RESEARCH IN THE SCHOOLS
2017, Vol.24, No. 1, 80-93

Comparison of IRT and CTT Using Secondary School Reading Comprehension Assessments

Joanne V. Coggins, Jwa K. Kim, and Laura C. Briggs
Middle Tennessee State University

The Gates-MacGinitie Reading Comprehension Test, fourth edition (GMRT-4) and the ACT Reading Tests (ACT-R) were administered to 423 high school students in order to explore the similarities and dissimilarities of data produced through classical test theory (CTT) and item response theory (IRT) analysis. Despite the many advantages of IRT demonstrated in simulation studies comparing CTT and IRT, it is still necessary to study the practical application of IRT to the measurement of reading comprehension assessment using actual test data in order to determine whether IRT's significant advantages persist. Accordingly, the following research questions were addressed: (a) To what degree does IRT offer a significant advantage over CTT when evaluating student reading comprehension ability? and (b) Which IRT model provides the best fit for the data? Results revealed that IRT's ability to provide more accurate information about item-level properties combined with independent trait calibration minimizing measurement error afforded significant advantages over CTT, thereby strengthening test design. It was determined that the 3-parameter logistic (3-PL) model provided better fit than did the 1-parameter (1-PL) and 2-paramenter (2-PL) models.

Keywords: IRT, reading comprehension measurement, reading comprehension assessment

From a young age, students are assessed regularly for myriad purposes, including evaluating individual educational growth, mastery of concepts, effectiveness of teacher instruction, individual areas of weaknesses for intervention, and prediction of abilities (Pearson & Hamm, 2005). Student assessment primarily focuses on three main areas: measuring the overall health of the school system, measuring the educational growth of the children (Seltzer, Frank & Bryk, 1994), and estimating and predicting future educational ability (Barnes & Wise, 1991; Hambleton & Jones, 1993). Accurate and reliable measurement of student reading comprehension is both important and difficult.

Reading comprehension initially appears simply to be whether or not the reader can understand the written words on a page; however, many factors impact reading comprehension (Perfetti, Landi, & Oakhill, 2005). Gough and Tunmer (1986) theorized in the *Simple View of Reading* that reading comprehension is the resultant combination of listening comprehension and word-level reading ability. Research has shown that the bottom-up decoding skills necessary for word-level reading must co-develop with the complex top-down language processing integral to listening comprehension skills for students to develop adequate reading comprehension (Cutting & Scarborough, 2009; Gough & Tunmer, 1986). Knowledge of semantic and syntactic language structure, vocabulary knowledge, background knowledge, schema development, and inferencing ability all impact reading comprehension (Cutting & Scarborough, 2009; Kintsch, 1998; Perfetti, 2007; Perfetti et al., 2005). In fact, Perfetti et al. (2005) aver that sensitivity to the nuances of the structure of text, comprehension monitoring, and the ability to make inferences are the major factors in comprehension. Additionally, background knowledge, understanding of hyperbole, vocabulary and idiomatic phrases all impact an individual's text comprehension (Kintsch, 1998; RAND Reading Study Group, 2002). Reading researchers agree that reading comprehension is multifaceted, composed of complex cognitive processes that support one another (Cutting & Scarborough, 2009; Kintsch, 1998; Paris & Stahl, 2005). Despite the many different processes involved, reading comprehension often is measured as a unitary construct using multiple-choice instruments due to ease of scoring and the need to measure large numbers of students (Mehta,

Correspondence concerning this article should be addressed to Joanne V. Coggins, Literacy Studies Department, Middle Tennessee State University, Murfreesboro, TN 37132.
E-mail: jvc2j@mtmail.mtsu.edu

Foorman, Branum-Martin, & Taylor, 2009; Pearson & Hamm, 2005).

The burgeoning need for reading comprehension measurement is being met at all grade levels through the increasing use of technology. Computer adaptive testing (CAT) software is now used to provide rigorous, frequent assessment and differentiated learning to diverse learners. As a result, psychometric methods for the analysis of test data must change to accommodate the movement away from paper-and-pencil test instruments toward CAT designs.

As educational measurement needs continue to grow, it becomes increasingly important to select psychometric methods that will maximize useful information. There are two major psychometric theories in education: classical test theory (CTT) and item response theory (IRT). Although CTT has been researched and used for more than 100 years, and is considered to represent a well-rooted test model, IRT models have received extensive attention for the past 45 years (Hambleton & Jones, 1993). Among the many differences between CTT and IRT are the management of test error and the issue of the bidirectional nature of the test instrument and student ability (Hambleton & van der Linden, 1982). Using CTT-based tests, an examinee's total score varies according to different item samples. As the item sample changes, so does an examinee's total score; however, IRT-based tests are not sample or item dependent, eliminating the bidirectional relationship between assessment and examinee ability found in CTT analyses (Rathvon, 2004).

