HHS Public Access

Author manuscript

Assess Eff Interv. Author manuscript; available in PMC 2015 September 02.

Published in final edited form as:

Assess Eff Interv. 2011 December ; 37(1): 17–25. doi:10.1177/1534508411407761.

Efficiency of Predicting Risk in Word Reading Using Fewer, Easier Letters

Yaacov Petscher and
Florida State University

Young-Suk Kim
Florida Center for Reading Research

Abstract

Letter-name identification has been widely used as part of early screening to identify children who might be at risk for future word reading difficulty. The goal of the present study was to examine whether a reduced set of letters could have similar diagnostic accuracy rather than a full set (i.e., 26 letters) when used as a screen. First, we examined whether a hierarchical scale existed among letters by using a Mokken scale analysis. Then, we contrasted diagnostic accuracy among the 5, 10, 15, and 20 easiest letters, with all 26 letters by using receiver operating characteristic (ROC) curves and indices of sensitivity, specificity, positive predictive power, and negative predictive power. Results demonstrated that a hierarchical scale existed among items in the letter-name knowledge test. In addition, assessing students on the easiest 15 letters was not statistically distinguished from all 26 letters in diagnostic accuracy. The implications of the results for the use of a Mokken scale analysis in educational research are discussed.

Keywords

Mokken scale analysis; letter name knowledge; screening

The early identification of students who are at risk for future difficulties with reading achievement is one of the most important goals from recent legislation, such as the No Child Left Behind Act. Considerable effort has been demonstrated in research to develop measures that will increase the diagnostic efficiency of an assessment, yet many assessment batteries contain large numbers of items, take a long period of time to collect data, and include items that are not appropriate given the students level of ability on the task. The primary goal of the present study was to examine and compare the diagnostic accuracy of several selected cut points on a letter-name knowledge assessment in predicting failure on a nationally norm referenced word reading test when using a full set letter-name identification test (i.e., testing all 26 letters) versus a reduced set with fewer and easier letter names. Children's letter-name knowledge was used as a predictor because it has shown to be one of the best predictors of early literacy achievement (Adams, 1990; National Research Council [NRC], 1998; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004), and has been included in

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

many assessments for young children such as the *Florida Assessment for Instruction in Reading* (FAIR; Foorman, Torgesen, Crawford, & Petscher, 2009), *Test of Early Reading Ability* (TERA; Reid, Hresko, & Hammill, 1981), *Test of Preschool Early Literacy* (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007), *Texas Primary Reading Inventory* (TPRI; Foorman, Fletcher, & Francis, 2004), and *Woodcock-Johnson III Diagnostic Reading Battery* (Woodcock, Mather, & Schrank, 1997).

BACKGROUND AND CONTEXT

Recent efforts to enhance children's literacy acquisition have focused on the early identification of children who are at risk of future reading difficulties and how to provide appropriate interventions to identified individuals. At the core of such a framework (e.g., response to intervention, RTI) is universal screening, which is typically administered at the beginning of the school year in order to identify children according to their risk status in target areas (e.g., reading). Effective screens are brief and target skills that are highly predictive of later reading outcomes (Jenkins, Hudson, & Johnson, 2007). For young pre-readers, prior research has shown that foundational skills such as phonological awareness, print concept, oral language (e.g., vocabulary), and letter knowledge are predictive of children's early literacy skills (Lonigan, 2006; Lonigan, Burgess, & Anthony, 2000; McCardle, Scarborough, & Catts, 2001; Scarborough, 1998; Schatschneider et al., 2004). As such, many widely used screeners such as the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS; Good, Kaminski, Smith, Laimon, & Dill, 2001) and *Phonological Awareness Literacy Screening* in kindergarten (PALS; Invernizzi, Meier, Swank, & Juel, 1999), and AIMSweb include tasks that assess children's phonological awareness and letter knowledge. Although the amount of evidence supporting classification or diagnostic accuracy of these screeners varies (see http://www.rti4success.org/), existing evidence provide support for the importance of these skills in predicting and accurately classifying students to whether they are at risk for developing reading disability or not (Good et al., 2001; O'Connor & Jenkins, 1999; Speece, Mills, Ritchey, & Hillman, 2003).

