (i.e., function, attributes etc.) of a pictured object
	Which One Doesn't Belong?	Identify the object, among three options, that is categorically differently
	Picture Naming	Name pictures of common objects

The second phase of the study included only the reduced set of measures. The purpose of this phase was to identify specific items within the measures which had standard errors in an acceptable range and which targeted desired locations on the ability scale. As discussed below, Sound Identification and Picture Naming were selected for the second phase of the study.

Data Collection

Students were recruited from preschools in four states (Kansas, Minnesota, Ohio, Oregon). Because of constraints on available classroom time and student attention span, the full set of items for a given measure could not be administered in a single sitting. Instead, the roughly 44 items in each measure were divided into overlapping subsets of 15 items which were administered as close together as possible. Roughly the same students were used at each of three waves – students missing an earlier wave were not excluded from subsequent waves, and some students participated in an earlier wave but were absent in subsequent ones. However, no single student saw all 44 items.

Item subsets were linked together onto a common scale, within each measure, using an anchor set of 5 common items. Though few, the anchor items were carefully selected based on pilot data to have high discriminations and a spread of item difficulties.

Calibration

All item and person calibrations were obtained using the Rasch model (Rasch, 1960), which can be formulated to describe the probability of correct response to item *i*, $P = P(x_i = 1)$, as an exponential function of the location of the item on the trait (π_i), i.e., the item difficulty, and the location of the person on the trait (r_j), i.e., the person ability:

$$P(x_i = 1) = \frac{1}{1 + e^{\pi_i + r_j}}.$$

This formula is often written in terms of the log odds of correct response, which results in a simple summation of item and person parameters:

$$\log \frac{P}{1 - P} = \pi_i + r_j$$

The final reported item locations and person abilities were based on a fixed calibration using the Rasch model. Item location parameters were estimated using only data from the first time point. These estimates were then treated as fixed in the subsequent person calibrations (in the Winsteps software this involves obtaining an item file with the first calibration and using it as an item-anchor file in subsequent calibrations). These steps resulted in a single Rasch measurement scale, defined at wave 1. According to the default setting in Winsteps, the scales typically ranged from about -4 to +4 with a standard deviation below or near 1, and were centered around the mean item location for the measure (i.e., zero represents the average item location).

An extension of the fixed calibration, referred to here as a concurrent calibration, was also examined. In this analysis the item parameters were again estimated at wave 1 and fixed at these values for subsequent waves. However, the initial calibration also included item responses from waves 2 and 3 for any students not present at wave 1. As a result the calibration sample increased on average by over 100 students.

Data Cleaning

Preliminary analyses of the datasets revealed some aberrant scores and outliers. Cases were omitted from further analyses if the total score was outside the possible range (i.e., below zero or above the number of items in the measure), if observed score growth between any two waves was below the 5th or above the 95th percentile, or if the total score on the common items was zero while total score on the remaining items was greater than the 80th percentile for the measure.

The first and second of these criteria address outlying scores, ones potentially resulting from data entry issues. The third criteria addresses individuals who were assumed to be off-task or distracted in some way during the administration of the common items, but who were then able to score higher than 80 percent of the sample. It is assumed that these scores would contribute mostly confounding information to the scaling of the measures. The screening step reduced each dataset by 50 to 100 cases (Table 2 includes final counts).

Criteria

Scale reliabilities for the fixed and concurrent estimations were obtained using the model-estimated reliability formula implemented in the Winsteps software:

$$r_M = \frac{\sigma^2 - MS}{\sigma^2},$$

where $MS = \sum se_j^2 / J$ and σ^2 is the observed variance of the person parameter estimates for measure *M*. People misfitting the model, as identified using Rasch person-fit statistics, were excluded from the estimation of scale reliabilities.

Though the scale reliability takes into account the standard errors of person parameter estimates, it does not involve standard errors for the item parameter estimates. These were investigated both on an item-by-item basis and overall for each measure using test information and standard error functions. The standard error function is simply the square root of the inverse of the test information and it can be used to describe the precision of estimation across the score scale.

