Inconsistency in Reading Comprehension Assessment: Application of Item Response Theory (IRT)

Wang Ping

［Qilu Universities of Technology (Shandong Academy of Sciences) Jinan 250353］

Abstract: This study explored the possibility of improving classification results with item response theory (IRT) for reading comprehension assessment. A sample of 91 fourth graders completed three standardized tests and three researcher-created Maze tests. The coherence of different tests was investigated by comparing the top 20% and bottom 20% of children identified by each test following the procedure by Keenan and Meenan (2014). For the three standardized tests, total test scores based on classical test theory (CTT) were attained. For the three Maze tests, both CTT and IRT scores were obtained to compare IRT and CTT in their classification results. The coherence of CTT scores for the three standardized tests was 28% in the high performer group and 22% in the low performer group, which was about the same level as that of previous studies. For three Maze tests, the coherence of CTT scores was 33% and 28 % for both the high and low performer groups, respectively, while the coherence of IRT scores revealed 39% for both groups. IRT scores demonstrated the same or higher level of measurement invariance in all pairwise groups and no inferior outcomes compared to CTT. Advantages of utilizing IRT scores in reporting student reading comprehension performance and classifying elementary students' reading level were discussed.

Key words: CTT; IRT; Maze test; reading comprehension; assessment

CLC Number：G424; H314 **Document Code**：A **Article ID**：2832-9317（2023）04-0074-10

DOI：10.12424/HA.2023.069 **Link**：https://www.oc-press.com/HA-202304-074.html

1.Introduction

Since reading comprehension is a non-unitary construct, assessment of the construct is a crucial matter. Accurate and invariant assessment of reading comprehension is a fundamental element for progress monitoring and diagnosis of reading difficulties. Without consistent and valid measurement of the construct under study, no further manipulation of variables, prediction of related behaviors, or interventions will be meaningful. Although various models and theories have been proposed by different researchers in order to discern the complexity of reading comprehension and its assessment, these theories and models are not apparent in most standardized comprehension assessments currently used. These theoretical approaches to test development have caused many issues in the identification and measurement of reading comprehension difficulties such as disagreements in diagnosis and classification of students with different reading comprehension levels (Betjemann, Keenan,

Funding Project: Innovative Research Team for Comparative Studies of Chinese and Japanese Folklore in Shandong Universities 'Youth Entrepreneurship Team'(2022RW028) - Comparative Study of Intangible Cultural Heritage Systems between China and Japan.

Author: Wang Ping, Ph.D., School of Foreign Languages in Qilu Universities of Technology (Shandong Academy of Sciences), Lecturer. Research interests: literacy studies, language assessment.

Olson, & DeFries, 2011; Cain & Oakhill, 2006; Keenan & Meenan, 2014; Nation & Snowling, 1997).

Previous research has explored the consistency of reading scores between different reading comprehension tests. Rimrodt, Roberts, Denckla and Cutting (2005) compared three reading assessments (i.e., *Wechsler Individual Achievement Test Reading Comprehension Subtest, Gates-MacGinitie Reading Test, Gray Oral Reading Test*) and reported the average agreement rate between each pair of tests was 43.5% in diagnosing reading deficit (RD) and the consistency rate was only 25% among all three tests. Similarly, Keenan and Meenan (2014) used a larger sample (n = 995) to compare four commonly used reading comprehension tests for the diagnosis of student reading deficit: Gray Oral *Reading Test-3* (GORT-3; Wiederholt & Byant, 1992), *Qualitative Reading Inventory-3* (QRI-3; Leslie & Caldwell, 2011), the *Woodcock-Johnson Passage Comprehension-3* (WJPC-3; Woodcock, McGrew & Mather, 2001), and the *Peabody Individual Achievement Test* (PIAT; Dunn & Markwardt, 1970). Among the lowest 10% of the children (n = 100), the average diagnosis agreement rate between each pair of the four tests was 43%, which means that on average, students who were diagnosed as having reading difficulties by one test may have around 57% chance of not being identified by another test. There was an even less consistent result of diagnosis for the top 10% performers, in which 33% on average were classified consistently for the pairwise comparison. When the four tests were taken into account simultaneously, the consistency rate decreased to 20% for the lower performing group.