The focus in the present study is children's knowledge of letter names. Numerous studies have shown that letter-name knowledge is highly predictive of early literacy acquisition (i.e., word reading and spelling; Foulin, 2005; NRC, 1998; Scanlon & Vellutino, 1997; Schatschneider et al., 2004; Speece et al., 2003) and accurately classifies children for their potential risks in word reading and spelling (O'Connor & Jenkins, 1999; Speece et al., 2003). Theoretically, letter names provide critical phonetic cues for letter sounds, and letter-sound knowledge is the foundation of early literacy skills (Foulin, 2005; Share, 2004; Treiman & Kessler, 2003). In addition, letter-name knowledge may be an indirect measure of early literacy-rich environment such as experiences with books and print (Adams, 1990; Scanlon & Vellutino, 1997). Thus, letter-name knowledge tasks are almost invariably included in many assessment batteries for emergent and early readers. The format and content coverage of letter-name knowledge varies across measures. However, a widely used format is identification of letter names in which children are asked to provide names of letters (upper and lower cases) presented in a random order. Assessment of letter-name knowledge does not typically take long even when sampling all the letters (e.g., 5 minutes) because letter knowledge is a very well-defined, small domain (i.e., in English there are a

total of 52 upper- and lowercase letters). However, there are at least two reasons to consider a reduced number of letters in a letter-name knowledge task. First, because a letter-name knowledge task is typically part of larger batteries of assessment, a reduced number of letters can help cut administration time of overall assessment. The second, and more important, reason pertains to the difficulty of individual letters. Previous studies have shown, and our data confirm, that children acquire names of certain letters more easily than others. For example, children in North America tend to acquire letter names for A and O earlier (or more easily) than V (Treiman & Kessler, 2003).

Several factors play a role in the variation of the ease of letter acquisition, which is beyond the scope of the present study, but interested readers should see a relevant discussion in Treiman and Kessler (2003). Although administering a letter knowledge task with a full set of letters is necessary and helpful for instructional planning or progress monitoring, excluding letters that tend to be acquired later and thus may be difficult to many children (e.g., V) might be more efficient for screening purposes, especially when typically administered to young children (preschoolers and kindergartners). That is, including fewer, easier letters might be of sufficient utility in identifying children who might be at risk for later reading problems.

The inclusion of easier items might be particularly important for many children from low-SES backgrounds who might not have sufficient learning experiences with all the letters (NRC, 1998), and inclusion of all the letters might unduly overwhelm young children, for whom maintaining attention span and motivation is critical for accurate assessment (Rathvon, 2004). The key, of course, is that diagnostic (or classification) accuracy is maintained in the reduced version. In the present study, we used Mokken scale analysis to examine whether an ordinal, hierarchical scale exists within a letter-name knowledge test and how full and reduced sets of letter names predict risk on an end-of-year, high-stakes reading test.

MOKKEN SCALE APPROACH

The purpose of the analyses was to examine the extent to which an item selection process relates to the prediction of student risk on an end-of-year, high-stakes achievement test using a nonparametric item response model. A Mokken scaling approach (Mokken, 1997) allows for testing whether an ordinal, hierarchical scale exists within a set of items. If a Mokken scale exists for a set of items, then a positive endorsement of a dichotomous item (e.g., “yes” or “correct”) will provide an indication as to what others have been positively endorsed. An implicit assumption to Mokken scaling is invariant item ordering; that is, the ordering of items according to the probability of a correct response for a given value of theta will be the same across all values of theta and subgroups from the population. Conceptually, Mokken scaling is a nonparametric model in item response theory that is a probabilistic version of Guttman scaling but is less deterministic in nature (Watson, 1996). Nonparametric item response theory models allow researchers to estimate models with less stringent assumptions, such as the ability to order individuals on an ordinal scale (Sijtsma, 1998).

From an applied perspective, Mokken scaling is very similar to the Rasch model. Both approaches make similar assumptions about local item independence, and neither technique is robust to violations of double monotonicity (i.e., every person generates an equivalent ordering of item difficulties) or monotone homogeneity (i.e., every item generates an equivalent ordering of ability scores for students; Meijer, Sijtsma, & Smid, 1990). The most distinguishing feature between Rasch and Mokken scaling pertains to the ordering of items. While Rasch models have strong assumptions about item ordering on a ratio basis, Mokken scaling is strictly ordinal. Under circumstances where one is interested in banking items or equating scores across different populations, more conventional item analyses such as the Rasch model are more appropriate. However, when a question of interest concerns hierarchies among items within a scale, Mokken scaling has been recommended as a desired procedure for item analysis. Moreover, research has shown that because of the strict model assumptions of the Rasch model, it is often applied when the number of items is large, and the fewer the number of items that exist within a scale, the worse the fit of the data to the Rasch model. As such, the Mokken scale fits these instances well and performs as a hybrid between the Guttman and Rasch models.

