Research Article

ITEM ANALYSIS OF MULTIPLE-CHOICE READING LITERACY INSTRUMENTS USING ITEM RESPONSE THEORY

Received: August 4, 2021 **Revised:** June 11, 2021 **Accepted:** June 16, 2021

Yanika Lunrasri[1] Kamonwan Tangdhanakanond[2*] and Shotiga Pasiphol[3]

[1,2,3]Faculty of Education, Chulalongkorn University, Bangkok 10330, Thailand

*Corresponding Author, E-mail: tkamonwan@hotmail.com

Abstract

Reading literacy instruments were designed and validated to assess students' reading literacy performance. The purposes of this study were 1) to validate the overall model fit and item fit of the reading literacy instruments, 2) to analyze the item discrimination and item difficulty parameters of the instruments, and 3) to analyze the reliability coefficients of the instruments. There were a total of 277 Grade 9th students in this study. The instruments consisted of 20 multiple-choice pretest items and 20 multiple-choice posttest items. Five measurement item response theory (IRT) models were fitted and compared as follows: 1) the one-parameter logistic model, 2) the two-parameter logistic model, 3) the three-parameter logistic model, 4) the multidimensional item response theory model, and 5) the 2PL bifactor model. The 2PL bifactor model was found to be the most appropriate model for the data. There were only three misfit items in the model. A majority of the items had good discrimination values, except three items needed to be modified. The difficulty estimates were in the acceptable range. Moreover, the instruments yielded highly internal reliability.

Keywords: Item Analysis, Item Fit, Item Response Theory, Model Fit, Reading Literacy

Introduction

Reading literacy is the constructive process involving interaction between the reader and the texts that measures how students understand the text, interpret the meaning of the text, evaluate the text, and apply their reading abilities into their real-life situations (OECD, 2019a). It is one of the major three domains measured in the Programme for International Student Assessment (PISA), an international survey aimed to evaluate international educational systems (OECD, 2019a). The adolescents, 15-year-old students, are assessed for their scholastic performances every three years. The processing aspects of reading literacy has been categorized into 3 processes, including 1) the ability to locate information, 2) the ability to understand, and 3) the ability to evaluate and reflect (OECD, 2019a). The tested contents are not based on the basic curriculum or what is taught in the classroom, but the wide range of contents necessary for real-life situations are tested (The Institute for the Promotion of Teaching Science and Technology, 2018). Thailand has participated in the international assessment in order to assess and evaluate the quality of Thai educational system in accordance with the international criteria and standard (The Institute for the Promotion of Teaching Science and Technology, 2018). Reading literacy was the major tested domain in 2000, 2009, and 2018. Even though reading literacy has been promoted as one of the indicators for the quality of Thai education, the results of Thai students' reading literacy are not yet satisfactory (OECD, 2019b; The Institute for the Promotion of Teaching Science and Technology, 2020).

Existing research on reading literacy in Thailand has paid attention to experimental research and survey research. The experimental design aimed to develop an instructional model or some methods to improve reading literacy, such as learning activities (Chandai, 2016; Diowvilai et al., 2012). Another is survey research intended to investigate the varying factors affecting students' reading literacy abilities (Jumnaksarn, 2013; Nilsawang, 2011; Praputtakun et al., 2013). Given that reading literacy is the broad and complex performance, the tests comprised several testlets, which is a group of items sharing the same reading passages. The response in one item may be correlated with responses to other items within the same item grouping (Baghaei & Ravand, 2016; Debelak & Koller, 2020; Fox et al., 2020). Thus, the passage-based assessment is involved with trait variance and content-related variances of reading passages. In order to improve student's reading literacy performance, more accurate assessment tools to measure student's reading literacy should be constructed and more studies are needed to investigate the measurement issue. Item analysis is essential in improving the quality of test instruments. Item analysis allows the researchers to see the item characteristics and make sure that the items are appropriate to be included in a test or need improvement (Kim, 2017; Quaigrain & Arhin, 2017). As a result, there is a need to develop psychometrically sound assessments to validate the overall test and item-by-item analysis of the reading literacy instrument.