One factor for inconsistent assessment results could stem from the impact of psychological theories, as well as theories in reading comprehension on the development of reading comprehension assessments (Duke, 2005; Gough & Tunmer, 1986; Lincoln & Guba, 1985; Pearson & Hamm, 2005; Waston, 1913). For instance, behavioral schools of thought emphasized quantification and objectivity of assessments; therefore, group-administered standardized tests with multiple choice items were widely used (Watson, 1913; Pearson, 2000). The test outcome was usually interpreted as how well a student performed compared to others in a national sample of scores. However, the increased test scores may not automatically indicate an improvement of cognitive reading level. As cognitive psychology contended, the conceptual criteria such as prior knowledge and test structure should be given more weight than psychometric criteria in developing new tests. Therefore, longer passages, various question formats, and more difficult questions were introduced to reading comprehension tests (Pearson & Hamm, 2005). These developments in psychological theories have caused an increase of complexity for interpretation of reading test scores.

Meanwhile, various reading theories and models provided more detailed explanations of the reading process. For instance, the constructivist view of reading (Anderson & Pearson, 1984; Graesser, Singer, & Trabasso, 1994) accentuated the importance of prior knowledge, comprehension strategies, test structure, and metacognitive monitoring in reading, but Simple View of Reading (SVR, Gough & Tunmer, 1986) indicated reading performance could be predicted by two components: decoding and language comprehension. Many reading models also exerted an influence on reading skill assessment, such as construction-integration (CI) model (Kinstch & van Dijik, 1978, Bell & McCallum, 2008) and landscape model (van den Broek, 1990). Before the establishment of an articulated theory or model of reading comprehension, test developers constructed various reading comprehension tests (Sweet, 2005). Even worse, some pre-theory tests produced test scores which are not clearly related to the construct of reading comprehension, for instance, using IQ test scores to identify reading disabilities (Das, Mensink, & Mishra, 1990; Siegel, 1988; Tiu, Thompson, & Barbara, 2003).

Other factors may contribute to the assessment inconsistency issue as well. Cain and Oakhill (2006) reviewed the test item format (e.g., cloze tasks, true/false sentence recognition, sentence verification tasks, multiple choice, and open-ended questions) of some reading test items and attributed the inconsistency to task sensitivity, which here means that different item formats might tap to different reading skills. For instance, cloze tasks were more related to students' word reading skill level; true/false sentence recognition measured students' memory for literal details; multiple choice measured inferential processing and required higher processing demand than true/false items. However, according to Keenan, Betjemann and Olson (2008), cloze format may not be the only questioning type that caused the discrepancy in assessing decoding skills. Keenan et al. (2008) speculated that short passages were also related to decoding because the success of comprehension might rely on decoding of a single word and there was lack of adequate context information to support understanding. In addition, test format might also influence the reliability of test scores, such as sentence verification test which was found more reliable in estimating the reading proficiency of average or below average students (Marcotte, Rick, & Wells, 2019).

Moreover, many well-validated tests seem to measure different aspects of reading comprehension. Nation and Snowling (1997) used two reading comprehension tests (*Neale Analysis of Reading Ability* (NARA) and Suffolk test) and discovered that NARA was more dependent on listening comprehension, while Suffolk was more related to single-word reading ability. Cutting and Scarborough (2006) compared three reading comprehension tests: *Wechsler Individual Achievement Test* (WIAT; Wechsler, 1992), *Gates-MacGinitie Reading Test-Revised* (GMRT; MacGinitie, MacGinitie, Maria, & Dreyer, 2000) and GORT-3 (Wiederholt & Byant, 1992). Among the three tests, through a regression analysis, WIAT was found to be more influenced by students' decoding skill with its unique contribution of 11.9% of variance—much higher than that in GMRT (6.1%) and GORT (7.5%). The results indicated that students with higher decoding skill could perform better in WIAT than in GRMT and GORT. In addition, through a factor analysis, Keenan et al. (2008) found that PIAT and WJPC loaded more on decoding factor, while GORT and QRI loaded more on listening comprehension factor. Betjemann et al. (2011) confirmed the findings of Keenan et al. (2008) by using oblique rotation of Cholesky decomposition. They classified the five reading comprehension tests (WJPC, PIAT, GORT-3, QRI_Questions, and QRI-Passage Retell) into two categories: reading comprehension-decoding and reading comprehension-listening comprehension. These findings clearly evidenced that different tests were tapping on different reading components of reading comprehension, which may lead to inconsistency in diagnosis results.

