Kasetsart Journal of Social Sciences 44 (2023) 257–266

Kasetsart Journal of Social Sciences

journal homepage: http://kjss.kasetsart.org

Analysis of parallel reading literacy instruments: Application of item response theory to statistical parallelism

Yanika Lunrasri, Kamonwan Tangdhanakanond*, Shotiga Pasiphol

Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Pathumwan, Bangkok 10330, Thailand

Article Info

Article history:
Received 19 February 2022
Revised 9 April 2022
Accepted 15 April 2022
Available online 15 March 2023

Keywords:
item response theory,
parallel form,
statistical parallelism,
test characteristic curve,
test information function

Abstract

This research aimed to: (1) investigate the statistical parallelism of reading literacy instruments at the level of the overall test sets; and (2) examine whether the characteristic curves and information functions as the statistical item response theory (IRT) methods for parallelism yield the comparable scores at the levels of test forms as well as individual items. The population was ninth graders in secondary schools under the Secondary Education Service Area Office 1. Samples were examined in a total of 525 students. The instruments were three parallel multiple-choice test sets designed to elicit reading literacy skills. Each test set consisted of two 20-multiple choice test forms. Parallelism was analyzed with respect to means, standard deviations, reliabilities, item parameter estimates, characteristic curves, and information functions. It was found that there was statistical parallelism among the three test sets in terms of means, standard deviations, reliabilities, and item parameter estimates. Cronbach's alpha coefficients were .81, .84, and .83, respectively. The one-way analysis of variance results presented similar mean scores ($F = 0.325$, $p = .723$), equal level of discrimination ($F = 0.194$, $p = .824$), and equal level of difficulty ($F = 0.761$, $p = .470$). Moreover, both characteristic curves and test information functions yielded similar degrees of parallelism at the test form and item-by-item levels. Interesting findings with respect to statistical parallelism are presented and discussed.

© 2023 Kasetsart University.

Introduction

Reading literacy is required as the foundation of other subject areas in the educational system. It is also a necessary skill needed for being successful as global citizenship (Organisation for Economic co-operation and Development [OECD], 2019a). Reading literacy measures student's reading ability in understanding the text, interpreting the meaning of the text, and evaluating the text in real-life situations (OECD, 2019a; The Institute for the Promotion of Teaching Science and Technology [IPST], 2020). It has been categorized as: (1) locate information; (2) understand; and (3) evaluate and reflect (OECD, 2019a). The Office of the Basic Education

* Corresponding author.
E-mail address: tkamonwan@hotmail.com (K. Tangdhanakanond).

https://doi.org/10.34044/j.kjss.2023.44.1.29
2452–3151/© 2023 Kasetsart University.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Commission has encouraged teachers and educators to promote reading literacy as stipulated in the National Education Plan of Thailand (2017–2036) (Office of the Education Council, 2017). However, it is still not sufficient to improve student's reading literacy performance. The evidence from the *Programme for International Student Assessment (PISA) in* 2015 and 2018 showed that the average reading literacy results of Thai students decreased steadily (Organisation for Economic Co-operation and Development [OECD], 2016, 2019b; IPST, 2020). More than half of Thai test-takers were defined as low performers in reading literacy (OECD, 2019b; IPST, 2020), resulting in the difficulty in solving complex problems that are faced in modern societies (OECD, 2019a).

Test parallelism is regarded as the similarities in terms of content-related and statistical properties (Ali & van Rijn, 2016; Debeer et al., 2017; Van Der Linden & Adema, 1998). In that sense, the two forms to be called parallel must be equivalent to meet specifications of test blueprints and have a specific type of statistical equivalence in order that the test scores are comparable (Bae & Lee, 2011; Hambleton & Jones, 1993; Lin, 2008). In accordance with the classical test theory (CTT), the parallel test forms refer to the tests that have equivalence in terms of equal means and standard deviations, factor structure, correlation, and reliability (Bae & Lee, 2011; Clause et al., 1998; Schmucker et al., 2019).