In CTT, the total score of the assessment is analyzed through comparisons of means and standard deviations to compare individual test scores to other students in the sample, or the total score is used to compare sample groups using alternate test forms (Park, Alonzo, & Tindal, 2011). Several assumptions in CTT appear unlikely; yet, these assumptions can neither be proven nor disproven. One assumption is that every student has the same standard error of measurement (SEM) on a test, which seems unlikely given that we know intuitively that individuals vary greatly in their ability levels and even their mental and physical preparedness on any given test day (de Ayala, 2009; Gall, Gall, & Borg, 2003). When using CTT, estimation of test score reliability and item characteristics is dependent upon the similarity of the currently assessed population to the population used during test development. Additionally, in CTT, score reliability should be determined by comparing performance on two parallel or alternate test forms to explore test item measurement error; however, this is seldom undertaken in practice because of the inherent difficulty of developing strictly parallel assessment forms (Gall et al., 2003).

In contrast, IRT offers many opportunities that CTT cannot leverage. By uncoupling the interdependence of the instrument and student ability, IRT makes it possible to estimate student latent ability more accurately (Hambleton, Swaminathan, & Rogers, 1991). IRT produces detailed item-level information that can be used to design item test banks, to calibrate and to scale assessments, and to examine test items for possible bias. Item-level examination of test data provides the opportunity to identify and to withdraw items not sufficiently discriminatory in order to refine the test instrument. Items demonstrating evidence of subgroup bias may be identified and withdrawn as well. Test banks can be created from items measuring and discriminating a targeted range of ability level. Conversely, item discrimination may be examined to determine that a range of items has been provided in order adequately to measure the latent trait of individuals with a wide range of abilities without encountering either ceiling or floor effects when an assessment is either too difficult or too easy for some examinees (Sébille et al., 2010).

Although CTT item difficulty measures are similar to those provided by IRT, the additional item parameters that IRT produces can provide information about item discrimination and examinee guessing. In educational measurement, IRT's additional item-level indices are widely used in computer adaptive testing (CAT) to provide individualized assessment and instruction differentiated for students' individual ability levels. This is undertaken by choosing from pools of items designed to measure different ranges of ability, then using those items comprehensively to assess students within their individually estimated ability ranges. In this way, IRT can be used to streamline testing and to provide detailed diagnostic information about both individual and group performance (Gall et al., 2003; Thomas, 2010). Reading comprehension assessments using this type of methodological analysis provides information about both the amount of ability that the individual items measure, as well as how much ability the students have. Finally, CTT is a tautology, whereas IRT provides a falsifiable model with empirical data.

It is important to evaluate both the test instrument and the types of analytical strategies employed because reading comprehension can be measured only indirectly through assessment of student performance. Therefore, determination of the most appropriate theoretical approach for estimating item difficulty, standard errors, and internal consistency on high-stakes reading comprehension assessments becomes more important as well. Studies comparing CTT and IRT methods in the fields of chemistry, medicine, psychology, and education suggest that the greater analytic benefits of IRT will eventually change the

nature of test design, selection, and scoring as its use becomes increasingly widespread and understood (Magno, 2009; Sébille et al., 2010; Seltzer et al., 1994; Sharkness & DeAngelo, 2011; Thomas, 2011). This is largely due to IRT's greater accuracy in measurement of clinically meaningful variance, reduced measurement error, increased objectivity in item calibration and equating, evaluation of item bias, analyses of item-person fit, and IRT's ability to estimate item difficulty, item discrimination, latent ability, and test difficulty using a single scale (Borg et al., 2003; Seltzer et al., 1994; Thomas, 2011).

Current Study

Purpose

Although there has been extensive research on the ACT Reading Tests (ACT-R; Allen, 2012; Topczewski, Cui, Woodruff, Chen, & Fang, 2013; Westrick & Allen, 2014; Woodruff, Traynor, & Cui, 2013), there has been relatively little psychometric evaluation of the Gates-MacGinitie Reading Tests-fourth edition (GMRT-4; W. MacGinitie, R., MacGinitie, Maria, Dreyer, & Hughes, 2000). Previous versions of the GMRT have been studied (Cooter & Curry, 1989; Graham, 1990; Johns, 1984; Jongsma, 1980; Powell, 1969), but there are no current studies evaluating the score validity or score reliability of the GMRT-4. The concurrent validity of the Dimensions of Self-Concept—level E (DOSC-E) and the Minnesota Reading Assessment both have been compared to earlier versions of the GMRT (Chang & Brown, 1983; Freeman & Hutchinson, 1989). Although secondary and post-secondary school administrators often use both tests to assess students' reading comprehension, no psychometric analyses of comparisons between the GMRT-4 and the ACT-R currently exist. Although designed for different uses, the ACT-R and GMRT-4 both measure adolescent reading comprehension. The ACT-R is designed to differentiate among freshman applicants to 4-year universities (ACT, Inc., 2006), whereas the GMRT-4 provides a measure of reading comprehension skills of secondary and post-secondary school students (www.riversidepublishing.com). A strong correlation between both the GMRT-4 and the ACT-R is implicit in the "professional development" link on the GMRT website that directs consumers to the ACT, Inc. (www.act.org;www.riversidepublishing.com). Both are group-administered tests that students are likely to take, and it is valuable to determine whether differing theoretical models of test analysis provide similar information and comparable measurement of adolescent readers. Simulation studies have demonstrated, in other fields of research, the improved item-level information that is gained using IRT analysis as opposed to CTT (Sébille et al., 2010; Topczewski et al., 2013); yet, IRT analysis of actual test data in the study of reading comprehension is relatively sparse. Thus, the goal of this study was to use actual test data to evaluate item and test characteristics of the ACT-R and GMRT-4 to explore the similarities and dissimilarities of analyses resulting from the use of CTT and IRT theoretical models.