Hierarchies are a useful convention to studying the relation among items on a given scale, and are especially popular in medical and psychological applications (Kempen, Myers, & Powell, 1995; Watson, Deary, & Austin, 2007; van Boxel, Roest, Bergen, & Stam, 1995), where individuals' scores on an assessment can be directly related to their ability on the latent construct of interest. Similar to Rasch models, Mokken scales have the advantage that the sum of items scores in a Mokken scale is a measure for the order of the latent trait being identified (Moorer, Suurmeijer, Foets, & Molenaar, 2001).

By using a combination of Mokken scaling and screening accuracy analyses, we aimed to study the following:

1. Does a hierarchy of difficulty exist for the FAIR letter knowledge task?
2. What is the screening accuracy of the letter knowledge task using the Mokken scale, for cut points of 5, 10, 15, and 20 letters compared to all 26 letters?

Given prior research about differential ease of letters (Kim, Petscher, Foorman, & Zhou, 2010; McBride-Chang, 1999; Treiman & Kessler, 2003), we expected that a hierarchy of difficulty would exist. However, we did not have a specific hypothesis regarding screening accuracy because of lack of prior research.

METHOD

Participants

Data were collected on 613 kindergarten students who were randomly selected from nine elementary schools across four major school districts in urban, semi-urban, and rural counties, which were representative of the demographics for the state of Florida. More than one third of students (36%) were White, followed by Black (27%) and Latino (17%); 52% of all participants were male; 57% were eligible for free or reduced-price lunch; and 38% were identified as English language learners.

Measures and Procedures

Students were assessed in the fall of 2009 on the *Florida Assessments for Instruction in Reading* (FAIR; Foorman, Torgesen, Crawford, & Petscher, 2009) letter-names and letter-sounds task in one of three randomized orders of the 26 letters. Testers asked the child for the name of the letter and then the sound. Both uppercase and lowercase letters were presented simultaneously. In the analysis, data on letter-name identification, not sound, were used. The maximum score was 26.

Students were also administered the Word Reading section (20 items) of the *Stanford Early School Achievement Test* (SESAT; Madden, Gardner, & Collins, 1983) at the end of the school year (i.e., May). Students' percentile rank scores on the SESAT were dichotomized for the purpose of statistical modeling where scores at or above the 40th percentile were coded as 0 and below the 40th percentile were coded as 1. Although this cut point might appear to be a low threshold for acceptable performance on a high-stakes outcomes test, 26 states have used the 40th percentile as the guideline for measuring proficiency in reading (American Institutes for Research, 2007). Thus, it was appropriate to use this here as well. The observed base rate of failure on the SESAT in the sample was 40%.

Data Analytic Plan

Mokken scaling analysis was used as a method to test the extent to which a hierarchy of item difficulty existed in the FAIR letter-name knowledge task. Once this level of hierarchy was ascertained, the screening accuracy for subsets of letters was used to examine how well identified subsets of letters predicted future risk status on the norm-referenced word reading test.

Compared to parametric item response theory models, which typically estimate items using a top–down approach, Mokken scaling uses a bottom–up item selection procedure (Meijer et al., 1990) that entails (1) selecting a pair of items from all possible items where the scalability coefficient for the pair (i.e., *H*) is greater than zero, (2) testing if *H* for the first pair of items is the largest among the coefficients among all possible pairs of items, (3) adding an additional item to the pair that correlates positively, (4) testing if the *H* of the added item is significantly larger than zero, and (5) testing if *H* for the added item is larger than 0.30. This process repeats as long as the conditions of Steps 3 to 5 are met and that it maximizes the overall estimate of *H*. Also known as Loevinger's coefficient, *H* measures the extent to which ordering of the items is consistent and is bound between zero and one (Molenaar & Sijtsma, 2000). Mokken, Lewis, and Sijtsma (1986) noted that as *H* increases, there is greater confidence about the ordering of persons and items using the total score from the tested items. Thus, *H* can represent the extent to which participants from the population can be ordered according to their performance on the set of items from the scale. Values greater than 0.30 have been used as the minimum acceptable threshold to indicate scalability, and $H > 0.40$ is an indication of a strong scale.

In addition to reviewing the magnitude of *H* for each item, it is important to examine other diagnostic information, including ρ (i.e., reliability) and violations of monotonicity and double monotonicity. The latter may be ascertained by an analysis where values >80

indicate violations and estimates between 0 and 40 indicate a scale where violations may be attributed to sampling error. In addition, a matrix containing the proportions of relative position responses to pairs of items, called the *P*(++) matrix, should be inspected. Similar to the way a Guttman data set would look like, the values in the *P*(++) matrix should increase from right to left, indicating that harder items are at the right of the matrix, and easier items are to the left of the matrix. Values should also increase from top to bottom. Although the scalability coefficient describes the nature of scaling for a set of items, the magnitude of *H* does not indicate the ordering of items for the scale. Rather, the means for the items can be used to rank order difficulty.

