HHS Public Access

Author manuscript

J Learn Disabil. Author manuscript; available in PMC 2016 May 01.

Published in final edited form as:

J Learn Disabil. 2015 ; 48(3): 227–238. doi:10.1177/0022219413495568.

Examining the Measurement Precision and Invariance of the Revised Get Ready to Read!

Amber L. Farrington and **Christopher J. Lonigan**

Florida State University

Abstract

Children's emergent literacy skills are highly predictive of later reading abilities. To determine which children have weaker emergent literacy skills and are in need of intervention, it is necessary to assess emergent literacy skills accurately and reliably. In this study, 1,351 children were administered the Revised Get Ready to Read! (GRTR-R), and an item response theory analysis was used to evaluate the item-level reliability of the measure. Differential item functioning (DIF) analyses were conducted to examine whether items function similarly between subpopulations of children. The GRTR-R had acceptable reliability for children whose ability level was just below the mean. DIF for a small number of items was present for only two comparisons—children who were older versus younger and children who were White versus African American. These results demonstrate that the GRTR-R has acceptable reliability and limited DIF, enabling the screener to identify those at risk for developing reading problems.

Keywords

assessment; reading; early literacy; preschool age

Children's academic skills affect a number of aspects of their lives. Children with stronger academic skills receive more opportunities for education, have more choices in terms of which career they pursue as an adult, and are more likely to have an overall higher quality of life. In contrast, children with weaker academic skills have lower rates of academic retention and are more likely to experience teenage pregnancies and to have behavioral problems (Bennett, Brown, Boyle, Racine, & Offord, 2003; Matson & Haglund, 2000). Academic skills are substantially influenced by children's reading skills (Chall, Jacobs, & Baldwin, 1990), which consist of decoding and comprehending written language (Lonigan, Schatschneider, & Westberg, 2008). In 2009, the National Assessment of Educational Progress (NAEP) reported that only two thirds of students were reading at or above a basic level, and, of these students, only half were reading at or above a proficient level. This low

Corresponding Author: Amber L. Farrington and Christopher J. Lonigan, Department of Psychology, Florida State University, 1107 W. Call St., Tallahassee, FL 32306-4301, USA. amlynnf@gmail.com and lonigan@psy.fsu.edu.

All differential item functioning (DIF) analyses were also conducted using “anchor” items in which one item was selected from the *Revised Get Ready to Read!* to be used as the control against which other items were compared to detect the presence of DIF. This technique was used with Items 5, 7, 14, 18, and 21. Using this technique to assess DIF did not change the results. The same parameters had DIF in all analyses, supporting the use of mean DIF to identify bias and nonuniform DIF.

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

rate of reading achievement was present for both fourth and eighth grade students and has been present since 1992, when the NAEP originated. The two thirds of students who are not reading at a proficient level would benefit from being provided with resources and interventions to improve their reading skills. Interventions that are offered to children at early ages can have lasting beneficial effects on children's academic skills and can enhance their cognitive and emotional development (Campbell & Ramey, 1995; Karoly, Kilburn, Bigelow, Caulkins, & Cannon, 2001). To provide children with early intervention, it is necessary to identify children who are at risk for having difficulty with reading achievement at young ages.

Reading-related skills begin to develop and can be measured as early as preschool in the form of “emergent literacy skills.” Emergent literacy skills are the developmental precursors of conventional reading skills (Whitehurst & Lonigan, 1998) and include print knowledge, phonological -awareness, and oral language. Print knowledge includes skills such as letter-name knowledge, directionality of print (e.g., text is read from left to right), and letter-sound knowledge. Phonological awareness is the ability to detect and manipulate the sound structure of words, independent of their meanings. Oral language skills involve a child's ability to convey and understand meaning in spoken language and include such skills as vocabulary, syntax, and narrative.

Emergent literacy skills are highly predictive of word decoding, reading comprehension, and spelling measured in elementary school (Lonigan et al., 2008). Assessment of emergent literacy skills can help in the identification of preschool children who are more likely to develop better reading skills and preschool children who are more likely to be at risk for having difficulties acquiring reading skills. Determination of children who are at risk for developing reading problems provides the opportunity to offer resources and interventions to these children with the goal of improving their literacy skills and preventing them from falling further behind in the classroom. Therefore, it is important to evaluate the means by which children are assessed to ensure that children with weaker literacy-related skills are being identified accurately.

There are a number of methods to assess whether a child is at risk for developing reading problems. These include diagnostic assessment, informal assessment, and screening, each of which has strengths and weaknesses (Lonigan, Allan, & Lerner, 2011). Informal assessments are often performed by the teacher and take little time to complete. However, unlike diagnostic assessment and screening, informal assessments are not standardized assessments. There is limited evidence that informal assessments are reliable and valid across administrations (Lonigan et al., 2011). Although diagnostic assessments are typically the most reliable and valid means of assessing a child, they are both time-consuming and costly to administer. In addition, diagnostic assessments often can be administered only by individuals with specialized training. Screening measures address the limitations of diagnostic assessments. Screening measures are brief standardized measures that evaluate a child's overall strengths and weaknesses in a specific skill area. Screening measures take less time to administer and are more cost-effective than diagnostic assessments, and screening measures can be administered by individuals, including a child's teacher or teaching assistant, with minimal training.

At present, there are few available screening measures to evaluate children's emergent literacy skills. One exception is the *Get Ready to Read!* (GRTR; Whitehurst & Lonigan, 2001), a screening measure designed to assess preschool children's print knowledge and phonological awareness skills. The original GRTR was a 20-item measure. Validation of this version of the GRTR (N = 342 preschool children) demonstrated that it correlated significantly with the *Developing Skills Checklist* (r = .69; Whitehurst, 2001). Molfese, Molfese, Modglin, Walker, and Neamon (2004) reported that the GRTR is significantly and concurrently correlated with measures of vocabulary, environmental print, phonological processing, and rhyming among a sample of preschool children from low-income backgrounds (rs ranged from .25 to .51). Molfese et al. (2006) found comparable results for children's performance on letter identification in the fall of their preschool year with the GRTR score in the spring of their preschool year (r = .48). With regard to predictive validity of the GRTR, Phillips, Lonigan, and Wyatt (2009) reported significant longitudinal correlations between GRTR scores and a number of emergent literacy and reading tasks (e.g., blending, elision, rhyming, letter knowledge, and word identification) among 3- to 5-year-old children across both short- and long-term prediction intervals (20–35 months; rs ranged from .25 to .40).

