Published in final edited form as:

Read Writ. 2015 January ; 28(1): 31–56. doi:10.1007/s11145-014-9518-z.

Improving the Reliability of Student Scores from Speeded Assessments: An Illustration of Conditional Item Response Theory Using a Computer-Administered Measure of Vocabulary

Yaacov Petscher, Alison M. Mitchell, and Barbara R. Foorman

Florida Center for Reading Research, Florida State University, 2010 Levy Avenue, Suite 100, Tallahassee, FL 32310, USA

Yaacov Petscher: ypetscher@fcrr.org

Abstract

A growing body of literature suggests that response latency, the amount of time it takes an individual to respond to an item, may be an important factor to consider when using assessment data to estimate the ability of an individual. Considering that tests of passage and list fluency are being adapted to a computer administration format, it is possible that accounting for individual differences in response times may be an increasingly feasible option to strengthen the precision of individual scores. The present research evaluated the differential reliability of scores when using classical test theory and item response theory as compared to a conditional item response model which includes response time as an item parameter. Results indicated that the precision of student ability scores increased by an average of 5 % when using the conditional item response model, with greater improvements for those who were average or high ability. Implications for measurement models of speeded assessments are discussed.

Keywords

Fluency; Conditional item response theory; Item response theory; Curriculum-based measurement

Central to the goal of assessment is the attainment of a reliable and valid score of an individual's measured skill. High levels of reliability and validity are necessary to ensure that a measured behavior can be confidently used for early identification of risk, relative student comparisons, mastery of instructional content, and/or growth in performance over time. Leveraging as much student response data as possible is likely to improve reliability and positively impact these goals. Assessments that have the capacity to capture speed of performance, as well as accuracy, have become increasingly more common due to the burgeoning use of technology in educational settings (Gray, Thomas, & Lewis, 2010). Incorporating response time into commonly used measurement models is one potential avenue that may facilitate greater sophistication in estimating the reliability of an individual's score; however, only direct comparison of measurement models that do or do not incorporate response time can speak to the extent to which this additional consideration may influence reliability in estimations of performance. The current paper compares

reliability estimates using multiple measurement models both with and without the inclusion of time information using a measure of computerized measure of vocabulary knowledge, with the goal of exploring whether adding information on student response time can provide added benefit to these estimations.

FLUENCY

Although a firm consensus on the construct of fluency has not been agreed upon (Wolf & Katzir-Cohen, 2001), it is frequently defined as a composite of speed and accuracy in performance (Chard, Vaughn, & Tyler, 2002). While fluency defined in this manner can relate to many areas of performance, it is frequently associated with time-limited assessments measuring proficiency in academic skills, such as reading, mathematics or writing. In the domain of reading, several types of fluency assessments exist, but can generally be distilled to those brief assessments that assess student skills using word or object lists (e.g., letter name fluency, word identification fluency) and those that measure reading using connected text (e.g., oral reading fluency, silent reading fluency). Reading fluency in general has been underscored as a significant factor in reading performance over the past several decades (Adams, 1990; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Kim, Wagner, & Foster, 2011) due to its relevance as a theoretical construct (LaBerge & Samuels, 1974; Wolf & Katzir-Cohen, 2001) its strong predictive relationship with reading comprehension in the primary grades (Kim et al., 2011; Petscher, Cummings, Biancarosa, & Fien, 2013), and its ability to measure skills in a relatively quick manner (Good, Simmons, & Kame'enui, 2001). Fluent reading and its component skills require the simultaneous processing of multiple word recognition and language comprehension skills (Scarborough, 2001). Thus, it is argued that as reading becomes increasingly more fluent and automatic, a reader's attention may become more directed to metacognitive skills involved in comprehending the meaning of text (LaBerge & Samuels, 1974).

Curriculum-based measures (CBM) are time-limited assessments that are designed to be used multiple times a year to measure growth in performance in academic skills (Deno, 2003). These fluency assessments are commonly used to identify students who are at risk for reading failure (Cummings, Atkins, Allison, & Cole, 2008) and to estimate change in developmental trajectories for differential subgroups of students over time (Al Otaiba et al., 2009; Christ, Silberglitt, Yeo, & Cormier, 2010; Logan & Petscher, 2010). However, recent measurement issues have been identified pertaining to curriculum-based measures of reading (CBM-R), such as form effects (Ardoyn & Christ, 2009; Christ & Ardoyn, 2009; Francis et al., 2008; Petscher & Kim, 2011) and error of measurement (Christ & Silberglitt, 2007). Related to the latter concern regarding error of measurement, one attractive property of CBM assessments is the purported high reliability of scores they provide (Petscher et al., 2013), yet such evidence should be evaluated with a note of caution. Similar to many assessments in educational research, CBM assessments are rooted in classical test theory. This approach assumes that the standard error of measurement is constant across the range of scores in a population. The implication of this assumption is that reports of high reliability for CBM assessments assume equal reliability for all students, which may not be a tenable assumption for many individuals within the population. Poncy, Skinner, and Axtell (2005) applied generalizability theory to CBM probes and found that reliability-like

coefficients ranged from 0.81 to 0.99 in their sample. This indicated that for some individuals the fluency scores were highly reliable, and nearly free from any error, while others' scores were observed to have relatively lower reliability. This variation in coefficients suggests that the reliability of one's fluency score may be related to where that score falls in the population distribution. Other studies of standard error pertaining to fluency and growth in fluency (Christ & Silbergitt, 2007; Mercer et al., 2012) have converged on a similar result, namely, that the standard error of measurement is not static in a population.

Such findings are important in considering how fluency scores from CBM tools, or resulting total scores from any educational assessment, are used for decisions regarding instruction or intervention for students. Differential reliability for students' scores has implications for the validity of the scores. A complexity that arises pertaining to differential reliability is that when scores are estimated from a classical test theory framework, it is not possible to evaluate whether each individual's score is minimally reliable. If one were interested in evaluating how to overcome the limitation of constant error for fluency tasks, it may be necessary to employ an alternative measurement framework to classical test theory and consider an alternative metric for a fluency score. This new metric, which we present later in this paper, is rooted in the idea that fluency can be measured at the item level as a function of speed and accuracy, but without the time-limited constraint. When a performance score is estimated as a function of item-level speed and accuracy, it is possible to use an alternative measurement models to provide a fluency score.

Measurement models that can account for both accuracy and speed at the item-level may be of special relevance to educational assessments of reading and language. Traditional measures of reading fluency present one clear way to capture both of these performance factors at a global level, as a student's accuracy level on a number of items is determined within a specified timeframe. Notably, with the advent of computerized assessments, it may be possible to accurately capture item response time information to improve estimations of student performance without being contingent on time as a limiting factor. In other words, while speed of response on an item may still be captured, students' are not limited to a certain timeframe within which to respond. It is important to state that this approach is not intended to replace the benefits of time-limited assessment, such as speed and convenience, but to add to a growing body of research that is modeling speed and accuracy from an alternative perspective (van der Linden & van Krimpen-Stoop, 2003). By leveraging response time information in computerized assessments we may gain greater flexibility in the types of assessments used to evaluate fluent performance.