In the context of item response theory (IRT), parallelism is the case when test forms have items with equal parameters in IRT models (Debeer et al., 2017). Item parameters in IRT are estimated using three unidimensional models (i.e. one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model) or multidimensional models (Baker, 2001). This study applied the two-parameter logistic model for calibrating the item parameters in statistical parallelism. The statistical specifications under IRT approach are stated in terms of the characteristic curves and information functions (Ali & van Rijn, 2016; Lin, 2008). Although there is still no strict agreement in IRT literature on parallelism, the test forms are parallel in the sense that the characteristic curves and the information functions are identical (Debeer et al., 2017; Lin, 2008).

CTT is associated with some limitations such as sample dependency and test dependency (De Champlain, 2010; Hambleton & Jones, 1993). On the contrary, the advantages of the of IRT are invariance properties, a person parameter and an item parameter. IRT can be estimated from different samples onto a common scale, and test scores of different test forms are comparable to each other (Davey & Hendrickson, 2010; De Champlain, 2010; Hambleton & Jones, 1993)

The use of parallel forms for repeated measure are generally preferred for intervention in classroom as one of the criteria for outcome measure (Petersen et al., 2020). The key advantages of the properly developed parallel forms result in fairness and security of a test (Bae & Lee, 2011; Clause et al., 1998; Lin, 2008). Moreover, they allow the direct comparisons, both individual and group levels, to track progress in learning and prevent the uncontrollable factors or threats to internal validity such as score increases from practice, previous exposure and memorization (Bae & Lee, 2011; Furtkamp et al., 1982; Morin et al., 2017; Raymond et al., 2007). In order to ensure the comparability of test scores on different forms, this paper described the construction of the parallel reading literacy tests and implemented several methods of test parallelism. The parallel instruments are required to improve the student's reading literacy skills and used as interchangeably alternate forms in the reading literacy assessment to monitor progress or capture changes in students' reading performances.

This area of investigation has received less attention by scholars in assessment. The majority of empirical research focused on parallelism in the sense of test construct and test score equivalence using CTT. Clause et al. (1998) studied the equivalence of parallel biodata forms and situational judgment forms using means, standard deviations, reliabilities, correlations, and factor analysis. Bae and Lee (2011) investigated the parallel picture forms in terms of the means, variances, and factor structures. Fitzpatrick and Clenton (2010) determined test equivalence for parallel vocabulary tests using means, variance and correlation. The study of Morin et al. (2017) determined the parallelism of the Music Self-Perception Inventory (MUSPI) using factor analysis, composite reliability, and convergent validity. The study by Quansah and Cobbinah (2021) developed parallel testing in a Basic Statistics course by comparing means, covariances, variances, and validity.

With the recent advances in IRT, there is an interest in the empirical research focusing on parallelism of test score equivalence using IRT. Moreover, there has been little empirical research focusing on comparing between characteristic curves and information functions for test parallelism. It is therefore appropriate to stress the necessity of using and verifying parallel test forms using IRT approaches in reading literacy research in this paper. The research objectives were as the following: (1) To investigate the statistical parallelism of the reading

literacy instruments at the level of the overall test sets; and (2) To determine whether the characteristic curves and information functions yield the comparable scores at the levels of test forms as well as individual items.

Literature Review

Parallel Literature

In establishing the parallelism of instruments, the original and alternate tests are compared in terms of content parallelism and statistical parallelism (Clause et al., 1998).

Content parallelism

Contents of parallel test forms are evaluated during pre-administration by a group of scholars and experts to check the parallelism of the tested contents. They are asked to evaluate their judgments and offer their opinions on the equivalence of the test items of parallel forms (Clause et al., 1998).

Statistical parallelism

In accordance with IRT, the characteristic curves and information functions are both used as criterion for statistical-related parallelism. Tests are defined to be weakly parallel if they measure the same trait and have the same information functions in accordance with Samejima (1977). Tests are strongly parallel if they have the same test length and the same characteristic functions as defined by McDonald (1999 as cited in Lin, 2008).