The original 20-item GRTR was revised to increase the range of the measure, enabling its use with preschool children with relatively high levels of emergent literacy skills. Six new items were added to the measure and one item was removed, resulting in a 25-item measure. Lonigan and Wilson (2008) reported that the revised GRTR (GRTR-R) had good reliability (alpha value of .88), moderate item– total correlations (rs greater than .30), and item difficulties that ranged from .37 to .81 in a development sample of 819 children 3 to 6 years old who were representative of the U.S. population. They also evaluated the criterion validity of the GRTR-R by comparing children's scores on the screener to their scores on a standardized measure of emergent literacy skills, the *Test of Preschool Early Literacy* (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007). The GRTR-R correlated significantly with all subtests (Print Knowledge, Phonological Awareness, and Definitional Vocabulary) and the overall Early Literacy Index score (rs ranged from .39 to .72; Lonigan & Wilson, 2008). In addition, Wilson and Lonigan (2010) compared the GRTR-R to the *Individual Growth and Development Indicators* (IGDIs) using receiver operating characteristic curves. They found that the GRTR-R had both higher accuracy and better overall correct classification than did the IGDIs when classifying children as at risk or not at risk for reading problems. Although the GRTR-R was better than the IGDIs at classifying children correctly, overall correct classification values ranged from .24 to .73 and the median shared variance of the GRTR-R with children's scores on the TOPEL was .19.

Although both classical test theory (CTT) and item response theory (IRT) techniques provide information regarding the psychometric characteristics of a measure, only CTT techniques have been used previously to provide information about the GRTR-R. The information obtained through IRT analyses has a number of benefits. IRT provides an estimate of error across the range of obtained scores, is sample invariant, and provides information regarding the utility of individual items as a function of test takers' latent abilities (measured as their level of theta) through the generation of item parameters

(Embretson & Reise, 2000). Up to three parameters can be generated for each item in IRT. The discrimination parameter, or a parameter, estimates the degree to which items discriminate children's ability on the underlying dimension (i.e., theta). A highly discriminative item provides a more precise estimation of theta. The difficulty parameter, or b parameter, for an item provides information about the level of skill indexed by that item. It represents the level of theta at which the item has an equal probability of being answered correctly or incorrectly (Embretson & Reise, 2000). The third parameter, the c parameter, is the pseudo-guessing parameter. This parameter provides an estimate of the probability that correct item endorsement occurred by chance.

In a one-parameter logistic (1PL) model, the item discrimination parameters are constrained across items whereas the difficulty parameters are allowed to vary across items. In contrast, a two-parameter logistic (2PL) model generates parameter estimates for both the difficulty and the discrimination parameters. A three-parameter logistic (3PL) model is used with multiple-choice assessments and generates the pseudo-guessing parameter in addition to the two parameters provided in a 2PL model. Simulation studies indicate that a sample size of at least 1,000 participants and a test with at least 50 items is required to generate a 3PL model accurately (Hulin, Lissak, & Drasgow, 1982; Lord, 1968). The current study included a sample of 1,351 children, which is large enough in size to generate a 3PL model if the test length is greater than or equal to 50 items (Lord, 1968). However, because the GRTR-R contains only 25 items, a 3PL model would have considerable difficulty reaching convergence (Hulin et al., 1982). Therefore, the current study did not use a 3PL model and no pseudo-guessing parameter was generated. Rather, both a 2PL model and a 1PL model were generated and a chi-square comparison of log likelihood values, as described in Embretson and Reise (2000), was performed to determine that a 2PL model provided a significantly better fit to the data than did a 1PL model.

For screening purposes, it is useful to know the range of abilities over which a screener has acceptable levels of reliability in addition to knowing the degree of precision it has in categorizing children. Ideally a screening measure should have high levels of reliability around the range of abilities at which decisions will be made. For a measure like the GRTR-R, which is used to identify children at risk of having difficulty reading, high levels of precision are needed at values of theta that are below the mean. This allows for the accurate identification of children who are likely to be at risk for difficulties developing adequate reading achievement skills. Identifying these children at early ages enables them to be provided with educational resources in an effort to prevent them from falling further behind their peers and to enable them to make more rapid gains.

In addition to providing precise estimates of the degree of error across the range of ability assessed by a measure, IRT allows for the evaluation of differential item functioning (DIF). DIF analysis can be used to determine the degree to which a particular item (or set of items) in a measure functions similarly or differently between groups. DIF analyses can determine which item parameters differ significantly between groups of participants and how these parameters differ. DIF on the discrimination parameter indicates that the degree to which the item discriminates a child's ability level differs between groups, whereas DIF on the

difficulty parameter indicates that children with the same underlying skill level differ in the likelihood that they will pass or fail an item as a function of group membership.

The purpose of this study was to use IRT analysis to examine the degree of measurement precision across the level of early literacy ability measured by the GRTR-R. An additional goal of the study was to examine whether items on the GRTR-R function similarly for different groups of children. Specifically, DIF analyses were performed to evaluate whether items on the GRTR-R function similarly between boys and girls, between older and younger children, and between children of different racial/ethnic groups.

METHOD

Participants

Data for this study came from three separate samples in which the GRTR-R was administered to preschool-age children. In one sample, 819 children completed the screener as part of the norming sample for the GRTR-R. These children were selected to be representative of the U.S. population in terms of geographic region, gender, race/ethnicity, maternal education, and child exceptionality status (Lonigan & Wilson, 2008). In the second sample, 268 children from northern Florida completed the GRTR-R at the start of their preschool year as part of a larger battery of assessments administered for a project examining children's academic development. All children whose parents completed and returned consent forms were selected to participate in the study. Most (78%) of these children were White, and both genders were equally represented. In the third sample, 264 children were administered the GRTR-R at the start of their preschool year to identify children who were at risk for later reading difficulties. Similar to the second sample, all children whose parents completed and returned consent forms were selected for participation in the study. These children attended preschools in northern Florida, were primarily White (66%), and boys and girls were equally represented. The combined sample consisted of 1,351 preschool children who ranged in age from 31 to 74 months at the time of testing (M = 53.87 months, SD = 7.74). The sample was approximately equal with respect to sex (51% male, 49% female). The majority of children in the sample were White (64%), 14% were African American, 12% were Hispanic/ Latino, 5% were Asian, and the remaining 6% of the sample were of another race/ethnicity.