Screening accuracy was tested using a Receiver Operating Characteristic (ROC) curve, and examining the area under the curve (AUC) differences between the different sets of items (i.e., derived Mokken scale compared to all letters). The AUC is a probability index ranging from 0.5 to 1.0 and provides the probability of the independent variable correctly classifying a pair of individuals where one student is at risk and the other is not. Values closer to 0.5 indicate that the independent variables do not classify well, whereas an AUC of 1.0 reflects test scores that perfectly classifies individuals. Swets (1988) indicated that the AUC is effective as an indicator of effect size, with values of at least 0.80 considered appropriate.

In addition, a series of 2 × 2 contingency tables are used to estimate conditional probability indices that provide differential perspectives on the accuracy of risk identification: sensitivity (SE), specificity (SP), positive predictive power (PPP), negative predictive power (NPP), and the overall percentage of students who were correctly classified (OCC). Sensitivity describes the proportion of all individuals who failed an outcome (e.g., <40th percentile on the SESAT) who were also identified as at risk according to a cut point on a selected screening assessment (e.g., scoring <10 letters correct on the FAIR letter task), specificity describes the proportion of all students who passed an outcome (e.g., ≥40th percentile on the SESAT) who were also identified as not at risk on the screener (e.g., scoring ≥10 letters on the FAIR), positive predictive is the percentage of all students who were identified as at risk on the screen who ultimately failed the outcome test, negative predictive power is the percentage of all students who were identified as not at risk on the screen who passed the outcome test, and the overall correct classification is the proportion of students who were correctly classified as either at risk (i.e., sensitivity) or not at risk (i.e., specificity).

Screening assessments often vary in classification accuracy because of the type of gold standard chosen, the base rate of the problem in the sample, and the method with which cut points are selected (Petscher, Kim, & Foorman, 2011), and though no universally agreed on threshold exists for indices of screening accuracy, many researchers attempt to meet estimates of at least .80 when maximizing a particular statistic (e.g., sensitivity or positive predictive power). However, others have recommended that both sensitivity and specificity should be at least .90, as this corresponds to a kappa agreement coefficient of .80, typically denoted as excellent agreement for pairs of ratings (Feuerman & Miller, 2008). Although this represents an idealistic threshold, many commonly used screens range in their screening accuracy from as low as .60 for sensitivity and specificity to greater than .90 (National Center on Response to Intervention, 2010).

In general, the approach to selecting the most appropriate cut score for the letter task was to balance the trade-offs in sensitivity, specificity, and predictive power as well as to account for any subgroup differences in performance on the screen. Initially, gender, race (i.e., White, Black, Latino), socioeconomic differences (i.e., free or reduced-price lunch [FRL]), and language status (i.e., English-language learner [ELL]) groups were examined for differential descriptive differences in the proportion of correct letters for each selected cut point on the screen. Finally, through a series of logistic regressions, we tested the interaction between risk identification on the screen and the different demographic variables in order to assess any differential accuracy in predicting risk on the SESAT for different groups of individuals. A fail-to-reject decision for an interaction term would indicate that no predictive bias existed for a subgroup at the cut point selected.

RESULTS

Descriptive statistics for the letters (Table 1) indicated that the easiest letters were O and X, with 87% of the participants correctly knowing the name. Conversely, the most difficult letter for the sample was the letter V (p = .60). The mean percentage correct across all letters was .73 (SD = .07). Using the procedure described above, the data were analyzed using Mokken Scaling Program 5.0 (Molenaar & Sijtsma, 2000). From the 26 letters entered into the analysis, only one scale was extracted, and all 26 letters were included in the scale. A summary of the individual H coefficients for each letter as well as the percentage of respondents correctly knowing the letter names are reported in Table 1. All items demonstrated H >0.30, ranging from 0.54 (letter W) to 0.78 (letter A). This indicated that scalability existed among the set of items. Scalability of the entire set of items was 0.63, with an associated classical test theory reliability estimate of ρ = 0.97. A check of the monotonicity among items showed that all item scores were far less than the critical value of 40, with the largest observed estimate of 4 for the letter W. Similarly, the P(++) matrix (available from the first author on request) corroborated that minimal violations of double homogeneity occurred. As can be gleaned from Table 1, a direct correspondence between H and the item difficulties (i.e., p) does not exist. In fact, the correlation between the two was only r = .08.