Besides the above factors, another possibility may be one of the inherited shortcomings of the classical test theory (CTT) in psychometrics: test-dependent true score. A test-dependent true score is the first and inevitable problem of CTT (Francis, Fletcher, Catts, & Tomblin, 2005; Hambleton & Van der Linden, 1982). Depending on item difficulty and other item characteristics of various reading comprehension tests, the same person's true score on reading comprehension may vary (Keenan & Meenan, 2014). Item response theory (IRT) offers the possibility of test-independent estimation of person true score from different tests. Unlike CTT which focuses on total scores from all items, IRT focuses on item level information. According to the 3-parameter logistic model (3-PLM), for each item there are three parameters: item difficulty parameter (b-parameter) which influences the possibility of the item to be answered correctly, item-discrimination parameter (a-parameter) which is related to distinguishing students from different levels of a latent variable, and pseudo-guessing parameter, or lower asymptote parameter (c-parameter) which takes into consideration whether examinees give a correct response by guessing (Thomas, 2011). IRT can provide comprehensive analysis, reduce the measurement error, and improve the reliability of conventional tests, thus bringing greater accuracy for assessment in diagnosis and clinical practice (Kim & Nicewander, 1993; Rotou, Headrick & Elmore, 2002; Thomas, 2011).

If a student's latent true score is estimated through a stochastic model in IRT, the estimated student true score will be invariant, at least in theory, from different tests of reading comprehension except estimation bias. Thus, a common latent score for reading comprehension can be obtained through IRT, even though different tests are employed. The common latent estimate can be used to test, compare, and predict other related scores of the constructs. It is worthwhile to apply IRT scores to reading comprehension tests to discern how the consistency rate improves. In this study, there are two primary aims. The first objective is to investigate the consistency rate of three standardized tests and three Maze tests in classifying the reading ability of the current sample in order to assure the consistency level from the current project is not dramatically different from that of previous findings. The second objective is to verify whether using IRT scores help achieve higher consistency rate than CTT scores among three Maze tests.

2.Method

Participants

A total of 100 fourth graders were recruited from four rural schools in the Southeastern U.S. All fourth-grade students from 19 classrooms were invited to participate regardless of gender, ethnicity, disability, or intervention. Only those students who completed all tests were included in the sample, consequently nine students' test results were removed because of incompleteness in one or several tests. For all reading tests, ninety-one test results were obtained for analysis. Table 1 presents the demographic information of the participants.

Table 1

Demographic Information for the Total Sample (N= 91)	
	Percentage of Total
	n (%)
Gender (male)	49 (54%)
Ethnicity	
% Caucasian	64 (70%)
% African American	12 (13%)
% Hispanic	14 (15%)
% Asian	1 (1%)
Special Education Services	10 (11%)
Intellectually Gifted	2 (2%)
Other Health Impairment	3 (3%)
Specific Learning Disability	5 (5%)

Measurements

The current study used three standardized reading comprehension tests and three researcher-created Maze tests. Two standardized tests (WJPC-IV & WIAT-III) were administered individually, and GMRT-4 RC was administered in group. All three Maze tests were created by one of the authors. Each test had three passages chosen from www. newsela.org, and the passage complexity met the Lexile level (Lexile range 620 to 690) for fourth graders. These passages covered science, arts, health, opinion articles and news that were interesting to elementary school students. Three Maze tests were all administered in group and the inter-rater agreement for grading was 100%.

Woodcock–Johnson Tests of Achievement–4th edition Passage Comprehension subtest (WJPC-IV) is a norm-referenced test that measures a student's understanding of written text by cloze format. Students were asked to supply an appropriate word to sentences and short paragraphs. Test items were increasingly difficult in sentence length, vocabulary complexity, and topic familiarity. There were 52 items in total, but the number of items taken by students depended on their age and continued performance. The test stopped when students made five consecutive incorrect responses. Students' raw scores were calculated by subtracting all incorrected items from the ceiling item. The reported reliability index for students between 9 to 10-years-old was .89 (McGrew, Laforte, & Schrank, 2014).