1. Information functions. Information functions are the amount of information providing the measurement precision at the specific ability level for either an individual item or the entire test (Ackerman, 1989). The test information function (TIF) is the sum of the item information functions (IIFs), which is the amount of information of a single item computed at any ability level (Baker, 2001; Boekkooi-Timminga, 1990). Equation (1) of TIF and Equation (2) of IIF are:

$$I(\theta) = \Sigma_{i=1}^{n} I_i(\theta) \quad (1)$$

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta) \quad (2)$$

where $I(\theta)$ is the amount of test information at an ability level of θ, $I_i(\theta)$ is the amount of information for item i at ability level θ, n is the number of items in the test, θ is the ability level, $P_i(\theta) = 1/(1+EXP(-a_i(\theta-b_i)))$, and $Q_i(\theta) = 1-P_i(\theta)$

The major index of statistical parallelism used in this study was a root mean square deviation (RMSD) statistic (Debeer et al., 2017). The RMSD is the average difference between the estimated values of the assembled test form and the reference test form. Equation (3) of $RMSD_{TIF}$ and Equation (4) of $RMSD_{IIF}$ are presented as follows (Debeer et al., 2017):

$$RMSD_{TIF} = \sqrt{\frac{\sum_{j=1}^{q}[I(\theta j)-I'(\theta j)]^2}{q}} \quad (3)$$

$$RMSD_{IIF} = \sqrt{\frac{\sum_{j=1}^{q}[i(\theta j)-i'(\theta j)]^2}{q}} \quad (4)$$

where q is the number of ability levels over which the RMSD is computed, $I(\theta j)$ and $I'(\theta j)$ are the TIFs at a particular θ level of the assembled test and the reference test, and $i(\theta j)$ and $i'(\theta j)$ are the IIFs at a particular θ level of the assembled test and the reference test.

2. Characteristic curves. Characteristic curves can be used to predict observed scores of examinees at given ability levels (Baker, 2001). Item characteristic curve (ICC) is the relationship between the probability of answering correctly to an item and ability level (Baker, 2001). Test characteristic curve (TCC) is the sum of the item response functions (Ali & van Rijn, 2016). Equation (5) of the TCC and Equation (6) of the ICC for two-parameter logistic model are:

$$T(\theta) = \Sigma_{i=1}^{n} P_i(\theta) \quad (5)$$

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \quad (6)$$

where $T(\theta)$ is the true score for respondents with ability level θ, $P_i(\theta)$ is the likelihood of a respondent correctly answering item i given the respondent's ability (θ), b_i is the item's difficulty, a_i is the item's discrimination, and D represented a scaling factor.

Equation (7) of TCC-parallel and Equation (8) of ICC-parallel are presented as follows (Ali & van Rijn, 2016):

$$RMSD_{TCC} = \sqrt{\frac{\sum_{j=1}^{q}[T(\theta j)-T'(\theta j)]^2}{q}} \quad (7)$$

$$RMSD_{ICC} = \sqrt{\frac{\sum_{j=1}^{q}[t(\theta j)-t'(\theta j)]^2}{q}} \quad (8)$$

where q is the number of ability levels over which the RMSD is computed, T(θj) and T'(θj) are the expected test scores of the assembled test and the reference test, and t(θj) and t'(θj) are the expected item scores of the assembled test and the reference test.

Methodology

Population and Sample

The population of this study was ninth graders in schools under the Secondary Education Service Area Office 1. Samples were 525 ninth graders of six secondary schools in Bangkok. They were selected by using a multi-stage random sampling procedure. In the first stage, the researchers used stratified random sampling to select schools under the Secondary Education Service Area Office 1 regarding their school sizes. Ninth-grade classrooms were randomly selected in the second stage. In addition, students in each classroom were randomly allocated into sub-groups and were assigned to take the reading literacy test sets differently. There were three sub-groups taking reading literacy test sets as follows: Test set 1 for sub-sample in Group One, Test set 2 for sub-sample in Group Two, and Test set 3 for sub-sample in Group Three.

The acceptable sample size for parameter estimates in IRT model is varied. The higher complexity of the model often requires larger sample sizes. A sample of approximately 500 subjects was found to be adequate for the parameter estimates in the two-parameter model (Goldman & Raju, 1986; Hulin et al., 1982; Liu & Yang, 2018). Accordingly, this study applied two-parameter logistic model as an appropriate model and administered with 525 respondents as the required sample size for the parameter estimation.