Although their applications differ, both the ACT-R and the GMRT-4 are designed to measure the reading comprehension of developmentally similar readers, and should provide similar person latent trait measures. Theoretically, IRT should offer more accurate item-based and person-based statistics due to the invariance property of IRT model parameters (Byrne, 2010). In other words, the bidirectional dependence of CTT's item-person relationship can be eliminated using IRT, and a more accurate picture of student ability and test function should emerge. In a 2009 study by Magno in which real data from a high school chemistry test was used, IRT's item difficulty indices remained more stable than did CTT's item difficulty indices (i.e., p-values) across differing samples. This stability is advantageous when measuring and comparing test performance of students from narrow demographic groups, such as those demonstrating significant economic disadvantage, to larger samples with wider demographic distributions.

Despite the many advantages of IRT demonstrated in simulation studies comparing CTT and IRT, it is still necessary to study the practical application of IRT to the measurement of reading comprehension assessment using actual test data in order to determine whether IRT's significant advantages persist. Accordingly, the following research questions were addressed: (a) To what degree does IRT offer a significant advantage over CTT when evaluating student reading comprehension ability? and (b) Which IRT model provides the best fit for the data?

Method

Participants

The sample consisted of 423 high school students from two rural high schools in the southeastern U.S. (M_{age} = 16.4; range 15-19) in Grades 9-12. Females constituted 57% of the sample and 22% of examinees spoke a language other than English at home. Ethnicity of the sample was 49% White, 22% Hispanic, 9% African American, 6% multiracial, and 14% not reported. One hundred percent of examinees qualified for two free meals daily, evidencing significant economic disadvantage.

Instruments

The test instruments analyzed were the Gates-MacGinitie Reading Comprehension Tests-fourth edition, form S and the ACT, Inc. (2013) Reading Comprehension Practice Tests. Both multiple choice tests were administered during the

same week in March 2015. The ACT-R (40 items) and the GMRT-4 (48 items) were scored dichotomously as either incorrect or correct, with unanswered items scored as incorrect responses. Publication companies for both assessments state that the single trait of reading comprehension is being measured (ACT, Inc., 2006).

Procedure

Both the ACT-R and GMRT-4 assessments were administered to 423 high school students from 18 classes during a 2-day window. One half of the students took the ACT-R on Day 1 and the remaining one half took the GMRT-4 on the same day. Students were given 35 minutes to complete the ACT-R and 40 minutes to complete the GMRT-4, as per test instructions. Students were asked to answer test items to the best of their ability and were debriefed regarding the purpose of the study after completing testing. Data for both ACT-R and GMRT-4 assessments were analyzed using Xcalibre 4.2 simultaneously to generate CTT summary statistics and IRT 1-parameter logistic (1-PL), 2-parameter logistic (2-PL) and 3-parameter logistic (3-PL) models.

Results

ACT-R

The CTT summary statistics for total scores indicated that the 40-item ACT-R yielded scores that represent a reliable (Cronbach's $\alpha = 0.83$) and difficult ($M_{ACT} = 13.54$, $SD = 7.73$; range 1-40) test of the sample group, $n = 423$ (see Table 1). Analyses of the sample group using the 1-PL, 2-PL, and 3-PL models showed a range mean true ability (theta) estimates: $M_{ACT\ 1\text{-}PLM} = -0.77$, $SD = 0.83$; $M_{ACT\ 2\text{-}PLM} = 0.01$, $SD = 0.88$; $M_{ACT\ 3\text{-}PLM} = -0.04$, $SD = 0.93$. Data indicated that the overall IRT model fit of the 2-PL and 3-PL IRT models was good. The ΔX^2 –test of the 1-PL model to the 2-PL model indicated a statistically significant difference between the models in favor of the 2-PL model, $\Delta X^2(40) = 1{,}034$; $p < .05$. The ΔX^2 –test of the 1-PL and 3-PL models revealed that the 3-PL model, $\Delta X^2(80) = 819$; $p < .05$, was a better fitting model to the data than was the 1-PL model. Finally, ΔX^2 –test of the 2-PL and 3-PL models determined that the 3-PL model was the best fit to the ACT-R data, $\Delta X^2(40) = 215$; $p < .05$.

Comparisons of item difficulty means of the 1-PL, 2-PL, and 3-PL models revealed that item difficulty level increased with the addition of each variable into the model. The 1-PL model had the lowest item difficulty level among the three models, $b = 0.00$; $SD = 1.00$. Mean item difficulty for the 2-PL model was greater than for the 1-PL model at $b = 0.98$; $SD = 0.57$; and mean item difficulty for the 3-PL model, $b = 1.75$; $SD = 0.53$, was the greatest. Comparison of item discrimination mean values revealed that the 3-PL model, $a = 0.98$; $SD = 0.15$, had higher discrimination and standard deviation values than did the 2-PL model, $a = 0.49$; $SD = 0.10$.