MEASUREMENT MODEL CONSIDERATIONS

Traditional approaches to the measurement of reliability include internal consistency, test-retest, parallel-form, and split-half methods. These multiple aspects of reliability, and the coefficients estimated to make inferences about them, are grounded in classical test theory, which purports that an observed score (X) is a function of a true score (T) and random error (e). The ratio of the true-score variance to observed-score variance is reliability, which separates out the random-error variance from the variance in scores attributable to the ability

of the individuals. Several seminal resources exist that outline the limitations of classical test theory (e.g., Hambleton, Swaminathan, & Rogers, 1991). A notable restriction to this theoretical approach is the assumption that the standard error of measurement for the test does not vary across a population. That is, whether an examinee obtains a low test score or a high test score, the standard error associated with the examinee's score is the same regardless of his or her actual performance (Embretson & Reise, 2000). Further, a particular limitation of classical test theory is that the total test score cannot be directly compared to the difficulty (i.e., p value) of the items in the assessment. A p value in classical test theory denotes the proportion of individuals who have correctly answered the item and ranges from 0 to 100 %; high values indicate easy items, while low values indicate a relatively difficult item. Because item p values and individual total test scores are often on different metrics, explicitly linking the two is not possible. While items may range in difficulty from 0 to 100 % (or 0.00–1.00), individual scores may appear in a raw metric (e.g., 0–26 for a letter naming task), an age-standardized metric (e.g., mean of 100 and standard deviation of 15), or a developmental-standardized metric to capture growth over time (e.g., mean of 500, standard deviation of 100). Relating an individual's standard score of 110 on a letter naming task to a particular letter's difficulty (e.g., "Z," p value for the sample = 40 %) can be quite challenging given this difference in item and person metrics.

Item response theory is a measurement framework that overcomes these described limitations. Its estimation framework is such that it places the item difficulty and the person's ability on the same scale (i.e., a z -score) via a function known as the item characteristic curve (ICC). Further, the assumption of equal measurement error for all individuals does not exist, thus, the reliability of scores for all students is not considered to be the same and is allowed to vary (Embretson & Reise, 2000). An ICC is a function of the item difficulty (i.e., the point on the curve where the probability of correctly answering the question is .50; b parameter) and the discrimination (i.e., the steepness of the slope of the ICC; a parameter).¹ Item difficulties range from approximately -3 to 3 and can be interpreted similarly to z -scores. Negative values indicate that items are easier while positive values denote harder items. While high p values in classical test theory denote easy items and low values indicate difficult items, in item response theory this pattern is reversed. The metric of item discriminations in IRT is $-\infty$ to $+\infty$, with optimal values ranging from 0.8 to 2.5 (de Ayala, 2009). This parameter is related to the item-to-total correlation; large values for item discriminations and item-to-total correlations suggest a strong relation between the item and the measured construct.

Person ability scores (known as θ scores) range from approximately -3 to 3 , and can be interpreted as z -scores, just as the item difficulty. By placing the item difficulties and individual abilities on the same scale, an explicit link is made between the two. For example, if a person who has $\theta = 0$ (i.e., average ability) encounters an item with a difficulty of 0 (i.e., average difficulty) there is a 50 % chance of correctly answering that item. When controlling for the difficulty of the item at 0, those with $\theta > 0$, have more than a 50 % chance of answering the question correctly because their ability exceeds the difficulty of the item, and

¹This assumes that the pseudo-guessing parameter has a value of 0.

those with $\theta < 0$ have less than a 50 % chance of answering the item correctly because their ability is lower than the difficulty of the item. Because the focus of IRT is on the item responses themselves, rather than the total test score, it is possible to estimate the standard error of each individual examinee's ability score rather than assume that it is the same for a sample or population of examinees of varying abilities, as is done in classical test theory.

Although item statistics in a traditional item response theory framework are comprised of the item difficulty, discrimination, and a guessing parameter, it is reasonable to consider that other item features may influence the likelihood of an individual correctly answering an item. Wainer, Bradlow, & Wang (2007) reported how testlet effects—when sets of items are administered as a bundle, such as in reading comprehension—influence the probability of a correct response. They proposed an item response model that accounted not only for the item difficulty, discrimination, and the guessing parameter but also the item dependency that may exist when items share a common stimulus. In a similar manner, it is possible to view item response latency, or the amount of time it takes a student to respond to an item, as a parameter that could influence the probability of a correct response. Response latency has long been recognized as an important source of data for evaluating individual differences in cognitive science (Sternberg, 1969) and is an important construct within neuroscience (Edgar et al., 2013) and psychopharmacology (Metrik et al. 2012) as well as psycholinguistics (Goodglass, Theurkauf, & Wingfield, 1984).

In a seminal study by Perfetti & Hogaboam (1975), students who were categorized as skilled or less skilled comprehenders differed in vocalization latencies for words they were shown, particularly when the stimuli were low-frequency words or pseudowords. Though skilled and non-skilled readers demonstrated similar response times when presented with high-frequency words, skilled students maintained faster response times to low-frequency and pseudowords compared to less skilled reader. These results led the authors to suggest that word meanings may be a less salient factor in response latencies for skilled readers than for their less skilled counterparts.

Further, they demonstrated the significance of response time, beyond accuracy alone, in providing a comprehensive measure of performance.

With the widespread availability and use of technology in educational contexts (Blackwell, Lauricella, Wartella, Robb, & Schomburg, 2013; Miranda & Russell, 2011; Pressey, 2013), researchers and practitioners are increasingly turning to computerized testing as an approach to administer academic achievement assessments. Along with recording item-level accuracy data, computers possess the valuable capability to accurately record item-level response times. By recording both time and accuracy, one is able to overcome the dichotomous distinction made by Cattell (1948) who noted that tests function to measure either power (i.e., accuracy given unlimited time) or performance (i.e., ability given limited time).

Several theoretical approaches have been proposed to account for time in response models (Schnipke & Scrams, 2002). One approach incorporates response time into a traditional item response theory model, where an assumed interaction exists between the parameters of response time and accuracy (van der Linden & van Krimpen-Stoop, 2003); that is, it is

assumed that more difficult items are related to longer response times. A second theoretical approach in the item response framework, advanced by Scheiblechner (1985), stated that the distribution of response time is relationally independent of the item accuracy. The limitation of this approach is that it ignores the ability of the individual, thus, the joint relation between speed and accuracy is unknown. A third method, proposed by van der Linden (2007) is called the conditional item response theory (CIRT) model. A hallmark of this model is that the variation in responses is due to two levels, a person/item level and a population/ domain level (Fig. 1). At level 1 (the individual level) there are two estimated vectors, one for the individual's item responses (i.e., U_{ij}) and one for the individual's response time (i.e., T_{ij}). The item response vector is defined as:

$$U_{ij} \sim f(u_{ij}; \theta_j, a_i, b_i)$$

where u_{ij} is the item response on item i for person j , θ_j is the latent ability of person, a_i is the item discrimination, and b_i is the item difficulty. This function is solved by a traditional two-parameter probability function. The response-time vector includes new parameters into the item response theory framework with

$$T_{ij} \sim f(t_{ij}; \tau_j, \alpha_i, \beta_i),$$

where t_{ij} is the response time on item i for person j , τ_j is the average speed of the individual, α_i is the discrimination parameter which provides information about the variation in response times across items, and β_i is known as the time intensity for the item. While the discrimination parameter is better known given its analog in the item response vector (i.e., α_i), response time, speed, and time intensity are additional parameters which require further explication. Though seemingly similar, each maintains its own operationalization (van der Linden, 2011). Of the three parameters, only response time is a measured variable. Time intensity may be defined as the amount of labor required by the item, or as the effect of an item on the mean log time. This component of the model and the average speed of the individual are considered to be latent constructs. van der Linden (2011) shows that the relation among these three can be expressed as a ratio, where response time is the ratio of time intensity and average speed. The relation between time intensity and average speed is similar to the association between item difficulty and individual ability. That is, in the same way that an individual has a higher probability of a correct item response when their ability exceeds the difficulty of the item, so it is also the case that it is more beneficial for the individual's speed to exceed the intensity of the item (i.e., $\tau_j > \beta_i$) than the reverse ($\tau_j < \beta_i$).