Measure

The GRTR-R is a 25-item multiple-choice test that takes 5 to 10 min to administer and assesses children's phonological awareness and print knowledge skills. Phonological awareness items require the child either to manipulate sounds in a way that forms a word and then to identify the picture represented by the word, or to identify which of four pictures rhymes with or begins with the same sound as a stimulus word. For example, Item 17 requires the child to blend together “sea” … “shell” and choose which of four pictures depicts a seashell. Print knowledge items require the child either to identify letters and sounds or to identify which of four pictures contains a word or letters. For example, Item 5 requires the child to find the picture that shows the name of a cereal from among four different pictures of cereal boxes— only one of which contains a word. Each item on the

GRTR-R is scored as correct or incorrect, and all items are administered (i.e., there are no basal or ceiling rules) and totaled to provide an overall score of children's emergent literacy skills. The alpha for the sample used in this study was .85.

Procedure

All children were tested while they were in preschool. Parents provided consent for their children to participate in the three studies. In the first sample listed above, the GRTR-R screener was administered to obtain norming data and evaluate the psychometric properties of the screener. In the remaining two samples, the GRTR-R screener was administered as part of a larger battery of emergent literacy assessments. Children in all samples were assessed individually in a quiet area at their school. For the first sample, research assistants, teachers, and classroom aides administered the GRTR-R. All of these individuals had received training in administration of the GRTR-R. For the second and third samples, undergraduate and graduate students in psychology or related fields and paid research assistants administered the GRTR-R. All of these individuals had undergone general training in administering assessments to young children and specific training in administering the GRTR-R.

RESULTS

Out of 33,775 possible data points, 169 were missing. A full information maximum likelihood algorithm was used in Mplus (Version 5.1; Muthén & Muthén, 2007) to account for the less than 1% missing data. In all IRT analyses, item-level data were treated as categorical. The average total score on the GRTR-R was 15.14 (SD = 5.50). Significant differences in performance on the GRTR-R existed between subgroups. Average scores for White children (M = 15.58, SD = 5.33) were significantly higher than average scores for both African American children (M = 14.05, SD = 5.88), $F(1, 1046) = 12.91$, $p < .01$, and Hispanic/Latino children (M = 13.66, SD = 5.41), $F(1, 1018) = 17.00$, $p < .01$. Scores for African American children did not differ significantly from scores for Hispanic/Latino children, $F(1, 336) = 0.38$, $p = .54$. There were no significant differences in GRTR-R scores between boys (M = 15.07, SD = 5.66) and girls (M = 15.21, SD = 5.34), $F(1, 1346) = 0.34$, $p = .56$. Data were examined for normality and were determined to have no significant skew; however, there was evidence for a platykurtic distribution of GRTR-R scores (i.e., scores were distributed such that the tails of the distribution contained fewer observations than would be expected in a normal distribution). Although the distribution of scores on the screener was platykurtic, IRT is robust to violations of normality (Embretson & Reise, 2000).

There are two primary assumptions of IRT analysis: The measure has to be unidimensional (i.e., measures only one domain), and the assumption of local independence has to hold (i.e., responses to each item are conditional on the latent ability level assessed and are not conditional on responses to other items; Embretson & Reise, 2000). Because local independence and unidimensionality are isomorphic with one another, a measure that meets the assumption of unidimensionality can also be assumed to meet the assumption of local independence (Lord, 1968). To test the dimensionality of the GRTR-R, we conducted a

modified parallel analysis. First, a simulated data set with known unidimensionality was constructed. Then exploratory nonlinear factor analysis was conducted in Mplus (Version 5.1; Muthén & Muthén, 2007) to compare the eigenvalues generated using the obtained data to those generated using the simulated data set with known unidimensionality. Scree plots of the obtained data and the unidimensional simulated data had a high degree of similarity, indicating that the GRTR-R is unidimensional. The first eigenvalue generated through exploratory factor analysis with the obtained data explained approximately 35% of variance on the GRTR-R, which is close to the recommended amount (40%) of variance that a unidimensional measure should capture (Reckase, 1979; Sinar & Zickar, 2002), and subsequent eigenvalues accounted for little additional variance (4% to 6% of the variance) in the model.

Fit statistics of the unidimensional model were compared to fit statistics of a two-factor model that was empirically derived through exploratory factor analysis and fit statistics of a two-factor model that was theoretically constructed by separating items into phonological awareness and print knowledge items (Table 1). No items were allowed to cross-load in either of the two-factor models. For the empirical two-factor model, Items 1 to 15, 17 to 19, and 22 clustered together to form the first factor, whereas Items 16, 20, 21, and 23 to 25 formed the second factor. These items did not differ systematically in content across the two empirical factors. The two factors were highly correlated in both the empirical (r = .81) and theoretical (r = .86) two-dimensional models. An incremental chi-square test indicated that both two-factor models provided a statistically significant improvement in fit over the unidimensional model; however, the incremental chi-square test is sensitive to large sample sizes, as in this study. Comparative fit index (CFI) and Tucker–Lewis index (TLI) values that are greater than or equal to .95 indicate a good fit of the model to the data (Hu & Bentler, 1999). In addition, small root mean square error of approximation (RMSEA) values indicate better fit, with RMSEA values less than or equal to .05 indicating good fit (Hu & Bentler, 1999). Although only the empirical two-factor model had a CFI value greater than .95, the CFI of the theoretical two-factor model (.949) and the TLI of the empirical two-factor model (.949) were close to meeting the recommended value to indicate good model fit. All RMSEA values indicated good model fit. Parameter invariance was evident on most items across the one-factor and two-factor models. Both discrimination and difficulty parameter estimates from the one-factor model were highly correlated with the estimates from both of the two factor models (rs from .95 to 1.0). However, discrimination parameters for Items 16, 21, and 23 to 25 (all of which loaded on to the second factor in the empirical model) differed across models, indicating that these items do not fit the unidimensional model as well as the other items on the GRTR-R. Although the findings were mixed regarding the unidimensionality of the GRTR-R, Stout's (1990) theory that essential unidimensionality is satisfactory to conduct an IRT analysis, in addition to the results of the modified parallel analysis and the size of the first eigenvalues, support the interpretation of the GRTR-R as an essentially unidimensional measure.