Evaluation of each of the three IRT models demonstrated the unique information provided by each model and delivered further evidence of the superior fit of 3-PL model to the ACT-R data. The 1-PL model identified 24 of 40 items as poorly performing due to high chi-square values. Accordingly, these items also demonstrated z-residual values greater than 2.0, resulting in p values smaller than the accepted limit of 0.05. Further analysis of the 1-PL model was unnecessary due to the number of poorly performing items and the clear indication by the ΔX^2 –test that the 1-PL model resulted in the poorest fitting model to the data of the three IRT models considered.

When evaluating IRT models, goodness-of-fit may be determined by comparing -2 log likelihood (-2LL) values, visual inspection of item-person maps and test information function plots, and analysis of item characteristic curves. The 2-PL model had acceptable item-level model fit. The item-person map of the 2-PL model and a comparison of examinee ability with the difficulty of the test visually showed a better match between examinee ability levels and item difficulty levels when compared to the item-person map of the 1-PL and 3-PL models. The test information function (TIF) of the 2-PL model showed the maximum information of 6.74 at *theta* = 0.70 with a conditional standard error of measurement (CSEM) = 0.39. Taken together, the TIF, CSEM, and theta values measure the precision of the assessment. Due to these factors, the test response function (TRF) of the 2-PL model showed a gently shaped s-curve, evidencing a lackluster number of items that an examinee would be predicted to answer correctly at different theta locations. This was demonstrative of the relatively few items at a difficulty range within the abilities of the examinees in this study. That the TRF curve did asymptote toward zero is another indication there might be relatively little examinee guessing occurring in this assessment.

All 40 test items fit the predicted item response of the 3-PL model. The additional item discrimination variable included in this model visually improved the location overlap between item difficulty and the ability of the examinees shown by the item-person map. Further, all assessment items were a good fit to the 3-PL model. The additional guessing variable significantly improved this model over the 1-PL and 2-PL models. Besides the visual mismatch evident in the item-person map, the maximum TIF= 14.55 and CSEM = 0.26 at *theta*= 2.10 further indicated that the 3-PL model was the best-fitting model to the ACT-R data. Summary statistics for item difficulty, item discrimination,

and summary theta estimates for the 3-PL model—the best fitting model to the data—tells the tale of how difficult this test is relative to examinee ability (see Table 2). The overall health of the test is visually evident when considering the comparison of person theta values to item difficulty demonstrated by the 3-PL model item-person map.

Comparisons between CTT and IRT 3-PL model item indices showed some similarities and dissimilarities between the two theoretical models. The *P*-value in CTT indicates the probability that an examinee will answer that item correctly. CTT's *P*-value correlates to the *b*-value in IRT that estimates an item's difficulty. Although scales differ between theoretical models, *P*- and *b*-values measure the difficulty of each item. IRT and CTT item indices failed to identify the same items as either the most difficult or the least difficult (see Table 3). CTT identified Item 2, $b = 0.96$; $p = .51$, as the least difficult item, whereas CTT identified Item 12, $b = 0.82$; $P = 0.44$ as being the least difficult item. CTT analysis identified the most difficult item as Item 39, $b = 2.37$, $P = .19$; whereas IRT identified Item 11, $b = 2.81$, $P = .20$ as the most difficult. Three of the six most difficult items identified by CTT also were the only three items with c-values (pseudo-guessing variable) greater than 0.25, indicating elevated estimates of examinee guessing.

Comparisons of IRT's item discrimination *a*-value and CTT's item-test correlation *R*-value provide further evidence of the structural dissimilarities between these theoretical models. CTT identified Item 29, $R = 0.45$; $a = 1.19$, as the item having the greatest discrimination value. In contrast, IRT identified Item 37, $R = 0.31$; $a = 1.22$, as having the greatest discrimination value. Item 14, $R = 0.16$; $a = 0.76$, and Item 38, $R = 0.16$; $a = 0.98$, both were identified by CTT as having the lowest discrimination values; whereas IRT identified Item 8, $R = 0.24$; $a = 0.72$, and Item 9, $R = 0.27$; $a = 0.72$, as providing the least discrimination information among examinees.

As with identification and ranking of item difficulty levels, CTT and IRT failed to rank items by magnitude of discrimination levels in the same order. The *R*-values provided by CTT only varied from a low of 0.16 to a high of 0.45, whereas IRT's *a*-values varied from a low of 0.72 to a high value of 1.22. The wider range of values provided by IRT's item discrimination *a*-values allowed for a greater degree of comparison among items compared to the range of CTT's item-test correlation *R*-values. Similarly, the b-values produced by IRT provided a greater range between the lowest and highest values (0.82 to 2.81) than did CTT's P-values (0.19 to 0.51). By providing a broader range of item indices values, IRT provided greater opportunity for differentiation among items than did CTT, thereby increasing analytical acuity. Although CTT and IRT analyses of ACT-R data initially appear to measure and to identify item difficulty and discrimination in a similar manner, further investigation revealed that IRT provided more precise item-level information than did CTT.

GMRT

Summary statistics for CTT revealed that the 48-item GMRT yielded good internal reliability, $\alpha = 0.91$. The mean total score was 26.82 ($SD = 9.74$) (see Table 1), with an examinee response range of 7 to 47 correct. Analyses of the sample group using 1-PL, 2-PL, and 3-PL models showed the following range mean true ability (theta) estimates: $M_{GMRT\ 1\text{-}PLM} = 0.31$, $SD = 1.01$; $M_{GMRT\ 2\text{-}PLM} = -0.01$, $SD = 0.95$; and $M_{GMRT\ 3\text{-}PLM} = 0.00$, $SD = 0.99$. These data indicated that the overall IRT model fit of the 2-PL and 3-PL IRT models was good; however, four items in the 1-PL model were identified as having *z*-residual scores greater than 2.0. At the item level, the 2-PL and 3-PL models had no items outside acceptable parameters for the item models.