Correlations with criterion tests provide some evidence for the validity of each measure as an indicator of student ability. The criteria for comparison were selected according to the domain for each measure, as follows: for alphabet knowledge, the Test of Preschool Early Literacy (TOPEL) print awareness subtest; for phonological awareness, the TOPEL phonological awareness subtest; and for oral language, the Peabody Picture Vocabulary Test-4 (PPVT). Values reported here were not corrected for attenuation or restricted standard deviations, which may explain in part their moderate sizes. Another limiting factor in the criterion correlations was time between administration of the IGDI and the criterion. In some cases more than a day passed between administrations and this temporal effect may have also reduced the values.

Finally, the overlap in person and item distributions was examined using descriptive statistics and what are referred to as item-person maps. These plots display the person and item distributions on a single scale (also known as the theta or logit scale) which aides in identifying items which are too easy or difficult and points in the scale which do not contain sufficient item coverage.

Results

As noted above, the final results were based on the fixed calibration. The concurrent approach provided valuable information; however, additional research is necessary before these models can be used for reporting purposes.

Reliabilities were estimated for both the fixed and concurrent calibrations (see Table 2). For four of the seven measures (sident, allit, which and letter), aggregating scores across waves resulted in slightly increased correlations. For the remaining three measures (picture, rhyme, vocab) the concurrent calibration led to lower reliabilities in comparison to fixed. Decreases and slight increases in correlations indicate that the addition of students in the concurrent models brings little increase in observed variance in comparison to error variance. These results suggest that the relationship between people and items may not be consistent across waves.

Table 2. *Criterion Correlations and Reliability Coefficients.*

			Fixed		Concurrent	
Measure	Domain	Criterion r	r	N	r	N
sident	Alphabet Knowledge	.56	.60	662	.62	765
letter	Alphabet Knowledge	.53	.70	663	.72	761
allit	Phonological Awareness	.41	.42	681	.45	781
rhyme	Phonological Awareness	.43	.56	706	.55	801
picture	Oral language	.68	.80	695	.70	956
which	Oral language	.49	.72	637	.73	766
vocab	Oral language	.61	.57	654	.50	805

Note: Criterion correlations involve a different criterion measure for each domain.

A simple way to assess consistency of the model across time is to examine consistency of item ordering from one wave to the next. Item location parameters were compared based on calibrations at wave 1 and wave 3. Plots of these values for four measures are contained in Figure 1. These plots and the correlations between parameter estimates (ranging from .74 to .96) show varying degrees of similarity in item location ordering.

Standard errors for item parameter estimates varied considerably by measure, and, as expected, were highest for outlying items with low variability in item response (i.e., having a high proportion either correct or incorrect) and fewer students seeing the item. Figure 2 contains standard error functions, the inverse of the test information function, across all items for the same four measures displayed in Figure 1. Picture, with the largest sample size, has the smallest standard errors across the theta scale. The remaining three measures have similar standard error plots, and all are centered close to zero and roughly symmetrical, indicating that the distributions of items were roughly symmetrical as well (this is also evident in Figure 1). The centering at zero is an arbitrary choice made in the calibration software and does not carry meaning from one measure to another. Standard errors functions were also examined for the concurrent calibration. The standard error functions for the fixed and concurrent calibrations were nearly identical, though concurrent values were slightly lower in the tails.

Correlations with criterion measures are included in Table 2. The IGDI measures were only correlated moderately positively with their respective criterion tests. Picture Naming had the highest criterion correlation of .68, and Alliteration had the lowest value at .41. To some extent these correlations are attenuated, due, at least, to the low reliabilities of the IGDI measure themselves. Note that no criterion value is higher than the reliability for the corresponding measure. These values are also limited to the extent that the criterion measures are unreliable to some degree.

Item-person distributions display the overlap between a set of items and a sample of students by plotting both on the same theta scale. Figures 3 through 10 contain plots for the seven measures. Person distributions are all skewed to some extent. For sident, allit, rhyme, and which the skew is noticeably positive, whereas for picture it is clearly negative. Letter and vocab have slightly longer tails for lower scores; however, they also both have a large number of scores bunched toward the top of the scale as well, making these distributions the least normal in appearance of all the measures.

The item and person distributions overlap to different degrees from one measure to the next. Table 3 contains descriptive statistics for all seven measures, based on raw scores and theta

estimates. Since the item distributions are centered at zero, the mean theta estimates for students indicate the mean difference between the item and person distributions. At wave 1, students had positive mean theta scores for sident, picture, which, letter, and vocab. Picture had the highest mean of 1.82. Overall, the items were easier for the students on these measures. The remaining measures, allit and rhyme, had negative means at wave 1, indicating that items were more difficult for students, on average.