Of these individuals, 52.19 percent were female and 47.81 percent were male. The majority of the participants (63.05%) were recruited from extra-large schools. Half of the participants (55.23%) indicated having a GPAX ranging from 3.00-4.00. Approximately 42 percent had obtained a GPAX of 2.00–2.99. Based on this data, they were randomly divided into three groups as sub-samples as presented in Table 1.

Research Instrument

The parallel reading literacy test sets for ninth graders had been constructed as measures to assess student's progress in reading literacy performance. The tests were scenario-based for some reading purposes to achieve the particular goals as proposed in OECD (2019a). The instruments were three parallel test sets written in Thai including Test set 1, Test set 2, and Test set 3. These items were constructed to measure reading literacy skills, requiring student's ability to access and retrieve information, the ability to integrate and generate information, and the ability to evaluate and reflect. All items were five options for multiple-choice questions. In each test set, two 20-item test forms were combined, with the total of 40 items as follows: Test set 1 (Form A and Form B), Test set 2 (Form A and Form C), and Test set 3 (Form B and Form C). The tests were scored dichotomously either correct or incorrect. Reading passages were selected based on real-life situations (i.e. personal, public, occupational, and educational), texts (i.e. text type, sources, and format), and level of complexity (i.e. near transfer, far transfer, and very far transfer). Each set of reading passage comprised five items.

Table 1 Characteristics of the samples (N = 525)

Characteristic	Sub-sample One (n_1 = 181)	Sub-sample Two (n_2 = 178)	Sub-sample Three (n_3 = 166)	Total Sample (N = 525)
Gender (%)				
Female	44.75	58.98	53.02	52.19
Male	55.25	41.02	46.98	47.81
School size (%)				
Extra large	61.32	61.79	66.26	63.05
Large	16.58	16.86	14.46	16.00
Medium	22.10	21.35	19.28	20.95
GPAX (%)				
3.00–4.00 (High)	51.38	60.11	54.22	55.23
2.00–2.99 (Mid)	47.51	36.52	41.57	41.90
0.00–1.99 (Low)	1.11	3.37	4.21	2.85

Parallel Test Construction

The following procedures were presented in order to develop parallel test forms.

Item writing

Reading literacy items were created from the general guidelines of reading literacy measures proposed by OECD (2019a), which were used as a basis for generating test specification. Three reading literacy test forms were carefully constructed to match with the same test specification. As a result, each reading literacy form had the degree of similarity in terms of the underlying constructs, the same number of items, the reading texts, and wordings in questions and response options.

Expert review

This step involved having a group of 20 experts in the field of reading and reading assessment to review the content validity and parallelism of items to ensure that the items were parallel. They were divided into four groups and asked to review and evaluate items for clarity, redundancy, and degree of similarity of the item pairs on evaluation forms, with -1 being 'not equivalent' and 1 'equivalent' on the sameness. The agreement of experts was calculated using the index of item-objective congruence (IOC). The index scores higher than .5 represented those items were parallel and were appropriate to be included in the test administration.

In each test set, there were a total of 20 item pairs. With respect to parallel Test set 1, the results suggested that all item pairs had an IOC index higher than .5, except for two pairs (i.e. Item pairs 3 and 15) that needed to be revised. Similarly, the results of parallel Test set 2 showed that all item pairs had an IOC index higher than .5, except for three pairs (i.e. Item pairs 11, 12 and 15). For parallel Test set 3, most item pairs also had IOC index higher than .5, except for the three pairs (i.e. Item pairs 3, 11 and 12) that needed correction. As a result, all the comments by the experts were considered for revising items. The revised parallel test sets were later administered with the sample groups.

Model and parameter estimates

Further analysis was conducted after data collection to select the best fitted unidimensional IRT model for the data. This study applied the two-parameter logistic model as the best fitted model to the data. The item parameter analysis was determined in terms of the item discrimination (a-parameter) and item difficulty (b-parameter).

Statistical parallelism check

Several approaches were applied to check parallelism of the parallel reading literacy instruments in terms of test sets, test forms, and individual items.