Research Objectives

The objectives of this research were as follows:

1. To validate the overall model fit and item fit of the reading literacy instruments
2. To analyze the item discrimination and item difficulty parameters of the reading literacy instruments
3. To analyze the reliability coefficients of the reading literacy instruments

Literature Review

Concept of Item response theory (IRT)

Item response theory (IRT) is the probability of a correct response on a test item as a function of the item characteristics and the ability levels of the test-takers. IRT estimates and interprets item statistics referred to parameters (Kanjanawasee, 2012).

Item parameter analysis

Item parameters are estimated using the data from students' responses to select the good items that have appropriate values. There are three item parameters as follows:

1) The a-parameter or item discrimination (slope) is the steepness of the item characteristic curve (ICC). A high discrimination value indicates that the item discriminates well between low-and high-level students (Baker,

2001). For the multidimensional IRT model, the multiple item discrimination estimates are combined into multidimensional item discrimination (MDISC) estimates (Cai & Kunnan, 2018) presented as:

$$\text{MDISC} = A_i = \frac{-d}{B_i}$$

2) The b-parameter or item difficulty is the location index that tells how easy or how difficult an item is. The negative item difficulty indicates that the item is easy, whereas a positive one shows the difficult item (Baker, 2001). For the multidimensional IRT model, the multiple item difficulty estimates are transformed into multidimensional item difficulty (MDIFF) estimates (Cai & Kunnan, 2018) presented as:

$$\text{MDIFF}=B_i = \frac{-d}{\sqrt{{a_0}^2+{a_s}^2}}$$

where d is the item intercept or item easiness, a_0 is the discrimination parameter on the general factor, and a_s is the discrimination parameter for the specific-related factors.

3) The c-parameter is known as a pseudo-guessing parameter to estimate the likelihood that an examinee with very low ability can guess the correct answer (Baker, 2001).

Item parameters in IRT are estimated and compared directly using three logical unidimensional models and two multidimensional models as follows: 1) the one-parameter logistic model (1PL) differs only in difficulty (b); the slopes (a) are constant and no guessing (c); 2) the two-parameter logistic model (2PL) shows that items are different in terms of difficulty (b) and slopes (a), without guessing (c); 3) the three-parameter logistic model (3PL) presents that items differ in terms of difficulty (b) and slopes (a), and pseudo-guessing (c); 4) multidimensional model allows each item to load only on specific dimension; and 5) bifactor model allows each item to load on the primary dimension and one specific dimension (DeMars, 2006, 2012).

Methodology

Participants

Grade 9 students at schools under the Secondary Educational Service Area Office 1 were recruited to participate in this study. There were a total of 277 participants from 6 schools. The two-stage random sampling was used to select the participants. The majority of the participants were female (60%) and 40% were male. Most of them were studying at extra-large schools (67%), followed by medium schools (17%) and large schools (16%). The participants were informed that their information was confidentially protected.

Research instruments

There were two tested instruments in this study as follows: 1) the reading literacy pretest and 2) the reading literacy posttest. The researchers studied the concepts and related documents related to reading literacy. The instruments were drawn upon the existing theoretical concept of reading literacy. The procedures of test construction were as follows: 1) review the literature on reading literacy, 2) develop the table of specification, 3) select reading passages, and 4) write test items (Creswell, 2012). The test content areas were based on three dimensions, namely 1) locate information, 2) understand, and 3) reflect and evaluate.

Both pretest and posttest comprised 20-item multiple-choice questions with one correct answer. In each test, the number of items were arranged in accordance with the PISA 2018 framework from OECD (2019a). The tests were written in Thai language. The reading passages were selected in accordance with the situational tasks (i.e., personal, public, occupational, and educational tasks). There were four reading passages in each test. The table of specification was shown in Table 1.