The response-time portion of the individual level model is estimated in a similar manner as the item response model using a lognormal distribution (Fox, Klein Entink & van der Linden, 2007; Klein Entink, Khun, Hornke, & Fox, 2009). Once the item response and response latency portions of the individual level model are established, a joint model is estimated as a function of the interaction between the products of item response and response time. The population portion of Fig. 1 (level 2) represents the estimation of the person and item components, as well as the covariances among them, as a function of the level-1 components. An attractive feature of the CIRT model is that it bridges the previously

reported theoretical approaches posited by van der Linden and van Krimpen-Stoop (2003) and Scheiblechner (1985). That is, though van der Linden and van Krimpen-Stoop (2003) posited that response latency can be directly modeled in an item response model, and Scheiblechner (1985) viewed the distribution of response time as independent from accuracy, the CIRT approach incorporates the theoretical notion of independence at the individual level of estimation, but then includes the joint relation between speed and accuracy at the population level.

Several benefits of this model compared to traditional item response models are worth noting. First, because the CIRT approach is rooted in item response theory, the resulting item parameters for difficulty and discrimination in the accuracy portion of the model can be compared to those estimated in traditional one-parameter and two-parameter item response models. Second, recent findings have suggested that accounting for the response time yields more precise estimates of ability due to the joint estimation of accuracy and response latency at level 2 of the CIRT model (Ferrando & Lorenzo-Seva, 2007; van der Linden, 2007). Third, the joint modeling of accuracy and response time has the potential for greater understanding of cognitive processes relative to the task. For example, it may be possible to evaluate the extent to which ability is related to speed, whether difficult items are the most time-intensive, and whether other aspects of the test might moderate the relation between an individual's ability and their speed.

Test format is a notable factor in considering which types of assessments may best lend themselves to inclusion of response time in their models. Connected text measures may not be well-suited to current item response models (with the exception of maze type tasks) as the data would likely violate the assumption of local item independence due to the repetition of many common words in a given text. Further, CIRT models would also be difficult to fit to the data from connected text measures of oral reading fluency, as obtaining response time per word would be challenging. List measures or maze type tasks that are limited to one sentence, on the other hand, may be well-suited to both item response accuracy modeling, as well as CIRT models, because the individual items can be programmed in a computer environment where response accuracy and response time could be captured accurately.

Assessments that use list or sentence-based delivery formats may expand the types of tasks that can be used to understand the degree to which speed facilitates performance on these measures for students at different performance levels. Component skills of reading, such as vocabulary knowledge, are established in the literature as important predictors of comprehension (Cunningham & Stanovich, 1997; Kamil, 2004; National Institute of Child Health and Human Development, 2000), yet outside of the Test of English as a Foreign Language (Educational Testing Service, 2007), relatively few measures of vocabulary which includes accuracy and speed currently exist. Vocabulary knowledge has been measured in multiple formats, including via computerized sentence-level maze tasks (Foorman et al., 2012). This format presents an opportunity to consider how response time may contribute to measurement precision using a novel domain.

PRESENT STUDY

Given the importance of the construct of fluency, wide utility of measures in the realm of curriculum-based measurement, and the broader context of educational assessment, a CIRT model may illuminate relations between response accuracy and time that were previously restricted to basic correlational evidence for other types of assessment measures. As reading skill assessments increasingly move toward modes of computer administration, it is of interest to compare estimates of reliability using classical test theory as well as item response theory analyses, which may or may not account for response time. Because a chief concern in educational assessment is minimizing the amount of time a student is assessed, a CIRT model for time-limited tests may improve the reliability of students' scores for accuracy tests that record response times.

A computerized maze vocabulary task was chosen because it constitutes an alternative to traditional fluency tasks that measure accuracy within a constricted timeframe and because of the empirical link between language skills and comprehension (Scarborough, 2001). The following research questions were explored: (1) What is the relation between classical test theory item difficulties, those estimated by traditional item response models, and CIRT models? (2) What is the relation between individual response accuracy and speed? (3) Does an IRT model provide differential precision (i.e., reliability) for estimated ability scores compared to classical test theory? (4) Does the CIRT model yield more precise estimates of student ability in a computerized test of vocabulary knowledge compared to either the IRT or classical test theory models?

METHOD

Participants

A total of 212 third grade students (110 boys, 100 girls, 2 not recorded) in the southeastern United States participated in the present study. The children came from predominantly low socioeconomic backgrounds, as 70 % of the students were eligible for free or reduced price lunch. The sample was primarily White (50 %), followed by Hispanic (28 %), Black (12 %), Asian (4 %), Multiracial (4 %) and Other (2 %). Four percent of students were identified as English language learners, and 13 % had an Individualized Education Program.

Measure

Vocabulary knowledge task (VKT; Foorman, Petscher, & Bishop, 2012)—In this task, students completed 30 sentences with one of three morphologically related words that best completed the sentence. Items were manipulated to test knowledge of prefixes and derivational suffixes (e.g., The student [attained*, retained, detained] a high grade in the class through hard work). Because this is a sentence-level task, there are concomitant word recognition, semantic, and syntactic demands in addition to the demands of the phonological and orthographic shifts. Target words in the task were selected on the basis of their printed word frequency (Zeno, Ivens, Millard, & Duvvuri, 1995) and sentences were assigned to grade level using the Flesch–Kincaid grade-level readability formula, along with our judgment about what topics would be familiar to students at different grades. This task was

constructed to maintain multiple forms with both common and unique form items (Kolen & Brennan, 2004). The task was group-administered to students in a computer lab. Because it was computer-administered, both the item response accuracy and response times were captured. The process of item administration was such that students were provided a fixed-order set of items and as each item was presented, students read the sentence, chose the option they believed was correct and submitted the response via a “submit” button on the screen. Response time was calculated (in seconds) as the amount of time which lapsed from the computer delivery of the item to when the student clicked the submission button.

Dimensionality for the scores in a larger sample were previously evaluated via factor analysis across grades 3–10 (Foorman et al., 2012) and demonstrated that a one-factor model provided the most parsimonious structure to the data. The present study used data from one form within third grade, allowing for exploration of both dimensionality of item responses within the selected form as well as an item response theory analogue to classical test reliability known as marginal reliability (Sireci, Thissen, & Wainer, 1991).

Data analysis

Prior to the estimation of the item parameters, the extent to which the items’ accuracy responses from the form yielded a unidimensional construct was evaluated. Due to the dichotomous scoring of the item responses, a combination of parametric and nonparametric tests of exploratory and confirmatory factor analyses were conducted in order to comprehensively evaluate the underlying structure of responses (Tate, 2003). Parametric exploratory and confirmatory factor analyses were run using Mplus (Muthen & Muthen, 1998–2012), where the ratio of eigenvalues, comparative fit index (CFI, Bentler, 1990), Tucker-Lewis index (TLI; Bentler & Bonnett, 1980), and the root mean square error of approximation (RMSEA, Browne & Cudeck, 1992) were used to evaluate model fit in the exploratory analysis. All but the ratio of eigenvalues were also used to evaluate the parametric confirmatory factor analysis model. CFI and TLI values greater than or equal to 0.95 are considered to be minimally sufficient criteria for acceptable model fit, and RMSEA estimates <0.05 are desirable. Non-parametric exploratory analysis was run using DIMTEST (Stout, 1987), where a non-significant *T* value indicates that the factor structure is essentially unidimensional. DETECT software (Zhang & Stout, 1999) estimated the nonparametric confirmatory model where a DETECT index less than 0.20 provides evidence of an essentially unidimensional model (Jang & Roussos, 2007).