Data Collection

The study was carried out in January-March 2020. All participants were randomly assigned to one of the three administration groups. Two out of three reading literacy test forms were organized into three test sets. The first group was administered the Test set 1 (parallel Form A and Form B). A second group completed the Test set 2 (parallel Form A and Form C). The final group was administered the Test set 3 (parallel Form B and Form C). They were administered for 70 minutes.

Data Analysis

Data analysis was computed using *R* statistical software version 3.6.0 (R Core Team, 2020). The two-parameter logistic model as the adequate model fit was applied in this study. Approaches for assessing parallelism comprise the same construct, the equal reliability, the indifferent means and standard deviations, and the equal parameter estimates (Petersen et al., 2020). The statistical parallelism was investigated into three levels, including test sets, test forms, and item pairs as follows:

1. The conformity between test sets was investigated by using the means, standard deviations, and reliability from the classical methods. One-way analysis of variance (ANOVA) was computed to investigate mean differences among the three groups. In addition, based on IRT method, One-way ANOVA was computed to examine the differences in terms of item discrimination and item difficulty among the three test sets. Moreover, the average TCC and the average TIF were presented.

2. Regarding the test forms, the statistical evidence for being parallel are TCC and TIF. RMSD statistic was used as an index to quantify the closeness of TCC and TIF between the test forms. If RMSD values were smaller than .50, both test forms were considered to be parallel (Debeer et al., 2017).

3. In examining parallelism at the level of the individual item, ICC and IIF were calculated for each individual item. Similar to the equivalence of test form, the RMSD was also calculated to determine the closeness of ICC and IIF of item pairs. The items were equivalent when RMSD values were smaller than .50 (Debeer et al., 2017).

Results

The Result of Statistical Parallelism between Test Sets

Means and standard deviations

The values of means and standard deviations lay in their similarity across test sets. The average mean and standard deviation of the Test set 1 ($M = 17.0$, $SD = 6.72$) were slightly greater than those of the Test set 2 ($M = 16.8$, $SD = 7.07$) and Test set 3 ($M = 16.4$, $SD = 6.91$), respectively. One-way ANOVA was conducted to determine the statistical differences between the means of reading literacy scores. The results yielded no difference in group means, $F(2,522) = 0.325$, $p = .723$, showing that examinees with different test sets had equal reading literacy mean scores. The boxplot for these data is shown in Figure 1.

Reliability

Cronbach's alpha coefficients of internal consistency were .81, .84, and .83, respectively, indicating good internal consistency of the test sets. This supports the conclusion that the scores across three parallel test sets were not different.

Item parameter estimates

One-way ANOVA suggested that there was no statistically significant variation in the mean discrimination parameter estimate among three parallel test forms, $F(2, 117) = 0.194$, $p = .824$. In addition, there were no statistically significant differences for the difficulty estimate among three test forms, $F(2, 117) = 0.761$, $p = .470$. It can be shown that the reading literacy test sets had similarity in terms of item parameter estimates.

Test characteristic curve and test information function

The results showed that the parallel test sets yielded similar conformity with respect to TCC and TIF as presented in Table 2. Figure 2 displayed the statistical plots of the three S-shaped curves and the precision of information of the three parallel test sets. The visual inspection showed that the three parallel test sets performed similar estimates.