The ΔX^2 –tests were performed comparing IRT models for goodness-of-fit to the data using -2LL. The 2-PL model was a better fit to the model data compared to the 1-PL model, $\Delta X^2(48) = 539$; $p < .05$. Comparison of the 1-PL to 3-PL model identified the 3-PL model as the better fitting model to the data, $\Delta X^2(96) = 612$; $p < .05$. Finally, a statistically significant difference was found in favor of the 3-PL model compared to the 2-PL model, $\Delta X^2(48) = 73$; $p < .05$. Due to the number of poorly performing items and the results of the ΔX^2 –test showing that the 1-PL model was the least fitting model to the data, no further analysis of this model was conducted.

In both 2-PL and 3-PL models, item-person maps visually indicated that item difficulty and examinee ability locations overlapped considerably, an additional indicator of goodness-of-fit of the model to the data. Summary theta estimates for the 2-PL and 3-PL models were very similar: $M_{2\text{-}PLM} = -0.01$ ($SD = 0.95$) and $M_{3\text{-}PLM} = 0.00$ ($SD = 0.99$). Test item functions for the 2-PL and 3-PL models showed that additional information was gained by adding the guessing parameter to the full model: $TIF_{2\text{-}PLM} = 15.26$ at *theta* = 0.65 and $TIF_{3\text{-}PLM} = 16.73$ at *theta* = 0.50. Visual inspection of the item characteristic curves of the 2-PL models shows that seven of 48 items (15%) did not asymptote toward zero. Item characteristic curves that do not asymptotically approach zero suggest some guessing components (Weiss & Von Minden, 2012). The graph of the 3-PL model visually showed the most steeply sloped test response function of the three models. In the 3-PL model, three items had *c*-values greater than 0.25, indicating that the pseudo-guessing variable contributed to the model. The strong s-shape of the test response function indicated that the test has

good discrimination among student abilities. After evaluating these data, the 3-PL model was selected as the best fitting model to the data.

Item statistics produced by CTT and item indices produced by the IRT 3-PL model revealed more dissimilarities than similarities (See Table 4). Initially, CTT and IRT jointly identified the most and least difficult items. Item 1 was identified as the easiest item, P = .87; b = -1.90; whereas both CTT and IRT identified Item 11 as the most difficult item, P = 0.24; b = 1.84. Analysis of item discrimination values, however, showed significant disagreement between the two analytical methods. Specifically, evaluation of the pseudo-guessing parameter estimates showed that seven items had *c*-values greater than 0.25. Of these seven items, Item 3 had the highest pseudo-guessing parameter value, c = 0.36, and it was identified by CTT as being an easy test item, P = .81. Item 42 shared both an elevated *c*-value, c = 0.27 and was one of the two items with the lowest item discrimination values as identified by CTT, a = 1.01; R = 0.20. Item 47 was co-identified by CTT as sharing the lowest item discrimination value, a = 0.73; R = 0.20. IRT, however, identified Item 2 as having the lowest discrimination value, R = 0.21; a = 0.40. CTT co-identified Item 22, R = 0.56, a = 1.57, and Item 26, R = 0.56; a = 1.78 as having the highest discrimination values, whereas IRT similarly identified Item 26.

CTT summary statistics revealed that the GMRT-4 mean percent correct was 55.87 (range 1-48) and the mean percent correct for the ACT-R was 33.86 (range 1-40). Using CTT to calculate the mean percent correct scores indicated that the ACT-R was a significantly more difficult test of reading comprehension than was the GMRT-4. A visual inspection of the *b*-values generated by IRT for both assessments indicated that the ACT-R was a much more difficult test than was the GMRT-4.

Table 1

CTT Total Score Summary Statistics for ACT-R and GMRT-4

Test	Items	Cronbach's α	*M* (*SD*)
ACT-R	40	0.83	13.54 (6.73)
GMRT	48	0.91	26.82 (9.74)

Table 2

IRT Model Theta Estimates and Item Parameter Statistics for ACT-R and GMRT-4

		Theta Estimates		Item Parameter Estimates		
Test	Model	*n*	*M* (*SD*)	Items	Parameter	*M* (*SD*)
ACT-R	3-PL	423	0.04 (0.93)	40	a	0.98 (0.15)
					b	1.75 (0.53)
					c	0.24 (0.03)
GMRT	3-PL	423	-0.01 (0.95)	48	a	1.03 (0.30)
					b	0.30 (0.80)
					c	0.24 (0.02)