Table 1 Table of specification of reading literacy pretest and posttest

Aspect	Sub-aspect	Percentage	Pretest	Posttest
Locate	Access and retrieve information within a text	15%	3	3
information	search and select the relevant task	15%	3	3
Understand	represent literal meaning	20%	4	4
	integrate and generate inferences	20%	4	4
Evaluate and	assess quality and credibility	10%	2	2
reflect	reflect on content and form	10%	2	2
	detect and handle conflict	10%	2	2
	Total	**100**	**20**	**20**

After reading literacy pretest and posttest were constructed, the researchers asked a group of five experts to validate the reading literacy pretest and posttest. With respect to reading literacy pretest, the IOC index ranged from 0.6-1.0, except for two items (i.e., Item 3 and 10) that needed to be revised. In addition, the IOC index of the reading literacy posttest ranged from 0.6-1.0, except one item (i.e., Item 30).

Data Analysis

Data were analyzed in terms of IRT model and item fits as well as item parameter estimates using the "Mirt" package of the freeware R (Chalmers, 2012). For model fit and model comparison, five IRT models were fitted to the data and compared: 1) the one-parameter logistic model (1PL), 2) the two-parameter logistic model (2PL), 3) the three-parameter logistic model (3PL), 4) the multidimensional item response theory model (MIRT), and 5) the two-parameter bifactor model (2PL bifactor). The first three models assumed independence between item responses, whereas the other two focused on multidimensionality. Several statistical tests were applied, including the likelihood ratio test, Akaike information criterion (AIC), Bayesian information criterion (BIC), deviance statistic ($\boldsymbol{G^2}$), and RMSEA. The model with the smaller statistical tests was preferable (Bock & Aitkin, 1981; Gibbons & Hedeker, 1992).

At an item level, items of reading literacy pretest and posttest were evaluated using the item fit index, S-X^2 statistics. It was calculated whether each item fitted the model well (Orlando & Thissen, 2000, 2003, as cited in Desjardins & Bulut, 2018).

Moreover, the item parameter analysis was determined in terms of the item discrimination (a-parameter) and item difficulty (b-parameter). If the best fitted model was multidimensional model, the multidimensional item discrimination (MDISC) and the multidimensional item difficulty (MDIFF) were estimated (Cai & Kunnan, 2018).

According to Baker (2001), good discrimination values ranged from 0.35-1.69 and difficulty values should lower than 2.00 as presented in Table 2.

The measure of the consistency of the test results was determined by using Cronbach's alpha coefficient and Omega coefficient.

Table 2 Labels for item discrimination and item difficulty estimates

Discrimination values		Difficulty values	
Label	Range of values	Label	Ranges of values
Very low	0.01 - 0.34	Very easy	< -2.00
Low	0.35 - 0.64	Easy	-2.00 – -0.51
Moderate	0.65 - 1.34	Medium	-0.50 – 0.49
High	1.35 - 1.69	Difficult	0.50 – 1.99
Very high	≥ 1.70	Very difficult	≥ 2.00

Results

Evaluation of model and item fits of reading literacy instruments

1. Model-based measures of fit

Table 3 provides some of the fit information of reading literacy pretest and posttest. For the result of the reading literacy pretest, although the BIC suggested that the 2PL provided better fit to the data, the AIC indicated better fit for the bifactor model. Moreover, the smallest Log-likelihood, RMSEA, and G^2 values of reading literacy pretest obtained by 2PL bifactor model, representing a relatively good fit. Thus, the 2PL bifactor model was used to examine the psychometric properties of reading literacy pretest.

Similarly, with respect to the reading literacy posttest data, the 2PL and 2PL bifactor models fitted well with the reading literacy posttest. Although the 2PL received the smallest information criterion values, the 2PL bifactor model obtained the smallest Log-likelihood, RMSEA, and G^2 for reading literacy posttest. The multidimensional 2PL bi-factor model was preferred to examine the psychometric properties of the reading literacy posttest in order to avoid the strictly local independence assumption of the standard IRT models, resulting from the multidimensional concepts of the passage-based instruments.