To assess the reliability of items on the GRTR-R, IRT analyses were used to generate item parameter estimates. A 2PL model provided a significantly better model fit than a 1PL model, as indicated by a chi-square comparison of log likelihood values from each model,

$\chi^2(1, 25) = 316.34$, $p < .001$ (Embretson & Reise, 2000). In addition, the 2PL model had lower Akaike information criterion (AIC; 37499.55) and Bayesian information criterion (BIC; 37775.49) values than did the 1PL model (AIC = 37762.69, BIC = 37913.67), indicating better model fit. Consequently, subsequent analyses used the 2PL model to generate discrimination and difficulty parameter estimates (see Table 2). The difficulty parameter values for 19 of the 25 items on the GRTR-R were less than zero, indicating that, on these items, the ability level required to have a greater than chance probability of correct item endorsement was below the mean. Of the 19 items with negative difficulty parameters, six items had difficulty values below −1, indicating that children whose emergent literacy skill levels were higher than one standard deviation below the mean had a greater than 50% chance of responding correctly to these items. Of the six items with difficulty parameter values greater than 0, none were greater than 1, indicating that children whose emergent literacy skill levels were between zero and one standard deviations above the mean had a greater than 50% chance of responding correctly to all items on the GRTR-R.

The only GRTR-R items with discrimination parameters greater than 1 were Items 6, 7, and 22, all of which assess print knowledge skills. These items have the greatest degree of precision in discriminating children's ability levels on the underlying dimension of theta. Five items, four of which assess phonological awareness, had low discrimination parameters (less than .50), thereby providing the lowest degree of precision in discriminating children's ability levels.

Standard errors across the range of ability are shown in Figure 1. Standard errors were below 0.42 for the range of theta from −1.8 to 0.80, indicating that approximately 72% of children can be assessed with adequate reliability, or reliability equal to or higher than .80 (Nunnally & Bernstein, 1994). Standard errors were below 0.37 for the range of theta from −1.5 to 0.50, indicating that approximately 62% of children can be assessed with moderate reliability, or reliability equal to or higher than 0.85 (Nunnally & Bernstein, 1994; Ponterotto & Ruckdeschel, 2007).

Differential Item Functioning

To identify DIF based on gender, age, and race/ethnicity on the GRTR-R, a multiple indicator multiple cause model in Mplus was used. DIF analyses were conducted under the assumption that the mean DIF of the entire measure was 0 (see Note 1). Follow-up DIF analyses were conducted using the IRTLRDIF program (Version 2.0; Thissen, 2001) to provide convergent findings of DIF. Whereas Mplus identified which items functioned differently across groups, the IRTLRDIF program provided information concerning whether the DIF for an item existed on the discrimination parameter, the difficulty parameter, or both. After conducting all DIF analyses and adjusting for Type I error using the linear step-up procedure (Benjamini & Hochberg, 1995), only 13 out of 125 comparisons (approximately 10% of comparisons; 25 comparisons each for gender and age and 25 comparisons for each of the three combinations of race/ ethnicity) yielded significant DIF.

Meade's VisualDF 1.3 program (Meade, 2010) was used to calculate effect sizes for DIF as well as several additional statistics related to differential functioning on the GRTR-R. This program provides only one effect size for each item; it does not provide separate effect size

estimates for each parameter generated. The calculated effect sizes follow the same guidelines as those for Cohen's *d*, such that an effect size of .30 represents a small effect, .50 represents a moderate effect, and .70 represents a large effect (Meade, 2010). Additional statistics reported using the VisualDF 1.3 included signed (SID) and unsigned (UID) item difference values. SID and UID values indicate the degree of difference in score between groups on each item. The SID values represent the average difference in the probability of correct endorsement of an item between groups and allow group differences in item functioning to cancel out. In contrast, the UID values represent the hypothetical differences between groups' scores. The UID is the sum of the absolute values of the differences in the probability of correct endorsement between groups over the range of abilities. The UID does not allow for cancelation of DIF and provides an index of the total amount of differential functioning present between groups on an item over the range of ability. If the UID and SID values are identical, the difference in the probability of endorsing an item correctly between groups is consistent across the range of ability.

DIF by child gender—To analyze the GRTR-R for the presence of DIF as a function of gender, male students were coded as the referent group and females as the focal group. After corrections for multiple comparisons, no items had significant DIF by gender.

DIF by child age—To evaluate the presence of DIF as a function of age, children were divided into older and younger preschool children, using the sample mean age (54 months) as a cut point. The distributions of gender and race/ethnicity in the younger group (52% male; 62% White) were similar to the distributions of gender and race/ethnicity in the older group (49% male; 65% White); however, average scores on the GRTR-R were significantly higher for older children ($M = 17.55$, $SD = 4.96$) than they were for younger children ($M = 12.96$, $SD = 5.07$), $F(1, 1335) = 271.67$, $p < .01$. Discrimination and difficulty parameters for older and younger children are presented in Table 3.

After correcting for multiple comparisons, seven items had significant DIF. All of these items had DIF on the b parameter such that they were easier for older children than they were for younger children with the same overall ability level. All items with DIF, except for Item 12, also had significant DIF on the a parameter. Of the items with DIF on the a parameter, all items except Item 17 had DIF such that there was a higher degree of precision in discriminating older children's skill levels than there was in discriminating younger children's skill levels. Item 17 had a higher degree of precision in discriminating younger children's skills levels than for discriminating older children's skill levels. All effect sizes for DIF were small. In addition, SID and UID values were identical, indicating that DIF was consistent across all ability levels.

DIF by child race/ethnicity—To assess for the presence of DIF as a function of race/ethnicity on the GRTR-R, item parameters were compared between African American and White children, Hispanic and White children, and African American and Hispanic children. Sample sizes were smaller for these comparisons. Therefore, a 1PL model and a 2PL model were generated and compared for each analysis of DIF by race. The first analysis compared the responses of children who were African American to the responses of children who were White. In this analysis, a 2PL model provided a significantly better model fit than a 1PL

model, as indicated by a chi-square comparison of log likelihood values, $\chi^2(1, 25) = 217.88$, $p < .001$ (Embretson & Reise, 2000) and a comparison of AIC and BIC values (2PL: AIC = 29010.98, BIC = 29387.60; 1PL: AIC = 29180.86, BIC = 29438.55). Item parameters and effect sizes from this analysis are shown in Table 4. After adjusting for multiple comparisons, six items had significant DIF by race/ethnicity for the comparison between African American children and White children. Based on their difficulty parameters, Items 5, 11, 13, and 18 were easier for White children than for African American children with the same ability levels, and Items 3 and 19 were easier for African American children than for White children with the same ability levels. Items 18 and 19 had significant DIF on the discrimination parameter and provided a higher degree of discrimination for African American children's skill levels than they did for White children's skill levels. Item 11 also had significant DIF on the discrimination parameter but provided a higher degree of discrimination for White children than it did for African American children. All effect sizes of items with DIF were small in magnitude. The SID and UID were similar, indicating that DIF was consistent across all ability levels.