Table 3

CTT and IRT 3-PL Model Item Parameter Estimates for ACT-R

Item ID	*P*	*R*	*a*	*b*	*C*
1	0.48	0.30	1.24	1.24	**0.35**
2	**0.51**	0.29	0.78	0.96	**0.32**
3	0.43	0.31	1.17	1.56	**0.35**
4	0.39	0.30	0.86	1.40	0.24
5	0.42	0.36	0.87	1.07	0.23
6	0.32	0.30	0.86	1.80	0.23
7	0.34	0.34	0.86	1.58	0.23
8	0.37	0.24	**0.72**	1.74	0.24
9	0.41	0.27	**0.72**	1.39	0.25
10	0.36	0.29	0.95	1.64	0.25
11	0.20	0.18	1.00	**2.81**	0.21
12	0.44	0.44	1.12	**0.82**	0.23
13	0.35	0.27	0.80	1.69	0.23
14	0.34	**0.16**	0.76	2.17	0.25
15	0.42	0.39	0.95	1.02	0.23
16	0.26	0.26	1.02	2.25	0.22
17	0.35	0.29	0.83	1.67	0.23
18	0.43	0.32	0.84	1.14	0.24
19	0.36	0.41	1.08	1.16	0.21
20	0.35	0.27	0.81	1.68	0.23
21	0.41	0.41	0.97	1.06	0.23
22	0.29	0.30	1.10	1.97	0.23
23	0.31	0.26	0.98	2.08	0.24
24	0.44	0.35	0.91	1.04	0.25

Item ID	*P*	*R*	*a*	*b*	*C*
26	0.22	0.20	1.12	2.54	0.21
27	0.41	0.35	0.90	1.19	0.24
28	0.32	0.35	1.07	1.73	0.23
29	0.37	**0.45**	1.19	1.19	0.23
30	0.24	0.30	1.13	2.15	0.21
31	0.31	0.29	1.03	1.99	0.24
32	0.28	0.28	1.11	2.08	0.23
33	0.31	0.21	0.86	2.25	0.25
34	0.24	0.26	1.05	2.37	0.21
35	0.26	0.30	1.01	2.23	0.22
36	0.30	0.36	0.94	1.89	0.22
37	0.26	0.31	**1.22**	2.14	0.22
38	0.25	**0.16**	0.98	2.73	0.23
39	**0.19**	0.29	1.36	2.37	0.19
40	0.29	0.26	1.08	2.14	0.24

Note. Highest and lowest values are in boldface. *c*-values > 0.25 are in boldface.

Table 4

CTT and IRT 3-PL Model Item Parameter Estimates for GMRT-4

Item ID	*P*	*R*	*A*	*b*	*C*
1	**0.87**	0.26	0.60	**-1.90**	0.24
2	0.75	0.21	**0.40**	-1.16	0.24
3	0.81	0.34	0.72	-0.94	**0.36**
4	0.58	0.26	0.61	0.38	**0.26**
5	0.67	0.27	0.55	-0.28	0.25
6	0.50	0.21	0.62	0.96	**0.27**
7	0.61	0.22	0.50	0.11	0.25
8	0.64	0.44	0.92	-0.08	0.24
9	0.83	0.36	0.86	-1.19	0.25
10	0.74	0.48	1.08	-0.61	0.24
11	**0.24**	0.27	1.15	**1.84**	0.21
12	0.76	0.44	1.03	-0.67	0.25
13	0.75	0.37	0.75	-0.71	0.25
14	0.65	0.50	1.15	-0.12	0.24
15	0.67	0.49	1.05	-0.22	0.25
16	0.37	0.30	1.09	1.31	0.24
17	0.32	0.26	1.29	1.46	0.23
18	0.54	0.30	0.69	0.55	0.25
19	0.58	0.35	0.77	0.28	0.25
20	0.54	0.38	0.96	0.45	0.25
21	0.51	0.47	1.31	0.47	0.24
22	0.57	**0.56**	1.57	0.17	0.23
23	0.64	0.48	1.22	-0.04	0.25
24	0.43	0.31	1.01	1.05	0.25
25	0.61	0.47	1.01	0.00	0.24
26	0.79	**0.56**	**1.78**	-0.74	0.24
27	0.43	0.38	1.25	0.89	0.24

Item ID	*P*	*R*	*A*	*b*	*C*
28	0.65	0.52	1.12	-0.15	0.23
29	0.60	0.49	1.16	0.13	0.24
30	0.58	0.41	0.92	0.21	0.25
31	0.53	0.38	0.86	0.51	0.25
32	0.48	0.52	1.28	0.51	0.22
33	0.31	0.42	1.46	1.20	0.20
34	0.52	0.41	1.03	0.49	0.25
35	0.66	0.53	1.69	-0.06	**0.26**
36	0.30	0.33	0.94	1.56	0.21
37	0.53	0.50	1.24	0.35	0.23
38	0.64	0.55	1.31	-0.10	0.24
39	0.43	0.38	1.12	0.88	0.23
40	0.58	0.38	0.97	0.32	**0.26**
41	0.52	0.38	1.11	0.60	**0.26**
42	0.41	**0.20**	1.01	1.36	**0.27**
43	0.38	0.37	1.42	1.05	0.23
44	0.55	0.49	1.01	0.27	0.23
45	0.53	0.45	1.02	0.42	0.24
46	0.32	0.26	0.91	1.67	0.23
47	0.42	**0.20**	0.73	1.42	0.26
48	0.47	0.49	1.16	0.60	0.22

Note. Highest and lowest values are in boldface. *c*-values > 0.25 are in boldface.