Wechsler Individual Achievement Test- 3rd Edition Reading Comprehension (WIAT-III RC) is a validated and norm-referenced test. This subtest included two expository texts and one narrative text with six to eight questions per passage. After reading the passages aloud or silently, students were asked to verbally answer the questions read by the examiners. There was no time limit for this subtest. All of the 21 items were administered to the sample, and students' answers were recorded by the examiners. Following the guidelines of protocol, scores of 2, 1, and 0 were assigned to complete, partial, or incorrect answers, and the total score was the sum of all points. The average reliability index for fourth grade students was .85 (Breaux, 2010).

Gates-MacGinitie Reading Test-4th edition Reading Comprehension subtest (GMRT-4 RC), a norm-referenced and group administered test, is used to provide a general assessment of reading achievement ability for individual students. For fourth grade, the test consists of two sections: Vocabulary and Comprehension. Form S was used in this sample and the test had 48 items with a 35 minutes limit. Students were required to answer multiple-choice questions in paper form. The score was graded by 1 point with correct answer and 0 point for incorrect response. Total raw score was obtained by adding up all item scores. The reported test-retest reliability index ranges from .87 to .92 (MacGinitie, MacGinitie, Maria, & Dreyer, 2000).

Multiple-choice condition (Maze Test). Multiple choice Maze test contained three intact passages and each passage had 10 items composed of a sentence question and four choices. There were 30 items in total. Students were required to read the passages and choose the best answer from four choices. The reliability index of the multiple-choice condition test for the sample is .783.

Word-feature deletion condition. Word-feature deletion Maze test was constructed by deleting words such as pronouns, conjunction words, and content words (i.e. nouns) which were related to the central meaning of passages. The deleted word was replaced with one correct word and two distractors. This test contains 75 items, and students needed to understand the meaning of the previous sentences in order to get a correct answer. The reliability index was .953 for the sample in this study.

Sentence deletion condition. Sentence deletion Maze test was constructed by deleting an entire sentence from the passage, with the students needing to choose the one correct sentence from two distractor sentences. Student are required to choose the best option based on their comprehension. There were 30 items in total, and the reliability index for the sample is .887.

Procedure

Defining high performers and low performers. Different from Keenan and Meenan (2014) who used 10% as a cutoff point for high and low performers, we used the cut-off score of 20% due to a small sample size compared to the sample size from the Keenan and Meenan's (2014) study. Therefore, there were 18 students in both the high and the low performer groups for all reading tests. For the three standardized tests, only CTT scores were employed for classification because of the unavailability of item-level information. For the three Maze tests, both CTT and IRT scores were applied.

Data Analysis. Descriptive statistics of the raw scores from each test were computed, along with a correlation matrix for all possible pairs of test scores. Then the number of students who were simultaneously classified by different tests into the high performer (top 20%) or the low performer (bottom 20%) group was computed based on the test score from each test. For the Maze tests, both CTT and IRT scores were used separately for comparison between CTT and IRT. The number and percentage of cases which were commonly classified by all three tests (such as WIAT & WJ & GMRT) were counted first for both the high performer group and the low performer group, respectively. After this, the agreement cases and agreement percentages in each pair of tests were calculated (such as WIAT & WJ, WIAT & GMRT, and GMRT & WJ) in both groups as well. For the three Maze tests, the same procedure was repeated, but a comparison of agreement cases and percentages between CTT and IRT scores was made.

3.Results

Descriptive statistics and correlations. Table 2 shows different reading test scores as well as descriptive statistics for all tests. Bivariate correlation coefficients for three standardized tests and three Maze tests were reported in Table 3. It was noteworthy that all the correlations were statistically significant. The average correlation among three standardized tests is .564; in contrast the average correlation among three Maze tests is .699.

Table 2

Descriptive Statistics of Each Comprehension Test

Reading Tests	# of items	M	SD
GMRT-4	48	25.14	10.80
WJPC-IV	52	31.01	4.63
WIAT-III	21	28.20	7.34
MC	30	14.63	5.26
Word	75	55.04	15.13
Sentence	30	17.56	6.70

Note: GMRT-4= *Gates-MacGinitie Reading* Test-4th edition Reading Comprehension subtest Form S; WJPC-IV= *Woodcock–Johnson Tests of Achievement–4th edition* Passage Comprehension subtest; WIAT-III=*Wechsler Individual Achievement Test- 3rd Edition* Reading Comprehension.