Following the tests of dimensionality, classical test theory statistics, including item *p* values, item-to-total correlations, and internal consistency of item responses via Cronbach’s alpha were estimated using SAS 9.3 software (SAS Institute Inc., 2011). Item response theory analyses (IRT) using Mplus (Muthen & Muthen, 1998–2012) with maximum likelihood estimation included the fitting of Rasch and two-parameter logistic models in order to estimate item parameters and person ability scores. The conditional item response theory analyses were fit using the CIRT package (Fox et al., 2007) in R (R Core Team, 2013). A total of four CIRT models were estimated to identify which best captured the data: (1) a one-parameter response, one-parameter response-time model (Model 1), (2) a two-parameter response, one parameter response-time model (Model 2), (3) a one-parameter response, two-

parameter response-time model (Model 3), or (4) a two-parameter response, two-parameter response-time model (Model 4). CIRT models were evaluated by using the Deviance Information Criteria (DIC), which estimates the data-model deviation penalized by the model parameters and is computed by the sum of the posterior mean of the deviation (i.e., $\bar{D}$) and the effective number of model parameters (i.e., p_D). Similar to other information criteria, such as the Bayesian Information Criteria (BIC), a DIC is evaluated based on its relative comparison to other DICs. As such, while a DIC may be large in magnitude, it is intended to be compared to others, and the model with the smallest DIC should be retained.

RESULTS

Dimensionality of scores

Results from the four methods of testing dimensionality all converged upon the same conclusion, namely, that the item responses were most parsimoniously represented by a unidimensional construct. The analysis of the correlation matrix for the parametric exploratory analysis yielded eigenvalues of 8.49, 2.00, and 1.77 for the first three estimated coefficients. When comparing the ratios amongst them, the ratio of the first to second eigenvalue was 4.25, which was larger than the ratio of the second and third eigenvalues (i.e., 1.13), suggesting that the structure was essentially unidimensional (Divgi, 1980; Lord, 1980). Moreover, the fit for a one factor solution was excellent, with CFI = .95, TLI = .94, RMSEA = .032 (95 % CI = .019, .043). The parametric confirmatory analysis resulted in identical fit indices as the exploratory model. Non-parametric analyses also provided sufficient evidence for a unidimensional structure. A T statistic of 1.32 was estimated from the DIMTEST model ($p = .095$), leading to a fail-to-reject decision of the null hypothesis that the item responses were unidimensional in the exploratory model. Similarly, a DETECT index of -0.05 was estimated for the confirmatory model, which was less than the desired 0.20 for a unidimensional model (Jang & Roussos, 2007).

Classical test theory

Given the evidence for the unidimensionality of the item responses, descriptive statistics for the accuracy of item responses and response times were calculated and reported in Table 1; there were no missing data for this sample. The item p values ranged from 0.23 to 0.81, indicating a range of difficult to easy items and the average proportion correct was 0.60. Internal consistency, as measured by Cronbach's alpha, was initially estimated as $\alpha = .80$. Item-to-total correlations were also estimated and broadly suggested that item responses were moderately associated with overall total test score performance. Three of the items were noted as either uninformative [i.e., item 28, $r(1) = .02$, $p = .42$] or mis-informative [i.e., item 29, $r(1) = -.15$, $p = .35$ and item 30, $r(1) = -.13$, $p = .46$]. Near-zero item-total correlations are not desired as they indicate students who correctly answer the item may obtain either low or high total scores. Similarly, negative scores are problematic as they suggest that students who correctly answer an item tend to have low overall test scores. Because of these poor item statistics, they were removed for both the traditional IRT and CIRT modeling. When these items were removed from the internal consistency analysis, Cronbach's alpha improved to .81.

The response-time data indicated that students spent an average of 15.21 s per item ($SD = 9.67$), and ranged from 12.20 s (item 17) to 20.36 s (item 28). Several observations concerning the response-time data are worth noting. The mean and standard deviations were correlated at $r(1) = .81, p < .001$ which suggested that items on which the students spent the longest also demonstrated the greatest variability in time spent across the sample, yet evaluating the data from Table 1 demonstrated that this correlation may vary conditionally on the mean response time. For items where the average response time was long (e.g., items 7 and 28), students tended to vary in their average responses to those items ($SD = 11.92$ and 13.49). Conversely, items with short average response times, such as items 19 and 20, presented with standard deviations which illustrates less variability in the average response ($SD = 8.96$ and 6.55 , respectively). In order to more fully explore this relation, quantile regression (Koenker & Bassett, 1978; Petscher & Logan, 2014; Petscher, Logan, & Zhou, 2013) via the quantreg package (Koenker, 2013) in R (2013) was used to test if the association between average response time and the variance in response time was conditional on the average response time. At the .20 quantile (or approximately 20th percentile) of mean response time, the correlation between response time and variance in response time was $r(1) = .58, p = .02$, compared with the .25 quantile [$r(1) = .65, p = .004$], the .75 quantile [$r(1) = .88, p < .001$], and the .80 quantile [$r(1) = .87, p < .001$]. This result confirmed what was seen in the descriptive association. At lower levels of mean response time (i.e., faster response), the relation between the mean and variance of response time was more variable compared to when average response time was slower.

Further, item p values were associated with the variance of response times [$r(1) = -.37, p = .04$] with a trend that easier items related to less spread across the sample in response time; however, the relation between p value and average response time was not statistically significant [$r(1) = -.23, p = .22$].

Item response theory (IRT) analysis

For the IRT analyses, Rasch and two-parameter logistic (2pl) models were estimated. A comparison of log likelihoods between the two models favored the 2pl ($\chi^2 = 56, df = 27, p < .001$). Item discrimination and difficulty parameters for both models are reported in Table 1. Item difficulties for the 2pl model ranged from -2.61 to 0.59 and correlated with the classical test p values at $r(1) = -.89, p < .001$. Despite the difference in metrics between the classical and item response approaches, the negative direction of the correlation indicates that items which are identified as easy in the classical framework (i.e., high p value) were also easy in the item response analysis (i.e., negative b value). The item discriminations in the 2pl model ranged from 0.30 to 2.22 ; values between 0.80 and 2.00 are often considered optimal (de Ayala, 2009). Similar to the relation between the classical test and item response difficulties, the 2pl discrimination parameter was strongly correlated with the item-to-total statistic at $r(1) = .89, p < .001$.

Conditional item response theory (CIRT) analysis

Each of the four CIRT models were estimated, and as part of the model evaluation, it was of interest to evaluate the fit of the models as well as the extent to which resulting theta scores differentially correlated with speed. Scatterplots for the relation between ability and speed

are presented in Fig. 2. It can be seen that the scatter did not meaningfully differ across Model 1 [Fig. 2a; $r(1) = .29, p = .003$], Model 2 [Fig. 2b; $r(1) = .31, p = .003$], Model 3 [Fig. 2c; $r(1) = .30, p = .003$], or Model 4 [Fig. 2d; $r(1) = .32, p = .003$], and that the relation was moderate in nature such that individuals with higher ability tended to respond to items more quickly.

Given the comparability of ability and speed, the model fit was evaluated (Table 2). Models 2 and 4 provided the most parsimonious fit as evidenced by the DIC (Model 2 = 3,216.20, Model 4 = 3,217.27), while Models 1 and 3 were comparatively worse (3,312.22 and 3,319.89, respectively). $\Delta DIC \leq 5$ suggests practically important model fit discrepancies, with the lower value model selected; however, $\Delta DIC < 5$ suggests both models should be considered. Given the present models, the DDIC for Model 2 and 4 compared to Models 1 and 3 was ~ 100 , but the ΔDIC between Model 2 and 4 was 1.07, suggesting that while both Models 2 and 4 were not practically differentiated from each other, they provided superior fit to Models 1 and 3. The primary difference between Models 2 and 4 in specification is that the former constrained the speed discrimination parameter to 1, while the latter freed this for estimation.