Discussion

This study has provided a comparison of CTT and IRT psychometric theories using actual examinee data that demonstrate what simulation studies have previously shown: IRT provides much more useful and valuable item-level indices than does CTT (Hambleton & Jones, 1993). Several issues are evident in this comparison of psychometric theory. CTT's item-test correlation *R*-values provide reduced information compared to IRT's item discrimination *a*-values (Hambleton & Jones, 1993). In this analysis, CTT's *R*-values ranged from 0.20 to 0.56, compared to the greater range of IRT *a*-values of 0.40 to 1.78. The narrow range of *R*-values provides less opportunity for differentiation among items, clouding fine-grain analysis of item discrimination ability (Hambleton & Jones, 1993). Similarly, although there is a general correlation between CTT *P*-values and IRT *b*-values (Hambleton & Jones, 1993), CTT is not able to provide the same degree of measurement accuracy that IRT is able to provide for item difficulty (Hambleton & Jones, 1993). IRT has the capacity to estimate the pseudo-guessing parameter, useful for determination of less desirable items that might measure examinee guessing more than examinee ability. CTT does not have the ability to provide pseudo-guessing estimates.

The dichotomy between the measures of test difficulty illustrates a fundamental difference between CTT and IRT analytical methods. The reliance of CTT on examinee total scores precludes estimation of the statistical properties of administered items as a component of the estimation of examinee ability (de Ayala, 2009; Hambleton & Jones, 1993; Hambleton et al., 1991). IRT incorporates both examinee "true scores" and latent trait measurements, such as item difficulty and discrimination values, into estimates of examinee theta, independent of variation among samples (Hambleton et al., 1991). The *true scores* obtained through CTT analysis are population dependent, fluctuating with each sample and allowing variation in the measurement of examinee ability (Rathvon, 2004; Sharkness & DeAngelo, 2011).

Summary and Conclusions

The 1-PL, 2-PL, and 3-PL IRT models were evaluated using comparisons of -2LL values, and by visual and statistical comparisons of item-person maps, test response functions, and item response curves. Evaluation of these data determined that the 3-PL model was the best fit to the data for both the GMRT-4 and ACT-R. Comparison of CTT and the IRT 3-PL model for the GMRT-4 and ACT-R yielded approximately similar test indices and item difficulty measures. Although CTT and IRT item indices sometimes identified the same items at the extremes of the difficulty and discrimination indices, there were substantial estimation differences among items between the two extremes. Items falling between the highest and lowest item trait were not similarly ranked by the CTT and IRT analyses. Most importantly, IRT's 3-PL model produced significantly more accurate item-level test indices necessary for assessment refinement.

The conflicting results of these theoretical methods highlight the importance of selecting the most appropriate psychometric validation process for evaluating item and test characteristics. IRT has both theoretical and practical advantages over CTT. Although CTT produces reasonable item and test indices, IRT models provide falsifiable models with empirical data (Birnbaum, 1968; Hambleton et al., 1991; Lord, 1952). Unlike CTT, IRT provides invariant estimation for both item and person parameters (Hambleton et al. 1991). CTT produces information about students' total scores and standard deviation, but it is unable to produce the fine-grained item information that IRT produces. The more detailed item-level information produced by IRT analysis can be used to make decisions regarding item construction, test selection, and educational methods. IRT also has the capacity to select items suitable for students' individual ability levels through CAT (de Ayala, 2009; Hambleton & Jones, 1993; Hambleton et al., 1991).

Additionally, IRT analysis places students' latent abilities on the same continuum and aids the identification of assessment items that might be especially difficult, discriminate poorly among examinees, or demonstrate bias toward examinee sub-groups. Therefore, it is beneficial to examine student performance on the single, infinite scale used by IRT rather than introducing error by norming or vertically scaling total response scores using CTT.

This study examined the appropriateness and benefits of using IRT to measure reading comprehension by comparing IRT and CTT analytical methods. Simulation studies comparing IRT and CTT previously have shown that IRT offers many advantages for data analysis over CTT (Kim & Nicewander, 1993). However, studies analyzing and comparing these two psychometric theories on examinee responses of different reading comprehension assessments are relatively rare. Validation of simulation study findings using actual test data is a very important next step in the evolution of assessment development using increasingly advanced IRT models.

In the present study, analyses of item difficulty, item discrimination, and examinee ability estimates revealed that IRT provided more accurate information about item properties than did

CTT. The independent calibration of item and person latent traits provided by IRT analysis minimizes measurement errors and can be used to strengthen test designs. In so doing, IRT allows assessment developers to differentiate maximally among examinees within a targeted area of the continuum.

Despite the many advantages that IRT methods provide, reading researchers rarely employ the more sophisticated IRT models in the measurement of reading comprehension. Although IRT Rasch modeling software is readily available and relatively easy to use, more advanced IRT models necessitate the use of software that is costly, and which requires relatively greater knowledge of data management and conversion. IRT software programs need to be made more accessible and user friendly in order to achieve the widespread use of advanced IRT models in reading research. IRT data analysis is increasingly being used in diverse fields of scientific research (Thomas, 2011), but it is still not as widely understood as are CTT analytical methods. As demonstrated by the magnified item-level information produced by IRT in contrast to CTT in this study's comparison of ACT-R and GMRT-4, more studies employing IRT analysis are needed to demonstrate the many benefits of this psychometric methodology in the field of reading research.