MC: multiple-choice condition; Word: word-feature deletion condition; Sentence: sentence deletion condition.

Table 3

Correlations of Reading Comprehension Tests

	GMRT-4	WJPC-IV	WIAT-III	MC	Word	Sentence
GMRT-4	-					
WJPC-IV	.485**	-				
WIAT	.481**	.726**	-			
MC	.526**	.500**	.417**	-		
Word	.618**	.571**	.479**	.702**	-	
Sentence	.600**	.617**	.565**	.660**	.734**	-

Note: Correlations are significant *at p < .01**; GMRT-4: Gates-MacGinitie Reading Test-4th edition* Reading Comprehension subtest Form S; WJPC-IV: *Woodcock–Johnson Tests of Achievement–4th edition* Passage Comprehension subtest; WIAT-III: *Wechsler Individual Achievement Test- 3rd Edition Reading Comprehension.* MC: multiple-choice condition; Word: word-feature deletion condition; Sentence: sentence deletion condition

IRT score of Maze tests. For the three Maze tests, since the item-level information was available, IRT estimated person parameter (θ) score was calculated for each student through the Xcalibre software (Version 4.2). Three IRT models (1-parameter logistic model (1PLM), 2-parameter logistic model (2PLM), and 3-parameter logistic model (3PLM)) were applied to test the model-data fit using the Δχ2-test. For the multiple-choice Maze test, it was found that 2PLM was statistically better than 1PLM ($\Delta\chi^2$ (30) = 140.94, p < .05), but not different from 3PLM ($\Delta\chi^2$ (30) = -28.59, p > .05). The same result was observed from the sentence deletion ($\Delta\chi^2$ (30) = 99.68, p < .05 between 1PLM and 2PLM, and $\Delta\chi^2$ (30) = -17.61, p > .05 between 2PLM and 3PLM) and the word deletion Maze tests ($\Delta\chi^2$ (75) = 387.79, p < .05 between 1PLM and 2PLM and $\Delta\chi^2$ (75) =42.29, p > .05 between 2PLM and 3PLM). Thus, 2PLM was selected for IRT analysis in each of the three Maze tests based on the law of parsimony (Thorburn, 1915).

Agreement cases in standardized tests. The focus of the current study was to discern the consistency of test scores among various reading comprehension tests in assessing student's reading ability and in classifying students into different reading level groups. Every student's reading comprehension score was obtained from each test and the top 20% and the bottom 20% of all students were identified by reading comprehension scores from each test. The number of agreement cases in the top 20% and the bottom 20% of students were identified by three standardized tests. Table 4 showed that using the CTT total scores, the agreement rate of three standardized tests in high performer group was 28%, which means that 5 students (28% among 18 students) were simultaneously classified as in the high performer group by all three tests. For the bottom 20% of students, agreement rate was 22% (4 students), which means 22% of students were classified into the low performer group by all three tests. The consistency level from three standardized tests for the lower performance group was similar to the results (20%) of diagnosing students with comprehension deficits from four tests in Keenan and Meenan (2014). The average agreement rate across all pairwise test comparisons was 39% in the low performer group, meaning that a student who was diagnosed as a low performer by one test has only a 39% chance to be identified by another test. The average agreement rate from three standardized tests in the current study was slightly lower than that of 43% from four tests in Keenan and Meenan (2014). By contrast, in the current study, the average consistency rate in high performer group (56%) is higher than that of low performer group (39%), which is different from Keenan and Meenan's (2014) results (high performer group showed less consistency than the low performer group).

Table 4

Number of Overlapping Cases in Top 20% and Bottom 20% Students with CTT Score in Three Standardized Tests

Test Pair	Overlapping Cases	
	Top 20% (N=18)	Bottom 20% (N=18)
GMRT&WJPC&WIAT	5 (28%)	4 (22%)
WIAT & WJ	11 (61%)	11(61%)
WIAT& GMRT	8 (44%)	6 (33%)
WJPC & GMRT	11 (61%)	4 (22%)
Average pairwise	56%	39%

Note: GMRT-4= *Gates-MacGinitie Reading Test-4th edition* Reading Comprehension subtest Form S; WJPC-IV= *Woodcock–Johnson Tests of Achievement–4th edition* Passage Comprehension subtest; WIAT-III=*Wechsler Individual Achievement Test- 3rd Edition* Reading Comprehension.