Not shown in the item-person plots are the number of students with scores of zero. Table 3 contains these values, as percentages, at each wave. At wave 1, Rhyming and Alliteration had the highest percentages, 25% and 28%, and Definitional Vocabulary and Picture Naming had the lowest, 2% and 3%. As expected, the numbers of zero scores decreased at waves 2 and 3.

Having considered each of these factors (reliability, item parameter standard error, descriptive statistics, criteria correlations, and overlap between item and person distributions) two measures were chosen to be included in the next phase of the study. The measures all showed some promise, with similar standard error functions and good overlap in item and person distributions. However, scale reliabilities were deemed acceptable (above .60) only for sident, picture, which, and letter. The final consideration was the anticipated longevity of the measure, in terms of feasibility of writing additional items of appropriate difficulty. As the item-person distributions show (Figures 3 through 10), for each measure there was considerably more spread in the person distributions than in the item distributions. This additional person spread also tended to be near the top of the scale, where there were fewer items to discriminate among higher ability students. After considering which measures could be most easily supplemented with additional items of low and high difficulty, Sound Identification and Picture Naming were selected for phase 2 of the study.

Conclusion

This paper presents the results of a large-scale administration and Rasch calibration of seven measures of early literacy assessment. Model-estimated reliabilities, criterion correlations, standard errors, and item and person distributions suggest that not all seven measures are suited for calibration and scaling using the Rasch model. Instead, two measures were selected based on their potential for flexibility and longevity in subsequent stages of development.

As suggested, the phases of analysis presented here are merely two components of a larger and more complex assessment design model (for details see Wackerle-Hollman et al., 2011; Bradfield et al., 2011). The purpose of this study was to reduce a large set of measures to a more limited and more promising subset. The next phases of development involve the refinement of item-sets to be used in the identification of students in need of instructional intervention. Another paper (Rodriguez, Albano, McConnell, Wackerle-Hollman, & Bradfield, 2011) describes the standard setting process, a slight modification of the contrasting group method, which was used to select a cut score range to support the placement of students into intervention.

Table 3. *Descriptive Statistics for Theta and Raw Scores.*

		Theta				Raw						
Measure	Wave	Mean	SD	Min	Max	Mean	SD	Max	Skew	Kurt	%Zero	N
Sident	1	0.18	1.76	-2.69	4.42	5.55	3.85	18	0.46	-0.26	14%	518
	2	1.20	1.85	-2.64	4.55	8.55	4.66	20	0.20	-0.46	6%	524
	3	1.42	1.93	-2.54	4.55	9.43	5.00	20	0.02	-0.66	7%	593
	2-1	1.05	1.61	-4.06	6.53	3.07	3.60	10	-0.04	-0.82		518
	3-1	0.22	1.45	-4.54	6.14	0.87	4.05	13	-0.29	0.38		524
Letter	1	1.22	1.94	-2.75	4.74	10.52	5.41	20	-0.37	-0.67	9%	518
	2	2.16	1.71	-2.75	4.71	14.19	4.82	20	-0.92	0.29	2%	518
	3	2.43	1.65	-2.71	4.71	15.12	4.76	20	-1.15	1.70	2%	583
	2-1	0.94	1.48	-3.41	5.46	3.68	3.99	12	0.16	-0.81		518
	3-1	0.32	1.42	-4.73	5.98	0.99	4.15	20	0.31	2.20		518
Vocab	1	0.98	1.59	-4.50	4.18	7.99	2.95	16	-0.26	0.12	2%	509
	2	1.82	1.46	-4.32	4.39	10.69	3.16	20	-0.18	0.33	1%	513
	3	2.41	1.36	-4.32	4.39	12.01	3.06	20	-0.72	1.31	1%	594
	2-1	0.86	1.45	-3.05	5.35	2.74	2.56	8	0.07	-0.80		508
	3-1	0.59	1.47	-6.52	5.64	1.31	3.27	9	-0.15	0.23		513
Picture	1	1.82	1.94	-4.99	5.89	16.97	8.57	40	-0.15	-0.53	3%	550
	2	2.24	1.72	-3.91	5.92	20.15	8.48	40	-0.41	-0.30	1%	549
	3	2.47	1.62	-4.97	5.96	22.13	9.06	40	-0.31	-0.43	1%	627
	2-1	0.43	1.25	-4.10	3.90	3.22	6.22	15	-0.17	-0.81		548
	3-1	0.22	1.15	-5.16	3.62	1.81	7.27	25	-0.28	0.01		549
Which	1	0.28	1.65	-3.55	4.51	9.53	4.89	20	-0.05	-0.61	7%	500
	2	1.04	1.51	-3.50	4.49	12.55	4.72	20	-0.37	-0.52	2%	505
	3	1.27	1.64	-3.42	4.49	13.61	4.99	20	-0.85	0.10	3%	570
	2-1	0.79	1.34	-3.13	5.00	3.12	3.62	11	0.11	-0.80		498
	3-1	0.28	1.36	-3.79	5.88	1.20	3.96	18	0.27	1.20		505
Rhyme	1	-0.46	1.79	-3.50	4.28	5.02	4.52	20	0.84	0.15	25%	559
	2	0.27	1.84	-2.78	4.29	7.97	5.47	20	0.29	-0.80	14%	562
	3	0.49	1.83	-2.75	4.29	9.07	5.62	20	0.14	-0.93	10%	634
	2-1	0.76	1.45	-3.84	4.44	3.01	3.96	11	-0.03	-0.70		553
	3-1	0.22	1.38	-4.82	6.54	1.13	4.08	15	-0.10	0.62		561
Allit	1	-0.90	1.56	-4.07	4.02	3.73	3.30	15	0.72	0.04	28%	532
	2	-0.28	1.42	-2.92	4.17	5.70	3.70	19	0.71	0.87	11%	541
	3	-0.14	1.50	-2.92	4.34	6.43	4.20	20	0.77	0.74	11%	617
	2-1	0.64	1.45	-3.19	4.90	2.02	3.15	9	0.07	-0.68		529
	3-1	0.12	1.35	-5.83	4.18	0.71	3.65	11	-0.20	0.85		540