References

ACT, Inc. (2006). *Reading between the lines: What the ACT reveals about college readiness in reading*. Iowa City, IA: Author. Retrieved from http://www.act.org/research/policymakers/pdf/reading_report.pdf

ACT, Inc. (2013). *Reading Test, ACT Practice Test*. Iowa City, IA: Author. Retrieved from http://www.act.org

Allen, J. (2012). Relationships of examinee pair characteristics and item response similarity. *ACT Research Report Series, 8*. Iowa City, IA: ACT, Inc. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED542023

Barnes, L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education. 4*, 143-157. doi:10.1207/s15324818ame0402_4

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

Chang, G., & Brown, J. (1983). The Minnesota Reading Assessment and the Gates-MacGinitie Reading Tests: Concurrent validity. *Journal of Studies in Technical Careers, 5*, 93-99.

Cooter, R. B., & Curry, S. (1989). Gates-MacGinitie Reading Tests, third edition. *Reading Teacher*, *43*, 256-258. Retrieved from http://www.jstor.org/stable/40030013

Cutting, L. E., & Scarborough, H. S. (2009). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*, 277-299. doi:10.1207/s1532799xssr1003_5

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Freeman, J. G., & Hutchinson, N. L. (1989). The concurrent validity of the Dimensions of Self-Concept (DOSC), Level E. *Educational and Psychological Measurement, 49*, 429-431. doi:10.1177/0013164489492015

Gall, M., Gall, J., & Borg, W. (2003). *Education research: An introduction*. Boston, MA: Pearson Education.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, *7*, 6-10. doi:10.1177/074193258600700104

Graham, D. M. (1990). Test review: Gates-MacGinitie reading tests. *Reading Improvement*, *27*(1), 21-23. Retrieved from https://eric.ed.gov/?id=EJ411521

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response* theory. Newberry Park, CA: Sage.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement. 6*, 373-378). doi:10.1177/014662168200600401

Jongsma, E. A. (1980). Test review: Gates-MacGinitie Reading Tests (second edition). *Journal of Reading*, *23*, 340-345. Retrieved from http://www.jstor.org/stable/40033215

Johns, J. L. (1984). Equivalence of forms 1 and 3, level E, Gates-MacGinitie Reading Tests. *Journal of Reading*, *28*, 48-51.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests.

Psychometrika, *58*, 587-599. doi:10.1007/BF02294829

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Lord, F. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN07.pdf

MacGinitie, W., MacGinitie, R., Maria, K., Dreyer, L., & Hughes, K. (2000). *Gates-MacGinitie Reading Tests (GMRT) fourth edition, forms S and T*. Itasca, IL: Riverside.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Education and Psychological Assessment*, *1*(1), 1-11.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2009). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading*, *9*, 85-116. doi:10.1207/s1532799xssr0902_1909021

Paris, S. G., & Stahl, S. A. (2005). *Children's reading comprehension and assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Park, B. J., Alonzo, J., & Tindal, G. (2011). *The development and technical adequacy of seventh-grade reading comprehension measures in a progress monitoring assessment system*. (Report #1102). Eugene, OR: Behavioral Research & Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED531667.pdf

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13-69). Mahwah, NJ: Lawrence Erlbaum Associates.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357-383. doi:10.1080/10888430701530730

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The Science of reading: A handbook* (pp. 227-247). Oxford, England: Blackwell.

Powell, W. R. (1969). Gates-MacGinitie Reading Tests. *Journal of Educational Measurement*, 6, 114-116.

RAND Reading Study Group. (2002). *Reading for understanding*. Santa Monica, CA: Rand Corporation.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York, NY: Guilford Press.

Sébille, V., Hardoin, J-B., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients—a simulation study. *BMC Medical Research Methodology*, *24*, 1-10. Retrieved from http://www.biomedicentral.com/1471-2288/10/24

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, *16*(1), 41-49. doi:10.3102/01623737016001041

Sharkness, J., & DeAngelo, L. (2011). Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. *Research of Higher Education*, 52, 480-507. doi:10.1007/s11162-010-9203-3

Thomas, M. L. (2010). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291-307. doi:10.1177/1073191110374797

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291-307. doi:10.1177/1073191110374797

Topczewski, A., Cui, Z., Woodruff, D., Chen, H., & Fang, Y. (2013). Comparison of four linear equating methods for the common-item equivalent groups design using simulation methods. *ACT Research Report Series*, *2*. Iowa City, IA: ACT, Inc. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED55559

Weiss, D. J., & Von Minden, S. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG*. Saint Paul, MN: Assessment Systems Corporation. Retrieved from https://www.assess.com/docs/Xcalibre_Bilog_Comparison_Report.pdf

Westrick, P., & Allen, J. (2014). Validity evidence for ACT Compass Placement Tests. *ACT Research Report Series, 2*. Iowa City, IA: ACT, Inc. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED546849

Woodruff, D., Traynor, A., & Cui, Z. (2013). A comparison of three methods for computing

scale score conditional standard errors of measurement. *ACT Research Report Series*, 7. Iowa City, IA: ACT, Inc. Retrieved from http://www.eric.ed.gov/contentdelivery/servlt/ERICServlet?accno=ED555593