In order to create subsets of items for a reduced scale, we opted to test the differences between the 5, 10, 15, and 20 easiest letters and all 26 letters in the ROC curve analysis. Because the Mokken analysis suggested that scalability exists among these letters, the sets of the easiest letters represented those items that had the highest nested probabilities of success based on item ordering. The ROC curve analysis (Figure 1) demonstrated differences in the prediction of risk on the SESAT depending on the cut-off score chosen. For most of the curves, there was a great deal of overlap and intersection when comparing the curves across the values of the y-axis; however, most noticeably, the curve for a cut point of five letters was well below that of the other curves for most points on the graph. The largest difference in the slopes of the curves among the sets of letters was observed at the sensitivity level of .48. However, this finding is practically unimportant as most researchers would not consider sensitivity <.80 to be a meaningful cut point. A post hoc analysis using the AUC index from Table 2 was conducted to test if any of the sets of letters was the most predictive of SESAT word reading risk using methods outlined by Hanley and McNeil

(1983) for comparing the area under ROC curves. Results indicated that significant differences were observed when comparing the set of 5 letters to 20 letters ($z = -2.14$) or all 26 letters ($z = -2.16$).

The results in Table 3 further highlight the relation between the selected cut points and the screening accuracy in predicting risk on the SESAT by way of screening indices. As previously discussed, it was important to strike a balance in the trade-off of index strength when selecting the most appropriate cut point. When administering any fewer than all 26 letters, the sensitivity values were well below the .90 or .80 threshold. Conversely, specificity only reached levels of .80 or greater when a cut point of 5, 10, or 15 letters were used. Negative predictive power remained relatively stable across all cut points, ranging from .71 for 5 letters to .89 for 26 letters, while somewhat more variable for positive predictive power, ranging from .46 for 26 letters to .76 for 5 letters. The overall correct classification index was fairly consistent in magnitude for the cut points of 5 to 20 (range = .72–.77), but dropped to .60 when all 26 letters were used. From these screening index results, it appeared that a cut-off score of 15 letters would be the most appropriate in predicting risk on the SESAT and provided the greatest balance across the five indices.

However, to comprehensively determine which cut point would be most appropriate, the descriptive statistics of the proportion correct by demographic subgroups (Table 4) and the results from the differential accuracy analysis (Table 5) were used. When examining the drop-offs in the proportion correct in Table 4, it appeared that the 15-letter cut point provided the best balance in drop-off. Although there is no established rule for an optimum level of difficulty, a cut point at 15 letters appears to have reasonable ease for many demographic groups. The exception is for Latino students and ELL students, who had a success rate of 38% and 42%, respectively, when 15 letters were used. Moreover, the results from the series of logistic regressions (Table 5) suggested that no differential bias was observed in the prediction of risk on SESAT for any subgroup.

DISCUSSION

The present study examined whether a hierarchical scale existed in a letter-name knowledge test, and if so, whether there is a difference in screening accuracy between a reduced set of letters and a full set of letters. A stochastic Mokken scale analysis and ROC curves were used for the former and latter, respectively. Results demonstrated that a hierarchical scale existed among items in a letter-name identification test and that only one scale was extracted. This finding was not surprising given the recent evidence regarding letter names as a unidimensional construct (Foorman, Torgesen, Crawford, & Petscher, 2009).

Although differences existed in the screening accuracy across the selected cut points, the trade-offs across the five studied indices suggested that fewer, easier letters could be used to predict failure on the SESAT when compared to all 26 letters. Using conventional methods from developmental psychologists, the scalability of the items as one construct allowed for a reorganization of the items into a hierarchy according to the difficulty of letters. These findings also suggested that the easiest 15 letters (i.e., O, X, A, B, S, C, Z, R, E, I, K, W, P, D, and T) may be used as a screen as effectively as all 26 letters with relatively minimal

losses in screening accuracy. Although the sensitivity for 15 letters was .55 compared to .90 for all 26 letters, the .45 gain in specificity, .26 gain in positive predictive power, and .17 gain in overall correct classification, as well as the minimal .10 loss in negative predictive power (i.e., .89 to .79) appeared to provide the best balance in selection of the cut point. Moreover, the 15 letters appear to achieve balance when considering performances of various demographic groups. Thus, a more parsimonious measure of letter-name knowledge with 15 letters may be an alternative when considering this screen in the battery of other assessments. The classical test theory estimate of internal consistency was observed to be .94 for 15 items and .96 for all items. Thus, although some level of average precision is lost by using fewer items, the overall estimate is still above the clinically important threshold of .90 (Nunnally & Bernstein, 1994). Although some assessments include a partial list of letters in a letter-name identification test (e.g., Kaufman Survey of Early Academic and Language Skills, and TOPEL), the findings of the present study are unique in that the selected 15 items are easiest and based on an empirically examined hierarchy of items. To our knowledge, this is the first study that examined hierarchy of items in an important predictor of early literacy skills. The organization of hierarchically ordered items provides an intuitive description of individual student performance as well as comparisons among students. By using the Mokken scale, when a student is able to correctly name one of the more difficult letters, there is the ability to say they have a higher probability to name the preceding items.