Agreement cases in Maze tests. The second goal of the current study was to compare the consistency of classification of students using CTT and IRT scores from three Maze tests to detect the advantages of using IRT over CTT. The same procedure was adopted in three Maze tests but using both CTT scores and IRT scores, respectively. Table 5 presented the classification results from three Maze tests. When CTT scores were used, the consistency rate among three Maze tests was 33% in high performer group and 28% in the low performer group. When IRT scores from the 2PLM were used, the agreement rate was 39% in both the high performer group and the low performer group. For pairwise comparison, IRT demonstrated similar advantages in improving the consistency in low performer group. Even though such a pattern does not appear in all the pairwise comparisons, we found no cases where CTT outperformed IRT in classifications. On average, the agreement rate in pairwise test comparisons increased to 60% (increment of 2% from CTT) in the high performer group and increased to 58% (increment of 6% from CTT). It means when IRT scores were used, there was an improved consistency level of student classification with different tests scores compared to using CTT scores. The χ2-test revealed non-significant results between CTT and IRT classification outcomes, which could be attributed to a small sample size (n = 18 in each comparison group).

Table 5

Number of Overlapping Cases in Top 20% and Bottom 20% Students with CTT and IRT Score in Three Maze Tests

Test Pair	Overlapping cases			
	Top 20% (N=18)		Bottom 20% (N=18)	
	CTT	IRT	CTT	IRT
MC & Word & Sentence	6 (33%)	7 (39%)	5 (28%)	7 (39%)
MC & Sentence	10 (56%)	10 (56%)	10 (56%)	10 (56%)
MC & Word	9 (50%)	10 (56%)	8 (44%)	10 (56%)
Sentence & Word	12 (67%)	12 (67%)	10 (56%)	11 (61%)
Average Pairwise	58%	60%	52%	58%

Note: MC: multiple-choice condition; Word: word-feature deletion condition; Sentence: sentence deletion condition

4.Discussion

This study examined the issue of invariant assessment of reading comprehension in measuring student reading ability. The results demonstrated a higher consistency rate from IRT than CTT in Maze reading tests. With testing scores of three standardized reading tests and three Maze tests, this study adopted the methods used by Keenan and Meenan (2014) to compare the consistency level in categorizing students into different reading level groups by various reading comprehension tests. When CTT scores were adopted for three standardized tests, the consistency rate was similar to the findings of Keenan and Meenan (2014) in the diagnosis of low performing students, which confirms the existence of inconsistencies among reading comprehension assessments. However, consistency rate has a higher likelihood to be improved by using IRT scores compared to CTT scores in three Maze tests. Although the difference between the two types of scores was not statistically significant, which might be due to small sample size, the improving pattern of higher consistency by adopting IRT score is present, especially in classifying students in the low performer group. The increasing consistency pattern was more evident when all three Maze tests were included, which indicates the potential stability of IRT score.

It should also be noted that unlike the Keenan and Meenan's results (2014), in the current study, the high performer group exhibits higher consistency than the low performer group, which achieves a higher agreement level regardless of which tests (standardized tests and Maze tests) we used or which psychometric theories (CTT or IRT) we adopted. This phenomenon implies that, for the current sample, students with low reading level might perform with larger variability in different reading tests.

Researchers have raised the question of inconsistency in reading comprehension test results and explored the reasons with respect to test format, test construction and reading comprehension theories. In spite of multiple suggestions for how to address this issue from previous studies, there is still a lack of effective ways to improve assessment invariance among different reading tests. Theoretically, the invariance could be hard to achieve because of the innate test-dependent nature of the CTT-based reading comprehension test scores. The findings in the current study offer a possible solution from the psychometric perspective: enhanced consistency level in classification of students by the adoption of IRT score in reading assessments. IRT has resolved all theoretical and practical problems of CTT in terms of test dependency, sample dependency and impractical assumption for identical error of measurement (Hambleton & van der Linden, 1982; Kim & Nicewander, 1993). IRT is less biased than CTT in measurement error and has the potential to improve the accuracy of clinical assessment, even for conventional tests (Kim & Nicewander,1993; Thomas, 2011). Although we could not make a firm statement on the advantages of IRT scores over CTT scores, we found that IRT scores displayed the same or higher level consistency in all pairwise comparisons with CTT scores.