Note: Minimum raw score values were all zero and have been excluded. Theta estimates are based on fixed calibration. Picture Naming included 40 items and remaining measures included 20.

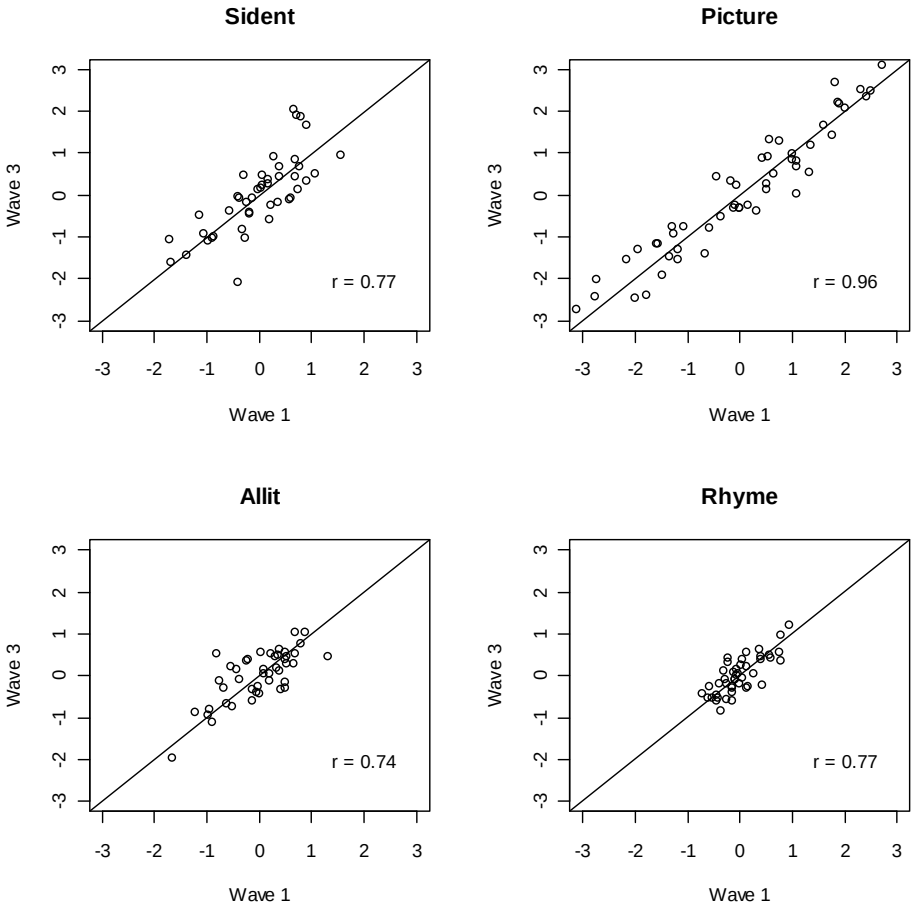

Figure 1. Item calibrations based on waves 1 and 3 for four measures.