The second analysis compared the responses of children who were White to the responses of children who were Hispanic/Latino. In this analysis, a 2PL model provided a significantly better model fit than a 1PL model, as indicated by a chi-square comparison of log likelihood values, $\chi^2(1, 25) = 262.46$, $p < .001$ (Embretson & Reise, 2000) and a comparison of AIC and BIC values (2PL: AIC = 28193.98, BIC = 28568.48; 1PL: AIC = 28408.45, BIC = 28664.68) from each model. After adjusting for multiple comparisons, no items had significant DIF by race/ethnicity for children who were White versus Hispanic/Latino. The third analysis compared the responses of children who were Hispanic/ Latino (focal group) with the responses of children who were African American (reference group). To determine whether to use a 2PL model or a 1PL model for this analysis a chi-square comparison of log likelihood values was conducted, $\chi^2(1, 25) = 91.52$, $p < .001$ (Embretson & Reise, 2000) and AIC and BIC values were compared (2PL: AIC = 9689.24, BIC = 9980.01; 1PL: AIC = 9732.76, BIC = 9931.71). Although the BIC value for the 1PL was less than the BIC value for the 2PL model, the chi-square significance test and the AIC values indicate that the 2PL model provided a better fit than did the 1PL model. In this analysis, no items had significant DIF after adjusting for multiple comparisons.

DISCUSSION

The purpose of this study was twofold. First, the study was designed to determine the degree to which the GRTR-R provided precise measurement of preschool children's emergent literacy skill across the range of abilities. Overall, the results of the study demonstrated that the GRTR-R has sufficient precision of measurement for a majority of preschool children assessed. Second, the study was designed to determine the extent to which properties of the test were influenced by the characteristics of the children (i.e., age, gender, and race/ethnicity). Item functioning on the GRTR-R did not differ significantly across child gender. Only 10% of comparisons resulted in significant DIF, and the effect sizes of this significant DIF were, in all cases, small.

IRT analysis was used to evaluate the degree of measurement precision of the GRTR-R across children's emergent literacy skill levels. Results using a relatively large sample of children demonstrated that, collectively, the items provided adequate ability estimates for approximately 72% of children assessed. Precision of measurement of the GRTR-R was better for children with below average abilities than it was for children with above average abilities. Although it would be possible to increase the precision of the GRTR-R for children with higher skill levels through the addition of items to the screener that have higher levels of difficulty, the primary purpose of the GRTR-R is to identify children in need of additional assessment to diagnose specific disability or who are likely in need of intervention and resources to mitigate possible problems in developing literacy skills. Consequently, it is advantageous for the GRTR-R to have greater precision of measurement for children whose ability levels are below average, and the addition of items to the measure is not necessary to identify these children with adequate levels of precision.

DIF analysis was used to determine the extent to which item properties were influenced by characteristics of the children. The results of these analyses revealed that most child characteristics did not influence item properties. In fact, there were no differences based on child gender. A number of items appeared to be easier for older children than for younger children when children were equated on ability levels. It is possible that older children have had a higher degree of exposure to printed material than younger children as a result of having had more opportunities to interact with print—such as more time in preschool—and, therefore, were better acquainted with the functions and meaning of print. The absence of exposure to print may prevent alignment of measured print knowledge with true underlying abilities. Alternatively, it is possible that items were easier for older than younger children because older children in this study had a greater vocabulary, better developed executive functioning skills, or more experience taking tests than did younger children. There was also evidence that some items operated differently between children who were White and children who were African American.

The items with significant DIF as a function of child race/ethnicity on the GRTR-R had differences that did not appear to be systematic or of the type indicating that the measure was biased toward one group or another. Two types of DIF can be present on a measure. Uniform DIF is said to be present when only the difficulty parameters of items vary across groups and often indicates that items are biased. Nonuniform DIF is said to occur when either the discrimination parameters or both the discrimination and the difficulty parameters of items vary across groups and does not indicate that items are biased (Hanson, 1998; Mellenbergh, 1982). Of the six items with significant DIF as a function of child race/ethnicity, only Items 3, 5, and 13 had DIF on just the b parameter in the comparison between White and African American children. Both Items 5 and 13 were biased toward children who were White (i.e., at the same level of ability, children who were White were more likely to respond correctly to these items than were children who were African American), whereas Item 3 was biased toward children who were African American (i.e., at the same level of ability, children who were African American were more likely to respond correctly to these items than were children who were White). Of the seven items with significant DIF as a function of child age, all seven items were uniformly biased such that they were easier for older children. However, six of the seven items also had significant DIF on the a

parameter; thus, this DIF was not uniform. Five of the six items had DIF on the a parameter such that they provided better discrimination for older children than for younger children, and one item provided better discrimination for younger children than for older children. Overall, these results indicate that the GRTR-R is, for the most part, not biased toward different groups of children.

Although the GRTR-R yielded good measurement precision across a wide range of ability, few of the items had high discrimination parameters. This may have been due, in part, to the fact that all items on the GRTR-R have a multiple-choice format. Consequently, there is some element of chance-correct responding when children are choosing their answers on each item of the GRTR-R. The alternative to a multiple-choice item format would be to use items that involve free responses. There are two primary reasons the GRTR-R uses multiple choice items rather than free-response items. First, multiple-choice items (i.e., recognition) are easier for children to respond to than are equivalent items presented in a free-response format (i.e., recall). Because a primary purpose of the GRTR-R is to measure the skills of children with lower ability levels, it is necessary to have items that are not too difficult. One goal of multiple-choice items is to reduce memory demands. Because multiple-choice items reduce memory demands and make it easier for children to respond, children are more likely to give answers than they would be if the measure consisted of free-response items. Second, multiple-choice items are easier to administer and score than are free-response items. A design goal of the GRTR-R was that it could be used by individuals with minimal training in assessment generally and the GRTR-R specifically (Whitehurst, 2001). Some items had particularly low discrimination parameters. For example, the four items that assess children's blending and elision skills had discrimination parameters less than 0.50. Because these types of phonological awareness items tend to be difficult for children, it is possible that some children whose phonological awareness skills were just emerging responded to the wrong parts of the questions (e.g., the foils) or resorted to guessing. For instance, on Item 25 children are instructed to select the picture of “Scar without /s/.” One of the foils on this item is a picture of a snake, which also contains the /s/ sound, possibly encouraging children to select this response rather than the correct response (a picture of a car).