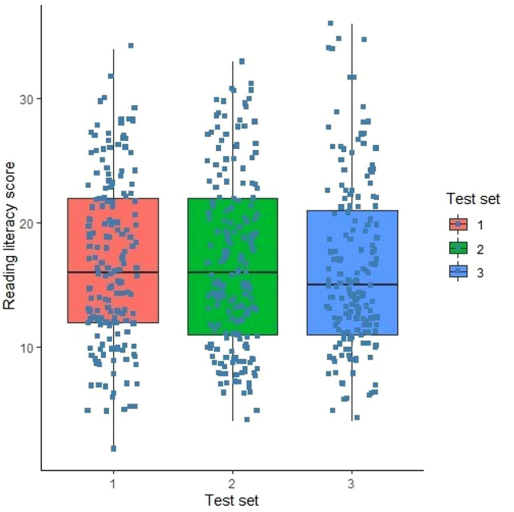


Figure 1 The boxplot of average reading literacy scores across test sets

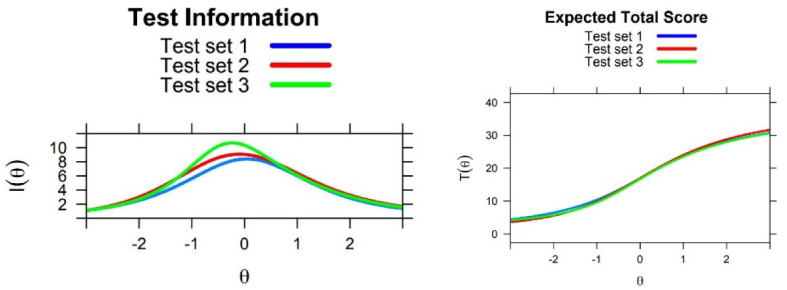


Figure 2 TCC and TIF plots of the three parallel test sets

Table 2 Test characteristic curve (TCC) and test information function (TIF) of the three parallel test sets

Theta	Test set 1		Test set 2		Test set 3	
	TCC	TIF	TCC	TIF	TCC	TIF
-3	4.40	1.10	3.59	1.12	4.17	1.11
-2	6.45	2.53	5.53	3.00	5.77	2.71
-1	10.38	5.62	9.83	6.83	9.57	7.41
0	16.89	8.39	16.93	9.03	16.68	10.33
1	23.65	6.00	23.98	6.37	23.50	6.09
2	28.12	2.87	28.75	3.27	27.95	3.08
3	30.78	1.38	31.65	1.66	30.62	1.55

The Result of Test Characteristic Curve and Test Information Function as Measures of Conformity between Test Forms

Test characteristic curve

Table 3 presents the TCC among the parallel test forms. The $RMSD_{TCC}$ values indicated that the test forms were parallel. The smaller TCC values represented greater conformity. The results showed that the $RMSD_{TCC}$ tended to be smaller in Test set 1 ($RMSD_{TCC}$ = .17) and in Test set 2 ($RMSD_{TCC}$ = .34). However, the highest statistical values were presented in Test set 3 ($RMSD_{TCC}$ = .71). To reveal more detailed information about the TCC congruence, the TCC for different test forms are plotted in Figure 3.

Test information function

Table 4 presents the TIF among the parallel test forms. The result found that the parallel forms within Test set 1 yielded acceptable value ($RMSD_{TIF}$ = .40). Similarly, Test set 2 represented the acceptable RMSD value ($RMSD_{TIF}$ = .17). On the other hand, $RMSD_{TIF}$ value of the Test set 3 was the greatest ($RMSD_{TIF}$ = .61). Figure 4 depicts the TIF showing the precision of the constructed test forms.

Table 3 TCC of the parallel test forms

θ	Test Set 1 (Form A*B)			Test Set 2 (Form A*C)			Test Set 3 (Form B*C)		
	TCC		Parallel	TCC		Parallel	TCC		Parallel
	Form A	Form B	$RMSD_{TCC}$	Form A	Form C	$RMSD_{TCC}$	Form B	Form C	$RMSD_{TCC}$
-3	1.40	2.14	.17	1.34	1.96	.34	1.33	2.06	.71
-2	2.38	2.77		2.31	2.64		2.12	2.79	
-1	4.66	4.51		4.68	4.37		4.37	4.37	
0	8.56	8.56		8.86	8.15		9.25	9.25	
1	11.90	12.74		12.82	11.76		13.53	13.53	
2	14.21	14.80		15.34	13.85		15.87	15.87	
3	15.85	15.70		16.76	15.09		17.09	17.09	