Table 3 The overall model fit of reading literacy pretest and posttest

Test	Model	Log-likelihood	Akaike's Information Criterion (AIC)	Bayesian Information Criterion (BIC)	RMSEA	Deviance Statistic (G^2)
Pretest	1PL	-3341.09	6724.18	6800.36	0.062	3609.05
	2PL	-3277.49	6624.99	6770.10	0.036	3471.86
	3PL	-3261.96	6643.92	6861.57	0.040	3450.78
	MIRT	-3274.84	6641.69	6808.56	0.075	3476.55
	Bifactor	**-3247.90**	**6615.81**	**6833.47**	**0.025**	**3422.67**

Test	Model	Log-likelihood	Akaike's Information Criterion (AIC)	Bayesian Information Criterion (BIC)	RMSEA	Deviance Statistic (G^2)
Posttest	1PL	-3271.73	6585.47	6661.65	0.087	3476.02
	2PL	-3142.26	6364.53	6509.63	0.024	3217.08
	3PL	-3137.3	6394.6	6612.25	0.023	3207.15
	MIRT	-3192.26	6476.53	6643.40	0.104	3317.08
	Bifactor	**-3128.08**	**6376.17**	**6593.82**	**0.021**	**3188.71**

2. **Item-level diagnostics**

Item fit was used to evaluate the fitted items. According to the fitted 2PL bifactor model, the result found that pretest items fitted well to the model. Out of 20 items, only 1 item showed misfit (i.e., Item 17). Likewise, those posttest items were well fitted to the 2PL bifactor model. There were only 2 misfitting items out of 20 items (i.e., Items 27 and 32) as depicted in Table 4. Figure 1-3 presents the graphical ICC of the three misfitting items by presenting the expected model-based ICC and the empirical ICC for each item.

Table 4 The item fit of reading literacy pretest and posttest

Pretest				Posttest			
Item	S-X^2	*RMSEA*	*p*-value	Item	S-X^2	RMSEA	*p*-value
1	10.72	0.03	0.21	21	10.47	0.00	0.48
2	10.66	0.00	0.47	22	15.67	0.03	0.20
3	7.20	0.00	0.78	23	7.77	0.00	0.65
4	7.48	0.00	0.82	24	11.51	0.00	0.48
5	12.70	0.02	0.31	25	17.17	0.03	0.14
6	11.46	0.03	0.24	26	8.24	0.01	0.41
7	9.81	0.03	0.19	27	34.31	0.09	0.00*
8	18.40	0.03	0.10	28	3.06	0.00	0.96
9	11.31	0.01	0.41	29	16.06	0.05	0.06
10	17.11	0.03	0.14	30	10.77	0.00	0.54
11	3.73	0.00	0.88	31	13.31	0.03	0.20
12	10.98	0.01	0.35	32	16.12	0.06	0.04*
13	13.62	0.03	0.19	33	11.57	0.03	0.23
14	9.04	0.00	0.52	34	13.34	0.03	0.20
15	11.33	0.02	0.33	35	15.29	0.04	0.12
16	10.07	0.02	0.34	36	5.68	0.00	0.77
17	25.08	0.07	0.00*	37	17.36	0.05	0.06
18	13.64	0.03	0.19	38	3.86	0.00	0.92
19	9.60	0.00	0.47	39	11.05	0.00	0.43
20	18.90	0.04	0.09	40	9.56	0.00	0.65