Although this study evaluated only the measurement precision of the GRTR-R, the validity of the measure has been examined in other studies. The original GRTR has been compared with a number of emergent literacy and reading skill measures and has been determined to have acceptable criterion and predictive validity (Molfese et al., 2004; Molfese et al., 2006; Phillips et al., 2009; Whitehurst, 2001). Wilson and Lonigan (2010) found that the overall correct classification of the GRTR-R ranged from 0.24 to 0.73 when categorizing children into groups based on child skill level as measured by a diagnostic assessment of early literacy skills. In addition, Wilson and Lonigan reported that, compared to the IGDIs, the GRTR-R was significantly more accurate in predicting later emergent literacy skills. Because validity has been established prior to this study, it was not necessary to examine the GRTR-R in relation with other measures of emergent literacy or reading skills in the current study.

The information obtained using the GRTR-R allows teachers to identify which children are most at risk for later academic difficulties and are most in need of additional resources.

Teachers can then provide the children identified with additional resources at an early age. Earlier identification and intervention helps to prevent a number of negative outcomes (Campbell & Ramey, 1995; Karoly et al., 2001). Enhancing children's emergent literacy skills through curricula and academic programs aimed at improving areas of weakness can also provide children with more positive academic experiences.

Limitations

Despite the relatively large number of children in the study, the sample was insufficiently large to use a 3PL model. Using a 3PL model could have provided information about the pseudo-guessing parameters and may have offered more accurate information about the discrimination parameters. However, the 2PL model, without accounting for a guessing parameter, provided evidence for an adequate level of measurement precision with the GRTR-R, supporting its use as a screener. In addition, in the DIF by age analyses, classifying children by years of school experience, rather than by age, may have provided stronger evidence in support of the postulation that some items appear to be easier for children with higher degrees of exposure to print than or for children with more testing experience; however, this information was not available for the samples. A final limitation involves the representativeness of the sample used. Although the sample was representative in terms of gender and race/ethnicity, children from the southeastern United States were overrepresented. It is unclear how this would affect item parameter estimates, given that one of the advantages of IRT analysis is sample invariance, but a more representative sample could have strengthened the ability to generalize the findings from this study.

Summary and Conclusions

The GRTR-R, which can be administered by individuals without formal training in the administration, scoring, and interpretation of assessments, is a brief and easily administered measure that can identify children who are at risk for having difficulties with reading achievement, thereby allowing children who are struggling with literacy skill development to be identified as in need of intervention. The results of this study indicate that the GRTR-R has adequate levels of precision as a screening measure that can be used to provide information about emergent literacy abilities for a diverse range of children. It has the highest degree of measurement precision for children with average to below average ability levels. Precision of measurement at this ability level enables the GRTR-R to function effectively as a screening measure to identify children with weaker emergent literacy skills who are in need of early intervention. In addition, most items on the screener function independently of child characteristics, including gender and race/ethnicity. Although there was evidence that the items on the screener functioned differently for older children than for younger children, this result was likely a result of degree of exposure to printed materials and, therefore, not indicative of bias. Therefore, the screener can be used to provide comparable results for most children assessed and to determine with accuracy which children are at risk for having difficulties with later academic achievement.

Acknowledgments

Funding The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Portions of this work were supported by grants from the National Institute of Child

Health and Human Development (HD052120) and the Institute of Education Science, U.S. Department of Education (R305B090021, R305B04074). The views expressed are those of the authors and have not been reviewed or approved by the granting agencies.

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57:289–300. doi:10.2307/1165312.

Bennett KJ, Brown KS, Boyle M, Racine Y, Offord D. Does low reading achievement at school entry cause conduct problems? Social Science & Medicine. 2003; 56:2443–2448. doi:10.1016/S0277-0536(02)00247-2. [PubMed: 12742607]

Campbell FA, Ramey CT. Cognitive and school outcomes for high-risk African-American students at middle adolescence: Positive effects of early intervention. American Educational Research Journal. 1995; 32:743–772. doi:10.3102/00028312032004743.

Chall, JS.; Jacobs, VA.; Baldwin, LE. The reading crisis: Why poor children fall behind. Harvard University Press; Cambridge, MA: 1990.

Embretson, SE.; Reise, SP. Item response theory for psychologists. Lawrence Erlbaum; Mahwah, NJ: 2000.

Hanson BA. Uniform DIF and DIF defined by differences in item response functions. Journal of Educational and Behavioral Statistics. 1998; 23:244–253. doi:10.3102/10769986023003244.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6:1–55. doi:10.1080/10705519909540118.

Hulin CL, Lissak RI, Drasgow F. Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement. 1982; 6:249–260. doi:10.1177/014662168200600301.

Karoly, LA.; Kilburn, MR.; Bigelow, JH.; Caulkins, JP.; Cannon, JS. Assessing costs and benefits of early childhood intervention programs: Overview and application to the Starting Early Starting Smart program. Casey Family Programs; Seattle, WA: 2001.

Lonigan CJ, Allan NP, Lerner MD. Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. Psychology in the Schools. 2011; 48:488–501. doi:10.1002/pits.20569. [PubMed: 22180666]

Lonigan, CJ.; Schatschneider, C.; Westberg, L. Impact of code-focused interventions on young children's early literacy skills. Report of the National Early Literacy Panel. National Institute for Literacy; Washington, DC: 2008.

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

Lonigan, CJ.; Wilson, SB. Report on the Revised Get Ready to Read! screening tool: Psychometrics and normative information (Tech. Rep.). National Center for Learning Disabilities; New York, NY: 2008.

Lord FM. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement. 1968; 28:989–1020. doi:10.1177/001316446802800401.