Table 3 reports the item response parameters (i.e., difficulty and discrimination) and response-time parameters (i.e., intensity and speed) for Models 2 and 4. The results for the item response portion of the models mapped on well to those estimated by the 2pl IRT models. Correlations between the 2pl and CIRT discriminations were near perfect for both Models 2 and 4 [$r(1) = .99, p < .001$] as were the difficulty values [$r(1) = .99, p < .001$]. Moreover, the absolute difference in magnitude for the discrimination parameters was 0.09, and for the item difficulties the absolute difference was 0.04 between the two approaches. The correlation between the ability score (i.e., θ) and response ability (i.e., ζ) was $r(1) = .31, p < .001$ for both Models 2 and 4, which suggested that a moderate association existed between accuracy and speed whereby individuals with higher ability responded to items more quickly than lower ability individuals. Interestingly, the estimated correlations among the item parameters indicated that no relation existed between the item difficulty and intensity [i.e., $r(1) = .02, p = .35$]. A review of the estimated intensity parameters (Table 3) shows that values do not considerably vary, thus, more difficult items did not require more time, yet individuals with higher ability spent less time per item.

An important ancillary consideration when evaluating the items from the CIRT model is the extent to which a good match existed between the sample and prior distributions used for the response-time model. These can be evaluated via P-P plots whereby the sample distribution (shown as points) is plotted against the prior distribution (shown as a line); the extent to which the plotted points deviate from the prior distribution provides evidence that the sample distribution was biased. Resulting plots (Fig. 3) demonstrated that little bias existed in the sample distribution for the VKT items.

Comparison of model-based reliability

As previously noted, an expected benefit of the CIRT model is that the standard error of the estimated ability score should be lower when compared to traditional IRT as well as classical test theory. It is possible to estimate the marginal reliability of scores, with the resulting

value allowing for a meaningful comparison to an estimate of internal consistency from classical test theory (Andrich, 1988; Embretson & Reise, 2000). Marginal reliability is computed as a function of the variance of the estimated h scores and the average of the squared standard errors of θ . In the present study, the marginal reliability was estimated at .83 for CIRT Models 2 and 4, and 0.78 for the IRT 2pl model. Though this is close to what was observed in the classical test model ($\alpha = .80$), it is only representative of the average relation between ability and error. A more useful heuristic for evaluating the relation lies in plotting the standard errors for the CIRT and IRT models, as it is possible to view where each model is differentially reliable across the range of abilities, as well as how they differ if one were to assume the fixed standard error from classical test theory. Figure 4 plots the standard errors of ability from the 2pl item response model (circles), CIRT Model 2 (crosses), and CIRT Model 4 (triangles). Additionally, three horizontal reference lines are included, corresponding to the observed classical test reliability in the current sample (i.e., $\alpha = .81$; solid line), $\alpha = .85$ (dashed line), and $\alpha = .90$ (dotted line). It should be noted that the standard error associated with each alpha index was converted to an IRT scale in order to allow for a direct comparison between the two theoretical approaches (Dunn, Baguley, & Brunson, 2013; McDonald, 1999).

Several characteristics of this graph are worth noting; first, both the classical test theory estimate of internal consistency and the IRT/CIRT marginal reliability coefficients assume that the error is constant for all students, as evidenced by the solid horizontal reference lines. Based on the plots from both types of item response models, this assumption was not tenable. The 2pl item response standard errors varied considerably and were the lowest (i.e., ability was most reliable) for individuals whose estimated ability was lower than average. It can be seen that for students whose 2pl IRT ability ranged from -2.00 to approximately 0.50 , their ability was under the dashed line indicating their scores were minimally reliable at $\alpha = .85$, and up to $\theta \sim .80$ reliability was equal to the classical test theory estimate of $\alpha = .81$. Conversely, when $\theta > 0.80$, the ability score was less precise, and thus less reliable, than that estimated by classical test theory. CIRT models approximated the IRT model in the reliability of ability scores when θ ranged from -2.00 to -0.80 . Even within this specified range it can be seen that the standard errors estimated by the CIRT models were below the dotted line, which corresponded to reliability of $\alpha = .90$. Further, for θ values > -0.80 , both CIRT models consistently outperformed the 2pl IRT model and yielded more reliable estimates of ability, especially for individuals whose ability was average or above average.

As the relative difference in standard errors between the IRT and CIRT models varied, conditional on h , it follows that an important contextual consideration is quantifying the conditional impact of the CIRT model on the reliability of resulting h scores for the sample. This was evaluated by first computing an estimate of efficiency reflecting the observed percentage change in the standard error of θ from the IRT model when response latency was accounted for in the CIRT model. Across the full range of ability scores, the CIRT model (e.g., Model 4), resulted in an average 4.9 % reduction ($SD = 3.8$ %) in the standard error of individual scores. Given the high variance relative to the mean, an analysis of variance was conducted to determine the extent to which efficiency for the CIRT model was greater for low ability ($\theta < -0.50$; $N = 67$), average ability ($-0.50 < \theta < 0.50$; $N = 81$), or high ability ($\theta > 0.50$; $N = 64$) individuals in the sample. Results indicated a strong effect for ability

groups in efficiency [$F(2, 209) = 160.29, p < .001$], with students who were categorized as low ability gaining little from the CIRT model ($M = 0.91\%$, $SD = 1.65\%$) compared to either the average ($M = 5.48\%$, $SD = 2.26\%$) or high ability ($M = 8.37\%$, $SD = 3.15\%$) students. All pairwise comparisons were statistically significant ($p < .001$), with Hedge's g effect sizes demonstrating that efficiency of the model was stronger for average ability compared to low ability students ($g = 2.26$), as well as high ability compared to either low ($g = 2.97$), or average ($g = 2.26$) ability differences.

Such stark differences in the model efficiency in favor of the high ability students, coupled with higher variance in efficiency for those individuals warranted further exploration. A scatter plot was generated (Fig. 5) which plotted ability against the CIRT model efficiency for the full sample. Within this plot, the three ability groups are denoted by different markers, but they are further distinguished by whether the student was below the mean in average response time (i.e., faster) or above the mean (i.e., slower), whereby fast students are represented by the open shapes, and slower students by the filled shapes. It can be observed that low and average ability students have an approximately equal number of individuals who were fast or slow, whereas the high ability students maintained a stronger representation of fast students. The non-linear shape of the scatter is largely marked by the variation of these high ability, fast response students, in that some of these individuals received a precision benefit to their ability score, while others did not.

To more fully explore this phenomenon, elements of Figs. 4 and 5 were combined to evaluate the relations among ability, the standard error of ability, and the percent efficiency for the CIRT model (Model 4; Fig. 6). This plot highlights what has been previously observed, namely, that CIRT model efficiency is strongest for the higher ability students, as well as that the standard errors for the full sample is lowest for lower ability individuals in the sample. What this further illuminated was that there appeared to be diminishing returns in response speed as it pertained to ability and its standard error. Note that as ability and standard error increased, so did the impact of the efficiency of the CIRT model, yet once ability exceeds 1, the efficiency decreased.