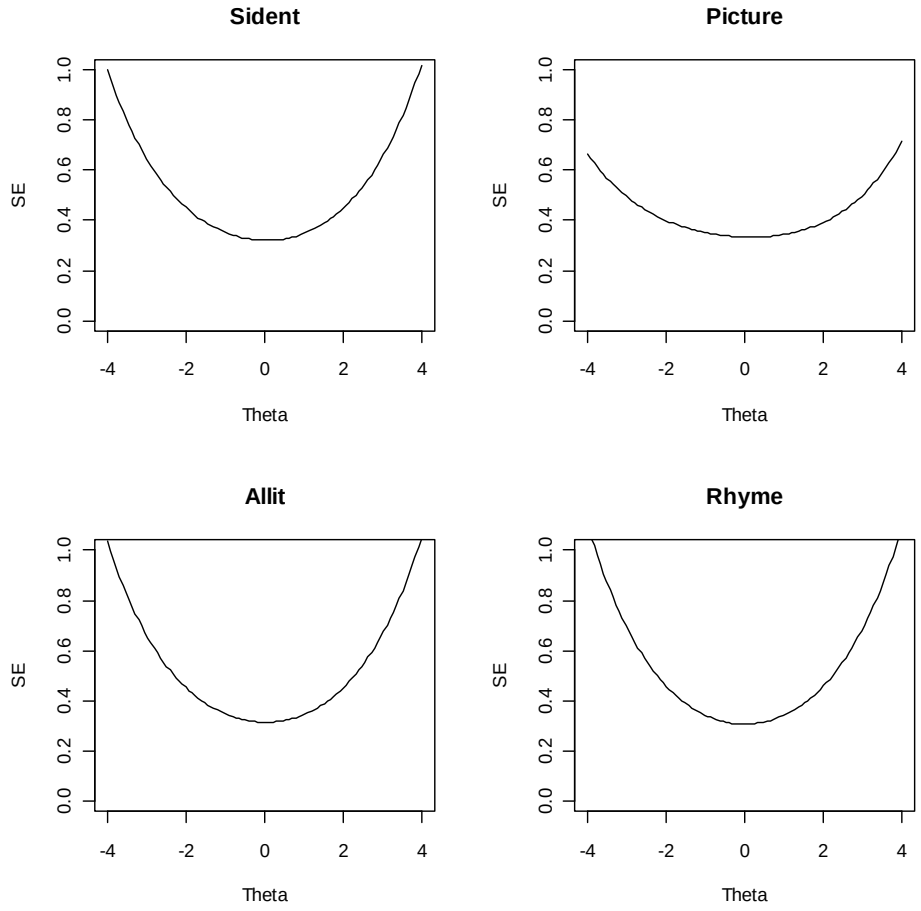

Figure 2. Test error functions (standard errors across the score scale) for four measures.

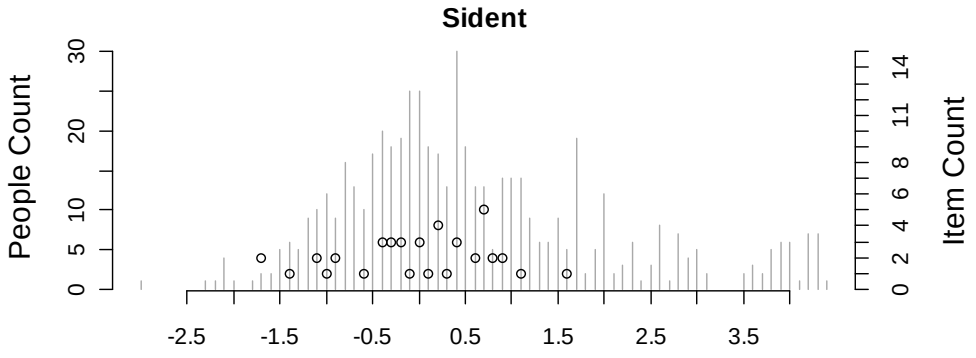

Figure 3. Distributions of items (points) and people (lines) for Sound Identification.