Note. * *p*-value < 0.05

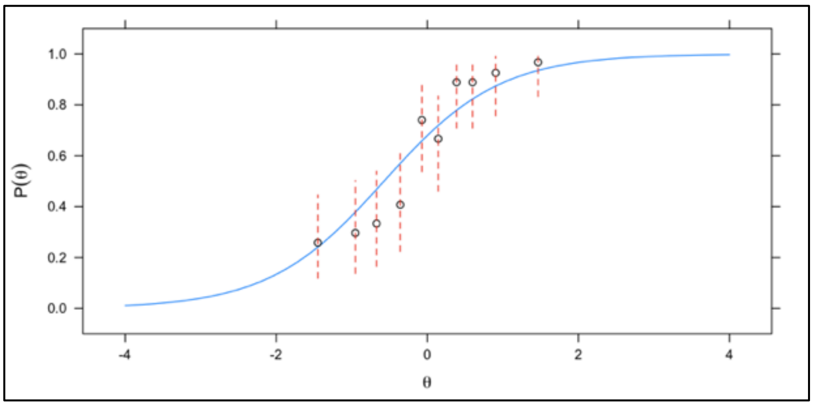

Figure 1 Empirical plot for item 17

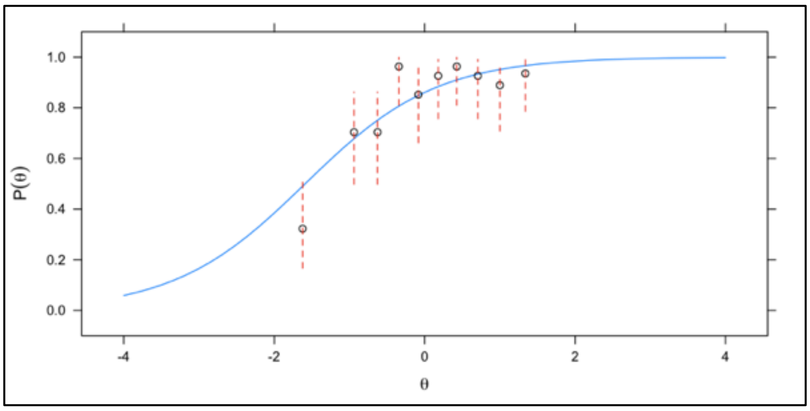

Figure 2 Empirical plot for item 27

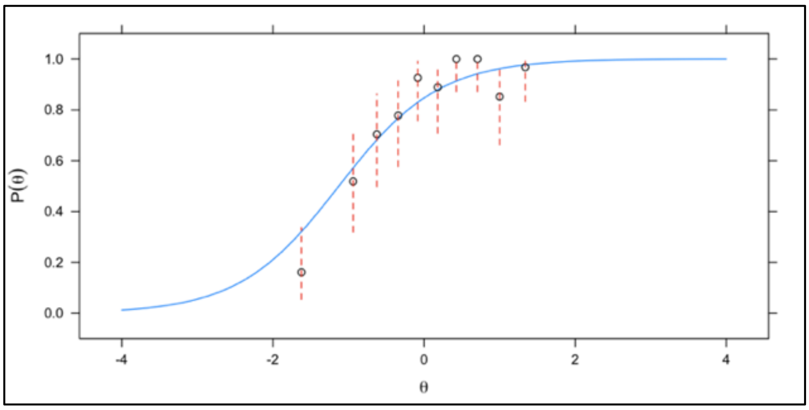

Figure 3 Empirical plot for item 32

Results of item parameter estimates

The responses of the pretest and posttest items were analyzed using item response theory (IRT) framework.

As depicted in Table 5, the responses of the pretest reading literacy items were analyzed. Most items obtained the larger discrimination parameter of the trait factor than the passage-based factors, except three items in the third passage (i.e., Items 13, 14, and 15). It indicates that most items assessed reading literacy performance than the background knowledge of the passages. For the multidimensional discrimination estimates, a majority of the pretest items had good MDISC, ranged from 0.37-3.59, except items 5 and 10. It was found that most items were at moderate level and five items were highly MDISC items (i.e., Items 1, 6, 7, 13, and 16). Only two items (i.e., Items 5 and 10) needed to be improved as their MDISC values were very low discriminating. Regarding the multidimensional difficulty estimates, all pretest items were produced in the acceptable range from -1.16 to 1.26, except for items 10 and 20 which had highly MDIFF values and needed to be improved.

Table 5 Item parameter estimates from 2PL bifactor model of reading literacy pretest