DISCUSSION

In the present study, the relation among estimated item parameters from classical test theory, item response theory, and conditional item response theory analyses were evaluated. Relations between response accuracy and speed were studied, evaluating the extent to which the three measurement approaches provided different information concerning the reliability or precision of student ability scores. Through this exploration, the value of adding response time above the information provided by accuracy alone was considered. Overall, the correlations for an item parameter value across the three approaches were very high. Classical test p values were strongly associated with IRT and CIRT item difficulties, as were the item discriminations and item-to-total correlations. Such relations were not surprising given the known approximations of item parameters in item response models to those in classical test theory (de Ayala, 2009). Similarly, the results demonstrated that an item response analysis of the data yielded ability scores with varying levels of precision dependent on where in the distribution the ability score was estimated. In this way, a

classical test theory approach to reliability limits the extent to which total scores can be viewed as reliable based on a traditional index of reliability.

Notable new findings in this study were that an empirical relation between item response ability and response time could be estimated, as well as that the CIRT model improved on the reliability of student scores by yielding lower overall standard errors associated with the individual ability scores. Yet it was also seen that the extent to which the reliability improved was dependent on the ability level of the individual. Consequently, it is possible that speed is a less important construct to account for when one's ability on the administered task is high and the precision of the ability score is low. Notwithstanding the conditional effect, these results indicated that response time is a valuable consideration that can be incorporated into a measurement model.

While several previously published studies have evaluated response-time based item response models, much of the literature has focused on simulations of estimation techniques (van der Linden, Klein Entink, & Fox, 2010; Wang & Hanson, 2005) or applied models in personality research (Ferrando & Lorenzo-Seva, 2007; Ranger & Kuhn, 2012). The current work adds to the body of applied research by evaluating CIRT models with data on a task tapping reading, vocabulary, and morphological elements. When assessing the relation between the individual ability score and speed, it was found that a moderate association existed ($r \sim .30$), but no association between the item difficulty and speed was found ($r = .02$). The lack of an association between difficulty and speed is inconsistent with previous research, which has found strong, positive correlations between the two (Prindle 2012; Verbic & Tomic, 2009); the lack of corroboration here presents a particular direction for future research.

Similar to other published studies (Ferrando & Lorenzo-Seva, 2007; van der Linden et al., 2010), the CIRT model was found to produce lower average standard errors for ability scores, suggesting that more reliable estimates of ability could be produced by accounting for item response times. Interestingly, when this model was disaggregated by ability groupings, it was shown that the CIRT model had little impact for those with lower ability scores and was only more beneficial for those who were average or high ability. This finding should not diminish the advantage that the CIRT model maintained, as lower standard errors were observed for 87 % of the total sample. However, such observations warrant future research to understand the extent to which the CIRT model is conditionally beneficially on ability. It is plausible that this phenomenon occurred due to a lack of more difficult items for the high ability individuals. When reviewing the range of item difficulties in Table 1, only items 3, 11, 13, 15, and 27 were positive, indicating they were more difficult; however the most difficult item was $b = 0.59$. Because approximately 30 % of the sample had an estimated ability score that was higher than the most difficult item, the further the ability of the individual is from the item, the less information that item contributes to the overall precision of the ability score. Subsequently, it may not be that response time is less important for individuals with high ability, but more that response time does not improve the precision of the ability score for individuals who are receiving items which are not optimally matched to their ability.

Given the potential positive psychometric benefits of using the CIRT approach to measure speed and accuracy based on the current results, it is of merit to note that the CIRT analysis is no more difficult to execute in practice than is a traditional IRT analysis, and may, in fact be easier due to the specificity of the software package in R requiring minimal programming to conduct the analysis. Further, because response times are an inherent component of time-limited testing, researchers do not need to expand testing time in order to capture data which could be used for modeling purposes. Rather, the response time simply needs to be recorded by the software program and recovered in a data file along with the accuracy of the student response.

The current findings have implications for the work of several groups. Educators in practical settings are continuously looking for time-efficient and reliable methods to determine the performance of their students. Likewise, both test developers and researchers are invested in determining methods to reduce testing time without compromising the psychometrics of scores. Thus, it is possible that models such as CIRT may lend themselves to using ancillary data, such as response time, to improve model estimations and increase testing efficiency. This application may be used to leverage both accuracy and speed in estimating performance in a number of literacy skills, such as decoding, vocabulary, spelling, etc., as well as those in other educational domains.

Limitations and future directions

Several limitations of the current study merit reporting. The sample used was appropriate for basic model-fitting purposes but a larger sample in a future study could evaluate the extent to which the findings can be replicated. Further, while a sample of third-grade students was selected for model comparisons, it is plausible that differential precision might be obtained when considering younger or older students; thus, future research should seek to not only improve on the sample size, but also the diversity of ages studied. An additional consideration is that scores from the present sample captured both response accuracy and response time, but students were neither told that they were limited in time, nor that their response time was being recorded.

Future research should extend the findings here in multiple ways. First, while this manuscript is largely pedagogical with a demonstration of IRT and CIRT compared to classical test theory, a direct comparison between CBM fluency scores and resulting ability scores from a CIRT model was not conducted. Consequently, differences in concurrent validity between the two types of measures, as well as differences in the prediction of proximal and distal outcomes are unknown with the present data. Second, because students were not primed to be aware that item-level response time was being collected by the computer, it is possible that differences in precision could be obtained when students are cognizant that their rate of response is being considered as a performance factor. Third, provided the availability of accuracy and response-time data at the item level, it would be of interest to test the dimensionality of both components to see if evidence is presented for a unidimensional or multidimensional representation of the data. Such findings would provide empirical evidence for the extent to which accuracy and speed could be parsed into multiple components of fluency.

Notwithstanding these noted limitations and considerations, findings from the present study affirmed that fitting an item response model to the accuracy data overcomes the limitation of constant standard error and reliability and allows for an evaluation of the differential reliability of scores. Further, an item response model that jointly models accuracy and response time could be used to improve the precision of an individual's estimated ability score. This may provide an opportunity for others to evaluate the extent to which such models may be useful for characterizing fluent performance across a number of literacy domains.

Acknowledgments

This research was supported by the Institute of Education Sciences (R305F100005, R305A100301) and the National Institute of Child Health and Human Development (P50HD052120).

References

- Adams, MJ. Beginning to read: Thinking and learning about print. Cambridge, MA: MIT; 1990.
- Al Otaiba S, Petscher Y, Pappamihel NE, Williams RS, Drylund AK, Connor CM. Modeling oral reading fluency development in Latino students: A longitudinal study across second and third grade. *Journal of Educational Psychology*. 2009; 101:315–329. [PubMed: 25132688]
- Andrich, D. Rasch models for measurement. Sage Publications; 1988.
- Ardoin SP, Christ TJ. Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMS web and an experimental passage set. *School Psychology Review*. 2009; 38:266–283.
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin*. 1990; 107:238–246. [PubMed: 2320703]
- Bentler PM, Bonett DG. Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*. 1980; 88:588–600.
- Blackwell CK, Lauricella AR, Wartella E, Robb M, Schomburg R. Adoption and use of technology in early education: The interplay of extrinsic barriers and teacher attitudes. *Computers & Education*. 2013; 69:310–319.
- Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociological Methods and Research*. 1992; 21:230–258.
- Cattell RB. Concepts and methods in the measurement of group syntality. *Psychological Review*. 1948; 55:48–63. [PubMed: 18910281]
- Chard DJ, Vaughn S, Tyler B. A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*. 2002; 35(5):386–406. <http://search.proquest.com/docview/619935634?accountid=4840>. [PubMed: 15490537]
- Christ TJ, Ardoin SP. Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology*. 2009; 47:55–75.
- Christ TJ, Silberglitt B. Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*. 2007; 36:130–146.
- Christ TJ, Silberglitt B, Yeo S, Cormier D. Curriculum-based measurement of oral reading: An evaluation of growth rates and seasonal effects among students served in general and special education. *School Psychology Review*. 2010; 29:447–462.
- Cummings KD, Atkins T, Allison R, Cole C. Response to intervention. *Teaching Exceptional Children*. 2008; 40:24–31.
- Cunningham AE, Stanovich KE. Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*. 1997; 33(6):934–945. [PubMed: 9383616]
- de Ayala, RJ. The theory and practice of item response theory. New York: Guilford; 2009.