Implications for Practice

When considering the elements of an effective screen for predicting failure on a subsequent outcome test, it is important to consider both statistical and practical needs. Statistically, the tasks in the screen should demonstrate sufficient screening accuracy across a set of different indices so that children are identified correctly with precision according to their risk level in their future reading difficulties. Practically, assessments in screens should be brief, easy to administer, score, and interpret (Johnson, Jenkins, Petscher, & Catts, 2009). Brevity is an important aspect in the context of RTI in which multiple screening measures are typically used. Although the use of multiple screening measures yields better classification accuracy (Foorman et al., 1998; O'Connor & Jenkins, 1999), the cost (time and personnel) of administering several measures, and potential student frustration level with the difficulty of measures are important considerations in schools where the allocation of limited resources is a critical decision. The findings of the present study indicate that a briefer assessment of letter names can be used to predict later reading ability as accurately as a full-length assessment at least for kindergartners. This offers practical utility for teachers and practitioners who use valuable class time for assessing students. Furthermore, because the selected 15 items are easiest of the 26 letters, this briefer assessment may prevent children from being unnecessarily exposed to difficult items and help ensure that children will perform with attention and motivation.

Acknowledgments

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences (R305A100301).

References

Adams, MJ. Beginning to read: Thinking and learning about print. MIT Press; Cambridge, MA: 1990.

American Institute for Research. *Reading First* state APR data. Author; 2007.

Feuerman M, Miller AR. Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. Journal of Evaluation in Clinical Practice. 2008; 14:930–933. [PubMed: 19018927]

Foorman B, Torgesen J, Crawford E, Petscher Y. Assessments to guide reading instruction in K-12: Decisions supported by the new Florida system. Perspectives on Language and Literacy. 2009; 35(5):13–19. [PubMed: 25598861]

Foorman, BR.; Fletcher, JM.; Francis, DJ. Texas Primary Reading Inventory. McGraw-Hill; New York, NY: 2004.

Foorman, BR.; Fletcher, JM.; Francis, DJ.; Carlson, CD.; Chen, D.; Mouzaki, A.; Taylor, RH. Technical report: Texas Primary Reading Inventory. 1998 ed. Center for Academic and Reading Skills and University of Houston; Houston, TX: 1998.

Foulin JN. Why is letter-name knowledge such a good predictor of learning to read? Reading and Writing: An Interdisciplinary Journal. 2005; 18:129–155.

Good, RH.; Kaminski, RA.; Smith, S.; Laimon, D.; Dill, S. Dynamic indicators of basic early literacy skills. 5th ed. University of Oregon; Eugene: 2001.

Hanley JA, McNeil BJ. A method for comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148:830–843.

Invernizzi, M.; Meier, JD.; Swank, L.; Juel, C. Phonological awareness literacy screening. University of Virginia; Charlottesville: 1999.

Jenkins JR, Hudson RF, Johnson ES. Screening for at-risk readers in a response to intervention framework. School Psychology Review. 2007; 36:582–600.

Johnson ES, Jenkins JR, Petscher Y, Catts HW. How can we improve the accuracy of screening instruments? Learning Disabilities Research & Practice. 2009; 24:174–185.

Kempen GIJM, Myers AM, Powell LE. Hierarchical structure in ADL and IADL: Analytical assumptions and applications for clinicians and researchers. Journal of Clinical Epidemiology. 1995; 48:1299–1305. [PubMed: 7490592]

Kim Y-S, Petscher Y, Foorman B, Zhou C. The contributions of phonological awareness and letter-name knowledge to letter sound acquisition—A cross-classified multilevel model approach. Journal of Educational Psychology. 2010; 102:313–326.

Lonigan CJ. Development, assessment, and promotion of pre-literacy skills. Early Education and Development. 2006; 17:91–14.

Lonigan CJ, Burgess SR, Anthony JL. Development of emergent literacy and early reading skills in preschool children: Evidence from a latent variable longitudinal study. Developmental Psychology. 2000; 36:596–613. [PubMed: 10976600]

Lonigan, CJ.; Wagner, RK.; Torgesen, JK.; Rashotte, CA. TOPEL: Test of Preschool Early Literacy. PRO-ED; Austin, TX: 2007.