Matson SC, Haglund KA. Relationship between scholastic and health behaviors and reading level in adolescent females. Clinical Pediatrics. 2000; 39:275–280. doi:10.1177/000992280003900503. [PubMed: 10826074]

Meade AW. A taxonomy of differential functioning effect size indices. Journal of Applied Psychology. 2010; 95:728–743. doi:10.1037/a0018966. [PubMed: 20604592]

Mellenbergh GJ. Contingency table models for assessing item bias. Journal of Educational Statistics. 1982; 7:105–118. doi:10.2307/1164960.

Molfese VJ, Modglin AA, Beswick JL, Neamon JD, Berg SA, Berg CJ, Molnar A. Letter knowledge, phonological processing, and print knowledge: Skill development in nonreading preschool children. Journal of Learning Disabilities. 2006; 39:296–305. doi:10.1177/00222194060390040401. [PubMed: 16895155]

Molfese VJ, Molfese DL, Modglin AT, Walker J, Neamon J. Screening early reading skills in preschool children: Get Ready to Read. Journal of Psychoeducational Assessment. 2004; 22:136–150. doi:10.1177/073428290402200204.

Muthén, LK.; Muthén, BO. Mplus: Statistical analysis with latent variables: User's guide. Muthén & Muthén; Los Angeles, CA: 2007.

National Center for Education Statistics. Reading 2009: National Assessment of Educational Progress at grades 4 and 8. U.S. Department of Education; Washington, DC: 2009.

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

Phillips BM, Lonigan CJ, Wyatt MA. Predictive validity of the *Get Ready to Read!* screener: Concurrent and longterm relations with reading-related skills. Journal of Learning Disabilities. 2009; 42:133–147. doi:10.1177/0022219408326209. [PubMed: 19074622]

Ponterotto JG, Ruckdeschel DE. An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. Perceptual and Motor Skills. 2007; 105:997–1014. doi:10.2466/PMS.105.3.997-1014. [PubMed: 18229554]

Reckase MD. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics. 1979; 4:207–230. doi:10.2307/1164671.

Sinar EF, Zickar MJ. Evaluating the robustness of graded response model and classical test theory parameter estimates to deviant items. Applied Psychological Measurement. 2002; 26:181–191. doi:10.1177/01421602026002005.

Stout WF. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika. 1990; 55:293–325. doi:10.1007/BF02295289.

Thissen, D. IRTLRDIF v.2.0.b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. University of North Carolina at Chapel Hill, L. L. Thurstone Psychometric Laboratory; Chapel Hill: 2001. Retrieved from http://www.unc.edu/~dthissen/dl.html

Whitehurst, GJ. The NCLD Get Ready to Read! screening tool technical report. State University of New York at Stony Brook, National Center for Learning Disabilities; Stony Brook: 2001.

Whitehurst GJ, Lonigan CJ. Child development and emergent literacy. Child Development. 1998; 69:848–872. doi:10.1111/j.1467-8624.1998.tb06247.x. [PubMed: 9680688]

Whitehurst, GJ.; Lonigan, CJ. Get Ready to Read! screening tool. National Center for Learning Disabilities; New York, NY: 2001.

Wilson SB, Lonigan CJ. Identifying preschool children at risk of later reading difficulties: Evaluation of two emergent literacy screening tools. Journal of Learning Disabilities. 2010; 43:62–67. doi: 10.1177/0022219409345007. [PubMed: 19822699]

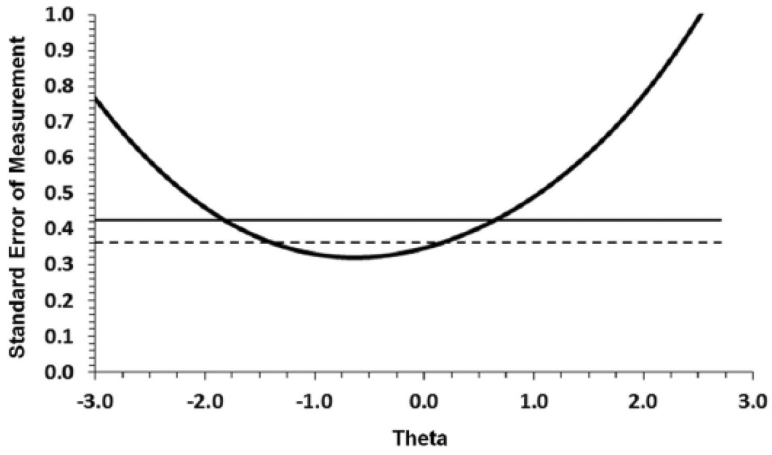

Figure 1.

Plot of the standard error of measurement of the *Revised Get Ready to Read!* (GRTR-R) over the range of theta.

Note. The solid horizontal line at y = .42 is equivalent to a reliability level of .80 in classical test theory (CTT), and the dashed horizontal line at y = .37 is equivalent to a reliability level of .85 in CTT.

Table 1

Model Fit Statistics of Unidimensional and Two-Factor Models of the *Revised Get Ready to Read!*

Model	χ^2	*df*	CFI	TLI	RMSEA
Unidimensional	962.32	275	.940	.934	.043
Theoretical 2-factor	859.71	274	.949	.944	.040
Empirical 2-factor	859.63	274	.953	.949	.038

Note. *N* = 1,351. CFI = comparative fit index; RMSEA = root mean square error of approximation; TLI = Tucker–Lewis index.

Table 2

Items, Item Response Theory (IRT) Parameters, and Classical Test Theory (CTT) Statistics of the *Revised Get Ready to Read!*