- Deno SL. Developments in curriculum-based measurement. *The Journal of Special Education*. 2003; 37:184–192.
- Divgi, DR. Dimensionality of binary items: Use of a mixed model; Paper presented at the annual meeting of the National Council on Measurement in Education; Boston, MA. 1980 Apr.
- Dunn TJ, Baguley T, Brunsden V. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*. 2013
- Edgar, et al. Neuromagnetic oscillations predict evoked-response latency delays and core language deficits in autism spectrum disorders. *Journal of autism and developmental disorders*. 2013:1–11.
- Educational Testing Service. Test and score data summary for TOEFL internet-based test. Princeton, NJ: Author; 2007.
- Embretson, SE.; Reise, S. Item response theory for psychologists. Mahwah, NJ: Erlbaum Publishers; 2000.
- Ferrando P, Lorenzo-Seva U. An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*. 2007; 31:525–543.
- Foorman BR, Petscher Y, Bishop MD. The incremental variance of morphological knowledge to reading comprehension in grades 3–10 beyond prior reading comprehension, spelling, and text reading efficiency. *Learning and Individual Differences*. 2012; 22:792–798.
- Fox JP, Klein Entink RHK, van der Linden WJ. Modeling of responses and response time with the package CIRT. *Journal of Statistical Software*. 2007; 20:1–14.
- Francis DJ, Santi KS, Barr C, Fletcher JM, Varisco A, Foorman BR. Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*. 2008; 46:315–342. [PubMed: 19083362]
- Fuchs LS, Fuchs D, Hosp MK, Jenkins JR. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*. 2001; 5:239–256.
- Good RH, Simmons DC, Kame'enui EJ. The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*. 2001; 5:257–288.
- Goodglass H, Theurkauf JC, Wingfield A. Naming latencies as evidence for two modes of lexical retrieval. *Applied Psycholinguistics*. 1984; 5:135–146.
- Gray, L.; Thomas, N.; Lewis, L. Teachers' use of educational technology in U.S. public schools: 2009 (NCES 2010–040. 2010. Retrieved from the U.S. Department of Education, National Center for Educational Statistics, Institute of Education Sciences. <http://nces.ed.gov/pubs2010/2010040.pdf>
- Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
- Jang EE, Roussos L. An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*. 2007; 44:1–21.
- Kamil, ML. Vocabulary and comprehension instruction: Summary and implications of the national reading panel findings. In: McCardle, P.; Chhabra, V., editors. *The voice of evidence in reading research*. Baltimore: Paul H Brookes Publishing; 2004. p. 213-234.
- Kim Y-S, Wagner RK, Foster E. Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading*. 2011; 15:338–362. [PubMed: 21747658]
- Klein Entink RH, Kuhn J-T, Hornke LF, Fox J-P. Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*. 2009; 14:54–75. [PubMed: 19271848]
- Koenker, R. Quantreg: Quantile regression. R package version 4.98. 2013. <http://CRAN.R-project.org/package=quantreg>
- Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978; 46:33–50.
- Kolen, MJ.; Brennan, RL. Test equating: Methods and practices. 2nd. New York: Springer-Verlag; 2004.
- LaBerge D, Samuels SJ. Toward a theory of automatic information processing in reading. *Cognitive Psychology*. 1974; 62:293–323.

- Logan JAR, Petscher Y. School profiles of at-risk student concentration: Differential growth in oral reading fluency. *Journal of School Psychology*. 2010; 48:163–186. [PubMed: 20159224]
- Lord, FM. Applications of item response theory to practical testing problems. New York: Enrlbaum Associates; 1980.
- McDonald, RP. Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum; 1999.
- Mercer SH, Dufrene BA, Zoder-Martell K, Harpole LL, Mitchell RR, Blaze JT. Generalizability theory analysis of CBM maze reliability in third- through fifth-grade students. *Assessment for Effective Intervention*. 2012; 37:183–190.
- Metrik, et al. Balanced placebo design with marijuana: Pharmacological and expectancy effects on impulsivity and risk taking. *Psychopharmacology*. 2012; 223:489–499. [PubMed: 22588253]
- Miranda H, Russell M. Predictors of teacher-directed student use of technology in elementary classrooms: A multilevel SEM approach using data from the USEIT study. *Journal of Research on Technology in Education*. 2011; 43:301–323.
- Muthen, LK.; Muthen, BO. Mplus. 7th. Los Angeles, CA: Muthen & Muthen; 1998–2012.
- National Institute of Child Health and Human Development. Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH publication no. 00-4769). 2000
- Perfetti C, Hogaboam T. Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*. 1975; 67:461–469.
- Petscher Y, Cummings KD, Biancarosa G, Fien H. Advanced (measurement) applications of curriculum-based measurement of reading. *Assessment for Effective Intervention*. 2013; 38:71–75. [PubMed: 26346551]
- Petscher Y, Kim YS. The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*. 2011; 49:107–129. [PubMed: 21215838]
- Petscher Y, Logan JAR. Quantile regression in the study of developmental sciences. *Child Development*. 2014; 85:861–881. [PubMed: 24329596]
- Poncy BC, Skinner CH, Axtell PK. An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*. 2005; 23:326–338.
- Pressey, B. Comparative analysis of national teacher surveys. 2013. http://www.joanganzcooneycenter.org/wp-content/uploads/2013/10/jgcc_teacher_survey_analysis_final.pdf
- Prindle JJ. A functional use of response time data in cognitive assessment. Doctoral dissertation. Retrieved from USC Digital Library. 2012
- R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. <https://www.R-project.org/>
- Ranger J, Kuhn J-T. Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*. 2012; 36:214–231.
- SAS Institute Inc. Base SAS 9.3 procedures guide. Cary, NC: SAS Institute Inc; 2011.
- Scarborough, HS. Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In: Neumann, S.; Dickinson, D., editors. *Handbook for research in early literacy*. New York: Guilford; 2001. p. 97-110.
- Scheiblechner, H. Psychometric models for speed-test construction: The linear exponential model. In: Embreston, SE., editor. *Test design developments in psychology and psychometrics*. New York: Academic Press; 1985. p. 219-244.
- Schnipke, DL.; Scrams, DJ. Exploring issues of examinee behavior: Insights gained from response-time analyses. In: Mills, CN.; Potenza, MT.; Fremer, JJ.; Ward, WC., editors. *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates; 2002.
- Sireci SG, Thissen D, Wainer H. On the reliability of testlet-based tests. *Journal of Educational Measurement*. 1991; 28:237–247.
- Stout WF. A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*. 1987; 52:589–617.

- Sternberg S. Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*. 1969; 57:421–457. [PubMed: 5360276]
- Tate R. A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*. 2003; 27:159–203.
- van der Linden WJ. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*. 2007; 72:287–308.
- van der Linden WJ. Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*. 2011; 53:334–358.
- van der Linden WJ, Klein Entink RH, Fox J-P. IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*. 2010; 34:327–347.
- van der Linden WJ, van Krimpen-Stoop EMLA. Using response times to detect responses in computerized adaptive testing. *Psychometrika*. 2003; 68:251–265.
- Verbic, S.; Tomic, B. Test item response time and the response likelihood. 2009. <http://arxiv.org/ftp/arxiv/papers/0901/0901.4356.pdf>
- Wainer, H.; Bradlow, ET.; Wang, X. Testlet response theory and its applications. New York: Cambridge University Press; 2007.
- Wang T, Hanson B. Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*. 2005; 29:332–339.
- Wolf M, Katzir-Cohen T. Reading fluency and its intervention. *Scientific Studies of Reading*. 2001; 5(3):211–239.
- Zeno, SM.; Ivens, SH.; Millard, RT.; Duvvuri, R. The educator's word frequency guide. NY: Touchstone Applied Science Associates Inc; 1995.
- Zhang J, Stout W. The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*. 1999; 64:213–249.