Madden, R.; Gardner, EF.; Collins, CS. Stanford Early School Achievement Test. Harcourt Brace Jovanovich; New York, NY: 1983.

McBride-Chang C. The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. Merrill- Palmer Quarterly. 1999; 45:285–308.

McCardle P, Scarborough HS, Catts HW. Predicting, explaining, and preventing children's reading difficulties. Learning Disabilities: Research & Practice. 2001; 16:230–239.

Meijer RR, Sijtsma K, Smid NG. Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. Applied Psychological Measurement. 1990; 3:283–298.

Mokken RJ, Lewis C, Sijtsma K. Rejoinder to "The Mokken Scale: A critical discussion. Applied Psychological Measurement. 1986; 6:279–285.

Mokken, RJ. Nonparametric models for dichotomous responses. In: van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. Springer; New York, NY: 1997. p. 351-367.

Molenaar, IW.; Sijtsma, K. User's manual MSP5 for Windows. IEC ProGAMMA; Groningen, The Netherlands: 2000.

Moorer P, Suurmeijer Th. P. B. M. Foets M, Molenaar IW. Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases, and COPD) in the Netherlands. Quality of Life Research. 2001; 10:637–645. [PubMed: 11822796]

National Center on Response to Intervention. Screening tools chart. Author; Washington, DC: 2010.

National Research Council. Preventing reading difficulties in young children. National Academy Press; Washington, DC: 1998.

Nunnally, JC.; Bernstein, IH. Psychometric theory. 3rd ed. McGraw-Hill; New York, NY: 1994.

O'Connor RE, Jenkins JR. The prediction of reading disabilities in kindergarten and first grade. Scientific Studies of Reading. 1999; 3:159–197.

Petscher Y, Kim Y-S, Foorman BR. The importance of predictive power in early screening assessments: Implications for placement in the RTI framework. Assessment for Effective Intervention. 2011; 36(3):158–166.

Rathvon, N. Early reading assessment: A practitioner's handbook. Guilford; New York, NY: 2004.

Reid, DK.; Hresko, WP.; Hammill, DD. The Test of Early Reading Ability. Pro-ED; Austin, TX: 1981.

Scanlon DM, Vellutino FR. A comparison of the instructional backgrounds and cognitive profiles of poor, average, and good readers who were initially identified as at risk of reading failure. Scientific Studies of Reading. 1997; 1:191–215.

Scarborough HS. Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. Annals of Dyslexia. 1998; 48:115–136.

Schatschneider C, Fletcher JM, Francis DJ, Carlson CD, Foorman BR. Kindergarten prediction of reading skills: A longitudinal comparative analysis. Journal of Educational Psychology. 2004; 96:265–282.

Share DL. Knowing letter names and learning letter sounds: A causal connection. Journal of Experimental Child Psychology. 2004; 88:213–233. [PubMed: 15203298]

Sijtsma K. Methodology review: Nonparametric IRT approaches to the analysis of dichotomous items scores. Applied Psychological Measurement. 1998; 22:3–31.

Speece DH, Mills C, Ritchey KD, Hillman E. Initial evidence that letter fluency tasks are valid indicators of early reading skill. Journal of Special Education. 2003; 36:223–233.

Swets JA. Measuring the diagnostic accuracy of diagnostic systems. Science. 1988; 240:1285–1293. [PubMed: 3287615]

Texas Primary Reading Inventory. McGraw-Hill; New York, NY: 2006–2008. TPRI

Treiman, R.; Kessler, B. The role of letter names in the acquisition of literacy. In: Kail, R., editor. Advances in child development and behavior. Vol. Vol. 31. Academic Press; San Diego, CA: 2003. p. 105-135.

Van Boxel YJJM, Roest FHJ, Bergen MP, Stam HJ. Dimensionality and hierarchical structure of disability measurement. Archives of Physical Medicine and Rehabilitation. 1995; 76:1552–1555.

Watson R. The Mokken scaling procedure (MSP) applied to the measurement of feeding difficulty in elderly people with dementia. International Journal of Nursing Studies. 1996; 33:385–393. [PubMed: 8836763]

Watson R, Deary I, Austin E. Are personality trait items reliably more or less "difficult"? Mokken scaling of the NEO-FFI. Personality and Individual Differences. 2007; 43:1460–1469.

Woodcock, R.; Mather, N.; Schrank, FA. Woodcock Johnson III Diagnostic Reading Battery. Riverside Publishing; Rolling Meadows, IL: 1997.