Item	Content	Content Domain	Discrimination Parameter	Difficulty Parameter	Item-Total Correlation[a]	Percentage Correct[a]
1	Find picture that shows back of book.	PK	0.55	−1.20	.39**	72
2	Find picture with letters.	PK	0.74	−0.63	.43**	64
3	Find picture with letters.	PK	0.73	−0.64	.44**	64
4	Find picture with words.	PK	0.82	−1.18	.38**	77
5	Find picture that shows the name of the cereal.	PK	0.72	−0.38	.38**	58
6	Find the letter R.	PK	1.16	−0.88	.54**	74
7	Find the letter G.	PK	1.02	−0.93	.45**	74
8	Find the letter that makes the /s/ sound.	PK	0.94	−0.85	.50**	71
9	Find the letter that makes the /t/ sound.	PK	0.97	−0.28	.56**	57
10	Find the letter that makes the /b/ sound.	PK	0.90	−0.69	.47**	67
11	Find the F that is written the best.	PK	0.60	−1.49	.38**	78
12	Find the name that is written the best.	PK	0.71	−0.30	.50**	56
13	Find the longest story.	PK	0.94	−1.33	.46**	82
14	Find the picture of the word that starts with the /b/ sound.	PA	0.90	−0.14	.51**	53
15	Find the picture of the word that starts with the /b/ sound.	PA	0.87	−0.52	.47**	62
16	Find the picture of the word that rhymes with “ball.”	PA	0.72	0.33	.49**	42
17	Find the picture of the word that is “sea - shell.”	PA	0.49	−1.66	.32**	77
18	Find the picture of the word that is “pen - guin.”	PA	0.49	−1.40	.37**	73
19	Find the picture of the word that is “m - oon.”	PA	0.43	−0.50	.37**	58
20	Find the picture of the word that rhymes with “arm.”	PA	0.76	0.77	.37**	32
21	Find the picture of the word that rhymes with “hat.”	PA	0.64	0.29	.36**	43
22	Find the picture that has numbers in it.	PK	1.09	−0.37	.46**	59
23	Find the one that shows how to write two words.	PK	0.40	0.69	.28**	39
24	Find the word that is written the best.	PK	0.62	0.06	.35**	48

Item	Content	Content Domain	Discrimination Parameter	Difficulty Parameter	Item-Total Correlation[a]	Percentage Correct[a]
25	Find the picture that is “scar” without “sss.”	PA	0.42	0.72	.05*	34

Note. N = 1,351. PA = phonological awareness; PK = print knowledge.

[a]CTT analogs to discrimination and difficulty parameters.

*$p < .05$.

**$p < .01$.

Table 3

Item Parameters Based on Age, Significance Levels, and Effect Sizes of Differential Item Functioning.

	Older[a]		Younger[b]					
Item	a	b	a	B	*p*	ES	SID	UID
1	0.44	−2.05	0.53	−0.80	.05	.13	−.15	.15
2	0.72	−1.01	0.68	−0.32	.04	.19	−.16	.16
3	0.80	−1.09	0.54	−0.21	.71	.18	−.22	.22
4	0.85	−1.67	0.62	−0.95	.71	.15	−.17	.17
5	0.75	−0.99	0.51	0.21	.04	.18	−.27	.27
6	1.02	−1.40	1.13	−0.54	.05	.22	−.19	.19
7	1.26	−1.15	0.84	−0.72	.01*	.20	−.15	.15
8	1.01	−1.23	0.78	−0.56	.64	.20	−.18	.18
9	1.05	−0.65	0.80	0.09	.11	.24	−.22	.22
10	1.02	−0.97	0.77	−0.44	.01*	.21	−.16	.16
11	0.66	−1.90	0.44	−1.36	.52	.11	−.14	.14
12	0.58	−1.19	0.56	0.44	.000*	.17	−.32	.32
13	1.07	−1.97	0.63	−1.08	.000*	.13	−.21	.21
14	0.93	−0.66	0.65	0.44	.67	.22	−.28	.28
15	1.20	−0.83	0.55	−0.18	.005*	.21	−.21	.22
16	0.99	−0.17	0.38	1.37	.003*	.21	−.24	.25
17	0.24	−4.02	0.58	−1.08	.000*	.11	−.12	.12
18	0.37	−2.23	0.51	−1.03	.03	.12	−.10	.10
19	0.32	−1.33	0.43	−0.07	.03	.13	−.15	.15
20	0.97	0.31	0.49	1.63	.02	.21	−.18	.18
21	0.81	−0.12	0.43	0.95	.12	.20	−.18	.18
22	1.14	−0.95	0.79	0.18	.03	.23	−.33	.33
23	0.45	0.21	0.29	1.57	.78	.13	−.14	.14
24	0.71	−0.45	0.39	0.87	.25	.17	−.24	.24
25	0.51	0.36	0.32	1.26	.05	.14	−.09	.09

Note. ES = effect size; SID = signed item difference; UID = unsigned item difference.

[a] $n = 639$.

[b] $n = 698$.

* Significant after correction for multiple comparisons.

Table 4

Item Parameter Estimates and Effect Sizes for Differential Item Functioning Analyses Comparing African American Responses With White Responses.

	African American[a]		White[b]					
Item	a	b	a	b	*p*	ES	SID	UID
1	0.57	−1.12	0.52	−1.33	.74	.15	−.02	.02
2	0.68	−0.47	0.67	−0.69	.86	.20	−.05	.05
3	0.82	−0.72	0.71	−0.62	.01*	.21	.04	.04
4	1.20	−0.77	0.77	−1.40	.17	.21	−.09	.09
5	0.75	0.20	0.72	−0.57	.005*	.23	−.19	.19
6	1.25	−0.66	1.05	−1.00	.86	.25	−.08	.08
7	0.92	−0.76	0.94	−0.99	.82	.22	−.06	.06
8	0.79	−0.55	0.92	−0.93	.41	.22	−.11	.11
9	0.96	−0.03	0.91	−0.37	.95	.27	−.10	.10
10	0.94	−0.55	0.93	−0.79	.78	.24	−.06	.06
11	0.39	−1.12	0.69	−1.53	.001*	.13	−.15	.15
12	0.70	0.14	0.72	−0.42	.17	.22	−.13	.13
13	0.96	−0.69	0.92	−1.66	.000*	.19	−.19	.19
14	0.90	−0.19	0.87	−0.17	.04	.26	.006	.007
15	0.78	−0.58	0.87	−0.55	.06	.23	.000	.02
16	0.56	0.57	0.76	0.25	.35	.21	−.05	.06
17	0.77	−0.97	0.40	−2.10	.14	.15	−.06	.08
18	1.05	−0.83	0.43	−1.51	.003*	.19	−.004	.10
19	0.78	−0.46	0.40	−0.39	.002*	.19	.05	.09
20	0.61	1.23	0.83	0.63	.23	.21	−.08	.08
21	0.57	0.74	0.70	0.08	.14	.21	−.13	.13
22	1.21	−0.29	1.14	−0.38	.12	.29	−.03	.03
23	0.49	0.79	0.40	0.64	.74	.15	−.04	.05
24	0.51	0.31	0.66	0.03	.35	.20	−.05	.05
25	0.62	0.55	0.43	0.60	.39	.18	−.02	.05

Note. ES = effect size; SID = signed item difference; UID = unsigned item difference.

[a] $n = 184$.

[b] $n = 865$.

*Significant after correction for multiple comparisons.