Fig. 1.
Graphical depiction of the conditionally independent item response model

Fig. 2. Scatterplots of the relation between estimated ability score and speed for the conditional item response model for Model 1 (a), Model 2 (b), Model 3 (c), and Model 4 (d)

Fig. 3.
Item P-P plots for CIRT Model 4

Fig. 4. Plotted standard errors of ability for a 2pl IRT model, CIRT Model 2, and CIRT Model 4 with *horizontal reference lines* corresponding to $\alpha = .81$ (*solid line*), $\alpha = .85$ (*dashed line*), and $\alpha = .90$ (*dotted line*)

Fig. 5. Scatterplot of the relation between estimated ability and CIRT model efficiency conditional on average response time

Fig. 6. Scatterplot of the relations among estimated ability, the standard error of ability, and CIRT model efficiency

Table 1

Classical test theory and item response theory item statistics

Item	<i>p</i> value	Item-total <i>r</i>	Mean <i>RT</i>	SD <i>RT</i>	Rasch IRT		2pl IRT	
					<i>a</i>	<i>b</i>	<i>a</i>	<i>b</i>
1	.75	.17	17.93	12.26	1.00	-1.30	0.43	-2.61
2	.63	.34	15.62	8.72	1.00	-0.64	0.99	-0.65
3	.43	.34	19.19	10.75	1.00	0.32	0.89	0.34
4	.83	.35	15.68	8.93	1.00	-1.93	1.36	-1.56
5	.64	.42	17.78	10.95	1.00	-0.72	1.19	-0.64
6	.78	.44	14.11	8.27	1.00	-1.51	1.73	-1.09
7	.73	.44	20.05	11.92	1.00	-1.19	1.59	-0.90
8	.53	.40	15.46	11.70	1.00	-0.17	1.11	-0.17
9	.65	.35	13.89	8.71	1.00	-0.74	0.96	-0.76
10	.85	.42	13.10	8.41	1.00	-2.09	2.22	-1.33
11	.45	.23	17.52	11.85	1.00	0.24	0.57	0.39
12	.67	.19	13.75	7.64	1.00	-0.84	0.47	-1.53
13	.39	.32	13.31	7.57	1.00	0.53	0.85	0.59
14	.80	.37	13.36	6.76	1.00	-1.64	1.44	-1.29
15	.42	.35	14.93	9.72	1.00	0.41	0.89	0.44
16	.70	.38	16.76	10.84	1.00	-1.05	1.18	-0.94
17	.69	.50	12.20	6.95	1.00	-1.00	1.87	-0.72
18	.58	.40	14.75	8.98	1.00	-0.38	1.18	-0.35
19	.81	.22	13.79	8.96	1.00	-1.71	0.92	-1.80
20	.60	.53	13.23	6.55	1.00	-0.50	1.80	-0.38
21	.73	.47	12.76	7.74	1.00	-1.19	1.81	-0.85
22	.65	.29	15.75	11.53	1.00	-0.74	0.86	-0.82
23	.65	.27	16.08	10.50	1.00	-0.74	0.64	-1.03
24	.56	.47	14.54	10.67	1.00	-0.29	1.50	-0.24
25	.60	.54	12.60	9.68	1.00	-0.50	1.84	-0.38
26	.52	.42	14.08	9.36	1.00	-0.13	1.13	-0.12
27	.49	.12	14.92	10.19	1.00	0.04	0.30	0.13

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Item	<i>p</i> value	Item-total <i>r</i>	Mean <i>RT</i>	SD <i>RT</i>	Rasch IRT		2pl IRT	
					a	b	a	b
28	.26	.02	20.36	13.49	–	–	–	–
29	.32	–.15	13.19	8.00	–	–	–	–
30	.23	–.13	15.72	12.41	–	–	–	–

RT = response time, a = item discrimination, b = item difficulty

Table 2

Conditional item response theory analysis model fit

Model	$\bar{D}$	p_D	DIC	log likelihood
1	2,843.17	469.05	3,312.22	-3,035.38
2	2,737.80	478.40	3,216.20	-2,981.77
3	2,826.94	492.95	3,319.89	-3,029.31
4	2,718.44	498.83	3,217.27	-2,974.34

Model 1 = one-parameter response and response-time, Model 2 = two-parameter response, one-parameter response-time, Model 3 = one-parameter response, two-parameter response time, Model 4 = two-parameter response and response time. $\bar{D}$ = posterior mean of the deviation, p_D = effective number of model parameters, DIC = deviation information criteria

Table 3

CIRT item response and response-time parameters for Models 2 and 4

Item	Model 2				Model 4			
	Response		Response		Response		Response	
	a	b	α	β	a	b	α	β
1	0.46	-2.52	1.00	1.19	0.4	-2.52	0.78	1.19
2	1.09	-0.61	1.00	1.14	1.09	-0.61	1.05	1.14
3	1.02	0.28	1.00	1.22	1.02	0.27	1.18	1.22
4	1.36	-1.51	1.00	1.14	1.38	-1.49	1.08	1.14
5	1.33	-0.62	1.00	1.19	1.33	-0.62	1.11	1.19
6	1.68	-1.07	1.00	1.09	1.70	-1.07	1.08	1.09
7	1.73	-0.84	1.00	1.24	1.72	-0.85	0.94	1.24
8	1.26	-0.18	1.00	1.11	1.26	-0.18	1.21	1.11
9	1.07	-0.71	1.00	1.08	1.07	-0.71	0.96	1.08
10	2.01	-0.32	1.00	1.05	1.99	-1.32	1.10	1.05
11	0.66	0.31	1.00	1.18	0.68	0.33	1.11	1.18
12	0.53	-1.45	1.00	1.09	0.51	-1.50	0.85	1.09
13	0.94	0.53	1.00	1.07	0.95	0.52	0.89	1.07
14	1.51	-1.22	1.00	1.07	1.51	-1.22	1.07	1.07
15	0.97	0.40	1.00	1.11	0.99	0.40	0.99	1.11
16	1.29	-0.88	1.00	1.16	1.31	-0.87	1.15	1.16
17	1.99	-0.67	1.00	1.03	1.99	-0.67	1.07	1.03
18	1.28	-0.35	1.00	1.11	1.28	-0.35	1.00	1.11
19	0.94	-1.76	1.00	1.07	0.95	-1.75	1.22	1.07
20	1.90	-0.36	1.00	1.07	1.89	-0.36	0.92	1.07
21	1.89	-0.81	1.00	1.05	1.89	-0.81	0.97	1.05
22	0.95	-0.79	1.00	1.13	0.95	-0.79	1.03	1.13
23	0.71	-0.98	1.00	1.14	0.71	-0.98	0.92	1.14
24	1.63	-0.26	1.00	1.08	1.63	-0.25	1.03	1.08
25	1.87	-0.36	1.00	1.02	1.87	-0.35	1.00	1.02
26	1.28	-0.13	1.00	1.08	1.28	-0.13	0.86	1.08

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Item	Model 2				Model 4			
	Response		Response		Response		Response	
	a	b	α	β	a	b	α	β
27	0.36	0.05	1.00	1.10	0.36	0.10	0.74	1.10

a = discrimination; b = difficulty; α = response time discrimination; β = time intensity