One of the limitations of this study is small sample size, which might be the reason that a significant difference between IRT scores and CTT scores was not obtained. Additionally, IRT scores were only available for the three Maze tests, thus the result is still uncertain when applying IRT to standardized tests. Nonetheless, the current study made a contribution to the reading comprehension field by providing an application of IRT scores to improve invariant assessment among different reading comprehension tests even some of those tests are not developed based on IRT. IRT sheds some light on the current issue and has the potential to propose a more accurate way to determine reading level placement. For future research, large sample size with item level information is suggested and more commonly used standardized tests could be examined. The results of this study may help educators evaluate students' reading proficiency with the outcome of various reading tests by utilizing IRT scores. Educational researchers may benefit from the outcome of this study for further exploration of the advantages of IRT over CTT.

References

[1] Anderson, Richard C., and P. David Pearson. 'A Schema-Theoretic View of Basic Processes in Reading Comprehension'. *Interactive Approaches to Second Language Reading*, edited by Patricia L. Carrell et al., Cambridge University Press, 1988, pp. 37–55.

[2] Bell, Sherry Mee, and R. Steve McCallum. *Handbook of Reading Assessment.* 2nd ed., Routledge, 2015.

[3] Betjemann, Rebecca S., et al. 'Choice of Reading Comprehension Test Influences the Outcomes of Genetic Analyses'. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, vol. 15, no. 4, Informa UK Limited, Jan. 2011, pp. 363–382.

[4] Breaux, K. C. *Wechsler Individual Achievement Test-3rd Edition (WIAT-III) Technical Manual with Adult Norms*. NCS Person, Inc, 2010.

[5] Cain, Kate, and Jane Oakhill. 'Assessment Matters: Issues in the Measurement of Reading Comprehension'. *The British Journal of Educational Psychology,* vol. 76, no. Pt 4, Wiley, Dec. 2006, pp. 697–708.

[6] Cutting, Laurie E., and Hollis S. Scarborough. '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: The Official Journal of the Society for the Scientific Study of Reading*, vol. 10, no. 3, Informa UK Limited, July 2006, pp. 277–299.

[7] Das, J. P., et al. 'Cognitive Processes Separating Good and Poor Readers When IQ Is Covaried'. *Learning and Individual Differences*, vol. 2, no. 4, Elsevier BV, Jan. 1990, pp. 423–436.

[8] Duke, N. K. 'Comprehension of What for What: Comprehension as a Nonunitary Construct'. *Children's Reading Comprehension and Assessment*, edited by S. G. A. Paris S, Lawrence Erlbaum Associates Publishers, 2005, pp. 93–104.

[9] Dunn, L. M., and F. C. Markwardt. *Examiner's Manual: Peabody Individual Achievement Test*. American Guidance Service, 1970.

[10] Francis, D. J., et al. 'Dimensions Affecting the Assessment of Reading Comprehension'. *Children's Reading Comprehension and Assessment*, edited by S. G. A. Paris S, Lawrence Erlbaum Associates Publishers, 2005, pp. 369–394.

[11] Gough, Philip B., and William E. Tunmer. 'Decoding, *Reading, and Reading Disability'. Remedial and Special Education: RASE*, vol. 7, no. 1, SAGE Publications, Jan. 1986, pp. 6–10.

[12] Graesser, A. C., et al. 'Constructing Inferences during Narrative Text Comprehension'. *Psychological Review*, vol. 101, no. 3, American Psychological Association (APA), July 1994, pp. 371–395, https://doi.org10.1037/0033-295x.101.3.371.

[13] Hambleton, Ronald K., and Wim J. van der Linden. 'Advances in Item Response Theory and Applications: An Introduction'. *Applied Psychological Measurement*, vol. 6, no. 4, SAGE Publications, Sept. 1982, pp. 373–378.

[14] Keenan, Janice M., et al. 'Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension'. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, vol. 12, no. 3, Informa UK Limited, July 2008, pp. 281–300.

[15] Keenan, Janice M., and Chelsea E. Meenan. 'Test Differences in Diagnosing Reading Comprehension Deficits'. *Journal of Learning Disabilities*, vol. 47, no. 2, SAGE Publications, Mar. 2014, pp. 125–135.