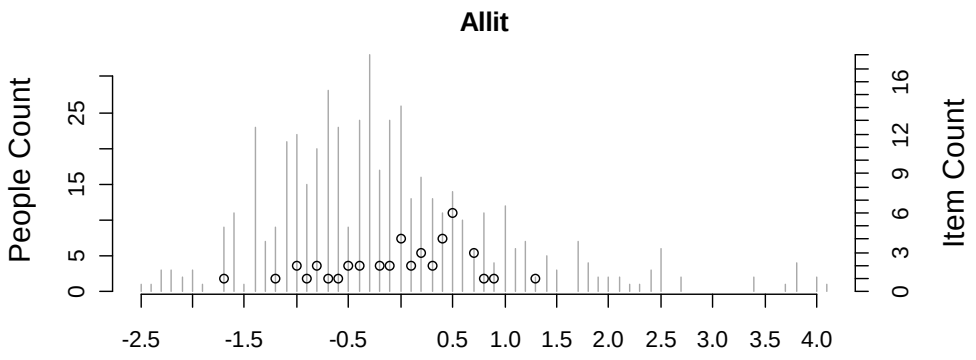

Figure 4. Distributions of items (points) and people (lines) for Alliteration.

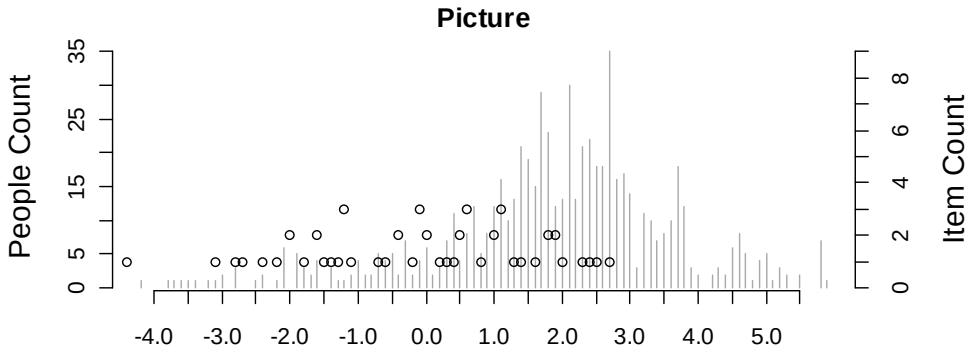

Figure 5. Distributions of items (points) and people (lines) for Picture Naming.

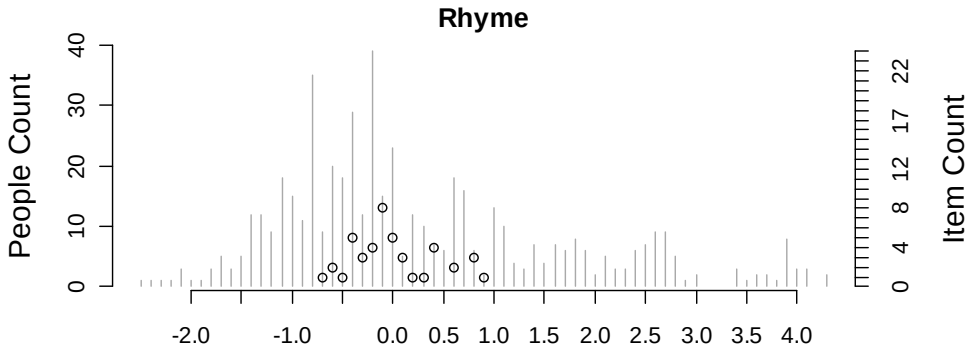

Figure 6. Distributions of items (points) and people (lines) for Rhyming.