Passage	Pretest Item	Trait-a_1	Passage-a_2	Difference	MDISC$_i$	Discrimination level	MDIFF$_i$	Difficulty level
Passage 1	1	2.44	2.09	0.35	3.21	Very high	-0.97	Easy
	2	0.80	0.46	0.34	0.93	Moderate	-0.14	Medium
	3	0.83	0.46	0.37	0.95	Moderate	-0.23	Medium
	4	0.27	-0.53	0.80	0.60	Low	1.18	Difficult
	5	0.14	-0.02	0.16	***0.14***	***Very low***	0.40	Medium
Passage 2	6	1.69	0.49	1.20	1.76	Very high	-0.11	Medium
	7	2.98	1.99	0.99	3.59	Very high	-1.02	Easy
	8	0.49	0.33	0.16	0.60	Low	1.26	Difficult
	9	0.71	-0.43	1.14	0.83	Moderate	1.01	Difficult
	10	-0.03	-0.11	0.08	***0.12***	***Very low***	***10.08***	***Very difficult***
Passage 3	11	1.50	0.29	1.21	1.53	High	-1.16	Easy
	12	0.81	0.50	0.31	0.95	Moderate	-0.16	Medium
	13	1.10	1.52	**-0.42**	1.88	Very high	-0.11	Medium
	14	0.75	1.00	**-0.25**	1.25	Moderate	0.45	Medium
	15	0.85	0.99	**-0.14**	1.31	Moderate	0.30	Medium
Passage 4	16	1.75	1.51	0.24	2.31	Very high	0.26	Medium
	17	1.29	0.33	0.96	1.33	Moderate	-0.56	Easy
	18	1.26	0.91	0.35	1.55	High	0.57	Difficult
	19	0.83	0.58	0.25	1.01	Moderate	1.26	Difficult
	20	0.37	-0.02	0.39	0.37	Low	***2.63***	***Very difficult***

Notes. Trait-a_1 = the slope parameter estimates from the general factor; Passage-a_2 = the slope parameter estimates from specific factors; Difference = Trait-a_1 – Passage-a_2

As shown in Table 6, the responses of the posttest reading literacy items were analyzed using a 2PL Bifactor model. Most items obtained the larger discrimination parameter of the trait factor than the passage-based factors, except four items (i.e., Items 24, 25, 35, and 40). For the multidimensional discrimination estimates, a majority of the posttest items had good MDISC, ranged from 0.40 – 5.81. It was found that most items were at very high MDISC. None of them needed to be improved. Regarding the multidimensional difficulty estimates, all posttest items were produced in the acceptable range from -1.58 to 1.39, except for two items (i.e., Items 24 and 25) that had highly MDIFF values and needed to be improved.

Table 6 Item parameter estimates from 2PL bifactor model of reading literacy posttest

Passage	Posttest Item	Trait-a_1	Passage-a_2	Difference	$MDISC_i$	Discrimination level	$MDIFF_i$	Difficulty level
Passage 5	21	1.12	0.38	0.74	1.19	Moderate	-0.08	Medium
	22	-0.52	-1.40	0.88	1.49	High	1.39	Difficult
	23	1.80	-0.32	2.12	1.83	Very high	-0.72	Easy
	24	0.26	0.30	**-0.04**	0.40	Low	***2.08***	***Very difficult***
	25	0.43	0.48	**-0.05**	0.64	Low	***2.03***	***Very difficult***
Passage 6	26	2.23	0.24	1.99	2.24	Very high	-0.90	Easy
	27	1.14	-0.15	1.29	1.15	Moderate	-1.58	Easy
	28	1.53	0.41	1.12	1.59	High	-0.13	Medium
	29	1.75	0.20	1.55	1.76	Very high	-0.22	Medium
	30	0.26	-3.94	4.20	3.95	Very high	0.54	Difficult
Passage 7	31	1.05	0.12	0.93	1.05	Moderate	-0.50	Medium
	32	1.55	0.15	1.40	1.55	High	-1.11	Easy
	33	1.52	0.50	1.02	1.60	High	-0.20	Medium
	34	0.96	0.39	0.57	1.03	Moderate	0.52	Difficult
	35	1.67	2.46	**-0.79**	2.97	Very high	0.30	Medium
Passage 8	36	2.53	0.27	2.26	2.55	Very high	-0.32	Medium
	37	0.98	0.20	0.78	1.00	Moderate	-0.43	Medium
	38	4.29	3.92	0.37	5.81	Very high	-0.17	Medium
	39	0.36	-0.40	0.76	0.54	Low	1.02	Difficult
	40	0.20	0.58	**-0.38**	0.61	Low	0.78	Difficult