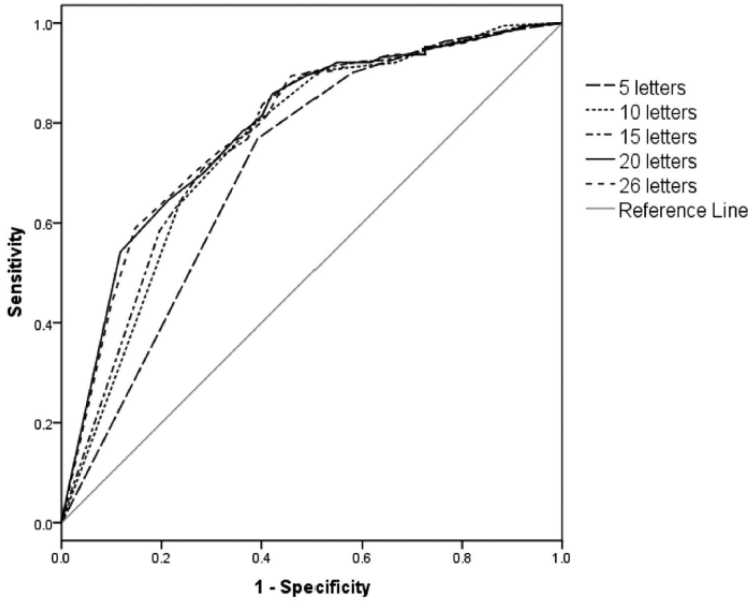

Figure 1.
Receiver operating characteristic (ROC) curve comparison of selected letter groupings

Table 1

Mokken Scalability Coefficient (H) Compared With Item Difficulty (*p*)

Item	Letter	Loevinger *H*	p	SD
1	O	0.63	.87	.33
2	X	0.64	.87	.34
3	A	0.79	.87	.35
4	B	0.67	.87	.37
5	S	0.63	.78	.42
6	C	0.62	.77	.42
7	Z	0.60	.77	.42
8	R	0.64	.76	.23
9	E	0.64	.75	.43
10	I	0.62	.75	.43
11	K	0.59	.75	.44
12	W	0.54	.74	.44
13	P	0.67	.73	.44
14	D	0.65	.73	.45
15	T	0.60	.72	.56
16	M	0.60	.71	.47
17	L	0.64	.70	.46
18	F	0.65	.70	.46
19	H	0.62	.70	.46
20	J	0.60	.68	.47
21	N	0.65	.68	.47
22	G	0.63	.67	.47
23	Y	0.60	.67	.47
24	Q	0.63	.64	.48
25	U	0.68	.72	.49
26	V	0.72	.60	.49
	Scale	0.63		

Table 2

Comparison of Area Under the Curve (AUC) Among Tested Item Clusters

			95% Confidence Interval	
Letter Cut Points	**AUC**	**SE**	**Lower Bound**	**Upper Bound**
5	.71	.03	.64	.78
10	.76	.03	.69	.82
15	.77	.03	.71	.83
20	.79	.03	.73	.84
26	.79	.03	.73	.84

Table 3

Screening Accuracy of Letter Clusters

Letter Cut Points	SE	SP	NPP	PPP	OCC
5	.27	.95	.71	.76	.72
10	.36	.93	.73	.73	.73
15	.55	.88	.79	.72	.77
20	.63	.78	.80	.60	.73
26	.90	.43	.89	.46	.60

Note. SE = Sensitivity, SP = Specificity, NPP = Negative Predictive Power, PPP = Positive Predictive Power, OCC = Overall Correct Classification.

Table 4

Proportion of Individuals by Subgroup Performing At or Above Selected Letter Cut Points

	Gender		Race			Lunch Status		Language Status	
Letter Cut Points	**Male**	**Female**	**White**	**Black**	**Latino**	**FRL**	**Non-FRL**	**ELL**	**Non-ELL**
5	85	94	94	91	66	87	88	63	92
10	78	90	92	86	53	79	84	53	88
15	67	84	85	77	38	67	77	42	79
20	59	73	75	70	25	57	68	26	71
26	30	37	43	32	12	23	35	16	36

Note. FRL = Student is eligible for Free or Reduced-Price Lunch; ELL = English Language Learner.

Table 5

F-Statistic for the Interaction Term in Detecting Differential Accuracy

Letter Cut Points	Gender	Race	Lunch Status	Language Status
5	0.84	0.30	0.71	0.89
10	0.26	0.62	0.30	0.02
15	0.28	2.22	0.60	0.02
20	0.28	1.21	1.44	0.22
26	0.28	0.68	0.21	0.91

Note. $p = ns$ for all *F*-statistic estimates.