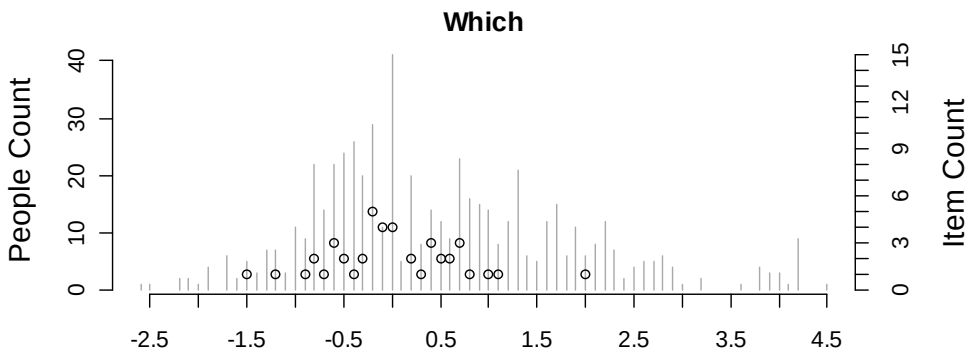

Figure 7. Distributions of items (points) and people (lines) for Which One Doesn't Belong.

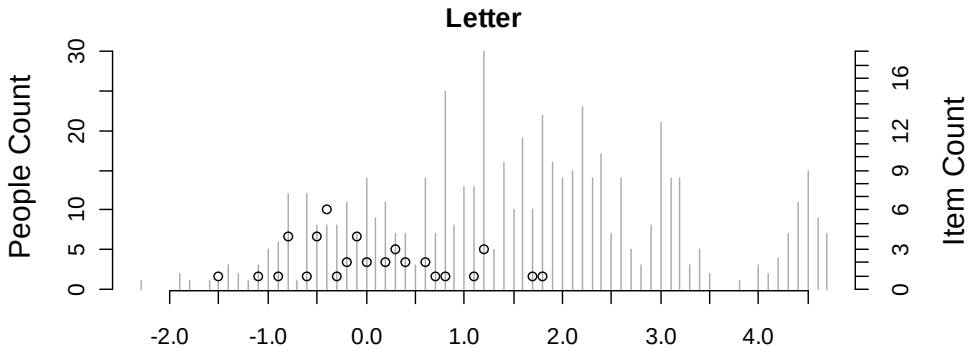

Figure 9. Distributions of items (points) and people (lines) for Letter Orientation.

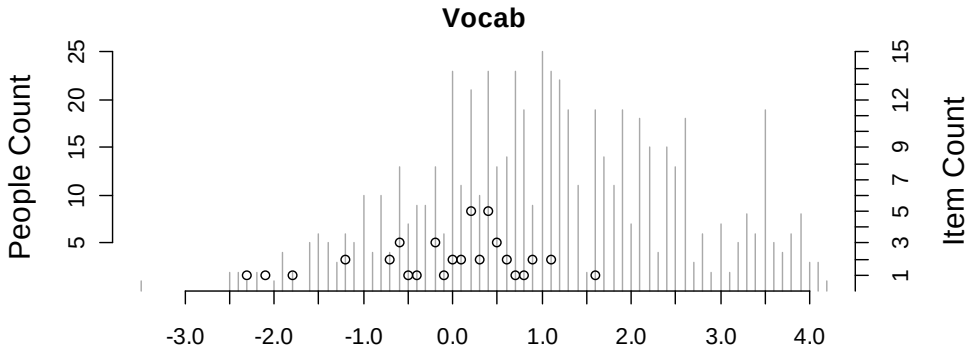

Figure 10. Distributions of items (points) and people (lines) for Definitional Vocabulary.

References

Bradfield, T., Wackerle-Hollman, A., McConnell, S., Rodriguez, M. C. & Albano, A. D. (2011, April). *Construct identification to support early literacy measurement.* Paper presented at the meeting of the National Council for Measurement in Education, New Orleans, LA.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L. S., & Deno, S. L. (1991). Effects of curriculum within curriculum-based measurement. *Exceptional Children, 58*, 232-243.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Rodriguez, M. C., Albano, A. D., McConnell, S., Wackerle-Hollman, A. & Bradfield, T. (2011, April). *Standard setting with innovative measures of early literacy: Contrasting Groups.* Paper presented at the meeting of the National Council for Measurement in Education, New Orleans, LA.

Wackerle-Hollman, A., Bradfield, T., McConnell, S., Albano, A. D., & Rodriguez, M. C. (2011, April). *Task development and item analysis in innovative measures of early literacy*. Paper presented at the meeting of the National Council for Measurement in Education, New Orleans, LA.