[16] Kim, J. K., & Nicewander, W. A. Ability estimation for conventional tests. *Psychometrika.* 1993, pp. 587–599.

[17] Kintsch, Walter, and Teun A. van Dijk. 'Toward a Model of Text Comprehension and Production'. *Psychological Review*, vol. 85, no. 5, American Psychological Association (APA), Sept. 1978, pp. 363–394.

[18] Leslie, L., and J. Galdwell. *Qualitative Reading Inventroy-3*. Addison Wesley Longman, 2001.

[19] Guba, Egon G., and Yvonna S. Lincoln. 'Epistemological and Methodological Bases of Naturalistic Inquiry'. *Evaluation Models*, Kluwer Academic Publishers, 2005, pp. 363–381.

[20] Macginitie, W. H., et al. *Gates-MacGinitie Reading Tests*. 2000.

[21] Magrew, K. S., et al. *Technical Manual. Woodcock-Johnson IV Tests of Achievement. Rolling Meadowns*. 2014.

[22] Marcotte, Amanda M., et al. 'Investigating the Reliability of the Sentence Verification Technique'. *International Journal of Testing*, vol. 19, no. 1, Informa UK Limited, Jan. 2019, pp. 74–95.

[23] Nation, K., and M. Snowling. 'Assessing Reading Difficulties: The Validity and Utility of Current Measures of Reading Skill'. *The British Journal of Educational Psychology*, vol. 67 (Pt 3), Sept. 1997, pp. 359–370.

[24] Pearson, P. D. *American Education: Yesterday, Today, and Tomorrow. Yearbook of the National Society for the Study of Education*. Edited by T. Good, University of Chicago Press, 2000, pp. 152–208.

[25] Pearson, P. D., and D. N. Hamm. 'The Assessment of Reading Comprehension: A Review of Practices-Past, Present, and Future'. *Children's Reading Comprehension and Assessment*, edited by S. G. A. Paris S, Lawrence Erlbaum Associates Publishers, 2005, pp. 13–69.

[26] Rimrodt, S., et al. *Are All Tests of Reading Comprehension the Same? Poster Presented at the Annual Meeting of the International Neuropsychological Society*. 2005.

[27] Rotou, Ourania, et al. 'A Proposed Number Correct Scoring Procedure Based on Classical True-Score Theory and Multidimensional Item Response Theory'. *International Journal of Testing*, vol. 2, no. 2, Informa UK Limited, June 2002, pp. 131–141.

[28] Siegel, L. S. 'Evidence That IQ Scores Are Irrelevant to the Definition and Analysis of Reading Disability'. *Canadian Journal of Psychology*, vol. 42, no. 2, American Psychological Association (APA), June 1988, pp. 201–215.

[29] Sweet, A. P. 'Assessment of Reading Comprehension: The RAND Reading Study Group Vision'. *Children's Reading Comprehension and Assessment*, edited by S. G. A. Paris S, Lawrence Erlbaum Associates Publishers, 2005, pp. 3–12.

[30] Thomas, M. L. The value of item response theory in clinical assessment: A review. *Assessment.* 2011, pp. 291–307.

[31] Thorburn, W. M. Occam's razor. *Mind.* 1915, pp.287-288.

[32] Tiu, Rolando D., Jr, et al. 'The Role of IQ in a Component Model of Reading'. *Journal of Learning Disabilities*, vol. 36, no. 5, SAGE Publications, Sept. 2003, pp. 424–436.

[33] Van Den Broek, P. 'The Causal Inference Maker: Towards a Process Model of Inference Generation in Text Comprehension'. *Comprehension Processes in Reading*, edited by D. A. Balota et al., Lawrence Erlbaum Associates, 1990, pp. 423–435.

[34] Watson, John B. 'Psychology as the Behaviorist Views It'. *Psychological Review*, vol. 101, no. 2, American Psychological Association (APA), Apr. 1994, pp. 248–253.

[35] Wechsler, D. L. *Manual for the Wechsler Intelligence Scale for Children-III*. Psychological Corporation, 1992.

[36] Wiederholt, L., and B. Byant. *Examiner's Manual: Gray Oral Reading Test-3*. 1992, p. PRO-ED.

[37] Woodcock, R. W., et al. *Woodcock-Johnson III Test of Achievement*. Riverside Publishing, 2001.