Notes. Trait-a_1 = the slope parameter estimates from the general factor; Passage-a_2 = the slope parameter estimates from specific factors; Difference = Trait-a_1 – Passage-a_2

Reliability of reading literacy pretest and posttest

Internal consistency using Cronbach's alpha coefficient and Omega coefficient were calculated for reading literacy pretest and posttest as shown in Table 7. The Cronbach's alpha reliability estimates for pretest and posttest were 0.72 and 0.77, respectively. For the Omega coefficient, the reliability estimates of pretest and posttest were 0.76 and 0.80, respectively. In order to be acceptable, the value of reliability should be more than 0.7. This indicated that the highly acceptable reliability values for the overall reading literacy pretest and posttest, which yielded highly internal reliability.

Table 7 Reliability of reading literacy pretest and posttest

Pretest		Posttest	
Cronbach's alpha	Omega coefficient (ω)	Cronbach's alpha	Omega coefficient (ω)
0.72	0.76	0.77	0.80

Discussion

Results of model and item fits to the reading literacy data

There was strong evidence that the bifactor model was the most adequate model for the reading literacy data. The model provided better understanding about the dimensionality of the tests because some items were grouped together in a form of reading passage called 'testlet' (DeMars, 2012). In each test, a domain-general factor represented the overall reading literacy and four domain-specific factors contained the particular passage-based knowledge. Previous studies have confirmed the advantages of the bifactor model for passage-based reading assessment (Byun & Lee, 2016; DeMars, 2006, 2012; Kim, 2017; Min & He, 2014; Sabbag & Zieffler, 2015). The model was assumed to accommodate the variability of the testlet effect within passage (Byun & Lee, 2016). Excluding the passage-related factors might lead to information loss (Cai & Kunnan, 2018).

In examining the item fit, the bifactor model had only one misfitting item for pretest and two misfitting items for posttest. As shown in the ICC plots (in Figure 1-3), the closer the empirical data follow the ICC, the better item fit (Desjardins & Bulut, 2018). For the misfitting items, the probability of answering items correctly in the empirical plot were lower for low ability students (θ = -1.8 to -0.2) and higher for high ability students (θ = 0 to 1.8) than those in the model-expected ICC. As a result, they were not performing well and needed to be modified. However, in terms of discrimination parameter estimate, they still discriminated moderately among students.

Results of Item analysis of reading literacy instruments

The findings revealed that most individual items obtained the larger discrimination parameters of the general factor than that those of the specific dimensions. This suggests that the test measured what it intended to measure, which was reading literacy performance (Byun & Lee, 2016). This is consistent with Byun and Lee (2016) that slope parameters of the individual items of the general dimension were higher than those of the testlet dimension. It is also interesting to point out some items received stronger influence from the passage factors more than the general factor (i.e., Items 13, 14, and 15). These items were grouped in the same passage. The reason for the passage-related influence might be because they had prior knowledge related to the reading passage.

The results showed that five items on pretest (i.e., Items 1, 6, 7, 13, and 16) and seven items on posttest (i.e., Items 23, 26, 29, 30, 35, 36, and 38) had great power for discriminating a student's reading literacy ability. Items 5 and 10 may not discriminate between low and high ability students. Item 5 addressed the student's ability in assessing the quality and credibility of the reading text. Item 10 addressed student's ability in detecting and handling conflict. Student was requested to read the additional text that contradicted the main reading text in task two, and then he/she was asked to figure out the conflicting issue represented in the additional text. Both items were grouped in the evaluate and reflect subscale. This indicates that it might be complex and more difficult to discriminate student's ability in evaluating and reflecting the reading texts. The possible explanation might be related to the wording of the item. The question might lead to wrong interpretations of what the question was conveying (Sabbag & Zieffler, 2015). Thus, revision in terms of content and grammatical structure are needed. Regarding item difficulty estimates, some items had very high difficulties (i.e., Items 10, 20, 24, and 25). In addition to having a very low discrimination value, Item 10 was also the most difficult item. It is noteworthy that all difficult items measured a student's ability in evaluating and reflecting the text. Thus, they needed to be reconsidered and improved.

Conclusion and implications for future research

The aim of this research was to investigate the item analysis of the reading literacy instruments. Several measurement models were examined to determine which was appropriate to model the data. It demonstrated the importance of the bifactor model for reading literacy assessment at the item and model level. Item-based information was provided for selecting appropriate items for reading literacy tests. Items with low discrimination or items with very high difficulty might be revised to improve the quality of the instruments.

Some strengths of this research should be highlighted. This research provides the empirical evidence supporting the validity of the instruments for measuring reading literacy. Also, this study demonstrated detailed procedures for evaluating model and item fits as well as item analysis by presenting all the competitive models and evaluating the psychometric properties of the reading literacy items. Following implications for future research, one of the limitations is related to the number of reading passages. Each reading consisted of four reading passages which may lack the variation of contents. Thus, a variety of passages are needed to lead to more precise information of the item estimates. Moreover, future studies can be designed to examine how factors, such as sample size, may affect the accuracy of item analysis of the instruments.

References

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension tests using testlet response theory. *Psicologica,* 37, 85-104.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika,* 46(4), 443-459.

Byun, J. H., & Lee, Y. W. (2016). The latent trait modeling of passage-based reading comprehension test: Testlet-based MIRT approach. *English Language Assessment,* 11, 25-45.

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approch combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly,* 15(2), 109-129.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software,* 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chandai, S. (2016). *Development of an instructional model based on scaffolded reading experiences approach and self-regulated learning for enhance reading literacy of lower secondary school students* (Doctoral dissertation). Bangkok: Chulalongkorn University.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Debelak, R., & Koller, I. (2020). Testing the local independence assumption of the Rasch model with Q3-based nonparametric model tests. *Applied Psychological Measurement,* 44(2), 103-117.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement,* 43(2), 145-168.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement,* 36(2), 104-121.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. Boca Raton, FL: CRC Press.

Diowvilai, D., Samranjai, J., Sommano, B., Janwong, V., & Mookham, T. (2012). *Development of elementary children Grade 4-6 with reading literacy through Lampang learning enrichment network* (Research report). Lampang: Lampang Rajabhat University.

Fox, J.-P., Wenzel, J., & Klotzke, K. (2020). The Bayesian covariance structure model for testlets. *Journal of Education and Behavioral Statisitcs,* 46(2*),* 219-243.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika,* 57, 423-436.

Jumnaksarn, S. (2013). The development of causal relationship model of factors influencing on reading literacy of 15-year-old students in Thailand. *Journal of education Research, Faculty of Education, Srinakharinwirot University,* 8(2), 213-230.

Kanjanawasee, S. (2012). *Modern test theory* (4th ed.). Chulalongkorn University Printery.

Kim, W. H. (2017). *Application of the IRT and TRT models to a reading comprehension test* (Doctoral dissertation). Tennessee: Middle Tennessee State University.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing,* 31(4), 453-477.

Nilsawang, P. (2011). *Factors related to reading literacy of grade 9 students under primary education service area office in Sisaket province* (Master thesis). Mahasarakham: Mahasarakham University.

OECD. (2019a). *PISA 2018 assessment and analytical framework*. PISA, OECD Publishing.

OECD. (2019b). *Thailand-Country Note - PISA 2018 Results*. OECD Publishing.

Praputtakun, P., Dahsah, C., Tambunchong, C., & Mateapinikul, P. (2013). The case study of prethomsuksa 6 students' scientific literacy and reading ability. *Journal of Education Thaksin University,* 13, 127-140.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education,* 4(1). DOI: 10.1080/2331186X.2017.1301013

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the goals-2 instrument. *Statistics Education Research Journal,* 14(2), 93-116.

The Institute for the Promotion of Teaching Science and Technology. (2018). *PISA 2015 results in science, reading, and mathematics: Excellence and equity in education.* Bangkok: Success Publication.

The Institute for the Promotion of Teaching Science and Technology. (2020). *PISA 2018 results: What student know and can do*. Retrieved from https://pisathailand.ipst.ac.th/issue-2019-48/