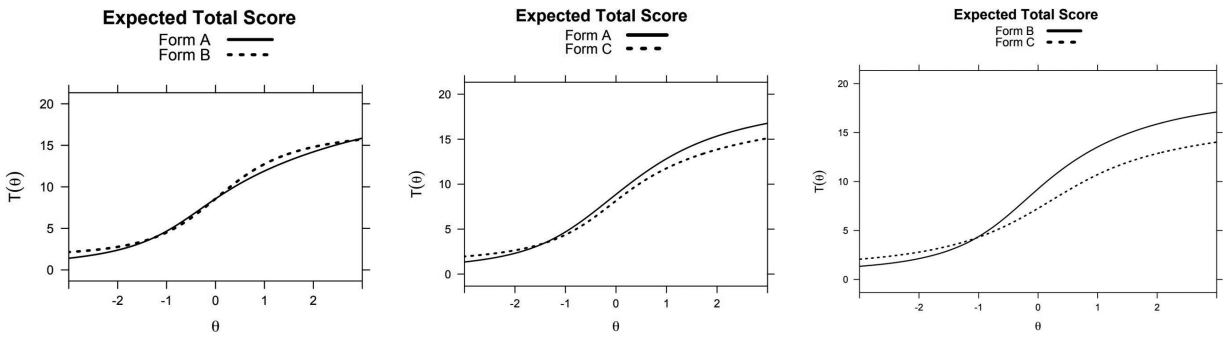


Figure 3 Plots of TCCs of the parallel forms

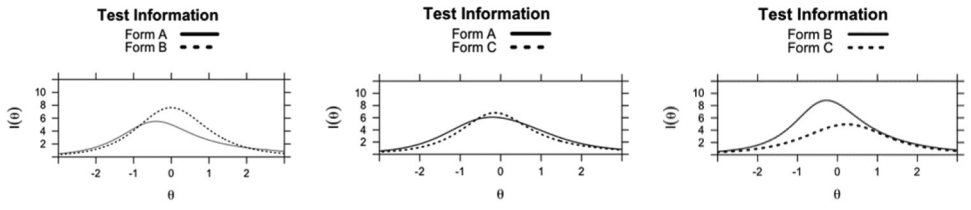


Figure 4 Plots of TIFs of the parallel forms

Table 4 TIF of the parallel test forms

θ	Test Set 1 (Form A*B)			Test Set 2 (Form A*C)			Test Set 3 (Form B*C)		
	TIF		Parallel	TIF		Parallel	TIF		Parallel
	Form A	Form B	$RMSD_{TIF}$	Form A	Form C	$RMSD_{TIF}$	Form B	Form C	$RMSD_{TIF}$
-3	0.49	0.36	.40	0.52	0.37	.17	0.51	0.40	.61
-2	1.51	1.18		1.68	1.13		1.56	0.97	
-1	4.36	4.08		4.62	3.88		5.66	2.52	
0	4.99	7.63		5.98	6.70		8.35	4.81	
1	2.60	4.24		3.89	3.29		3.92	3.65	
2	1.43	1.44		1.74	1.44		1.67	1.44	
3	0.86	0.52		0.79	0.73		0.75	0.62	

The Result of Item Information and Item Characteristic Curve as Measures of Item-By-Item Parallelism

With regard to Test set 1, the $RMSD_{IIF}$ ranged from .00–.14 and the $RMSD_{ICC}$ values were .01–.11., showing satisfactory closeness in IIFs between the two test forms. For Test set 2, the result also showed the conformity of item pairs, although the statistical values were larger than the Test set 1, ranging from .01–.20 and .02–.13 for $RMSD_{IIF}$ and $RMSD_{ICC}$, respectively. Similarly, the $RMSD_{IIF}$ and $RMSD_{ICC}$ of Test set 3 were relatively comparative, ranging from .00–.15 and .01–.14, respectively. This provides evidence that the test forms are parallel at the level of the individual items. The item-by-item closeness is presented in Table 5.

Discussion

We developed three parallel test sets for the assessment of reading literacy, each consisting of two test forms with 20 item pairs. A critical issue for development of parallel instruments is the replication of the tests. This study relied upon several statistical methods in establishing that the reading literacy tests are parallel.

The primary analysis of parallelism relying upon CTT showed that the three test sets yielded similar degrees in terms of means, standard deviations, or internal consistency reliability estimates. The classical criteria of parallel forms may be justified by equal means, standard deviations and standard errors of measurement. This is coherent with several previous studies (Bae & Lee, 2011; Clause et al., 1998; Furtkamp et al., 1982; Petersen et al., 2020; Tang &

Table 5 Item information function (IIF) and item characteristic curve (ICC) of the test forms

Pair	Test Set 1 (Form A*B)		Test Set 2 (Form A*C)		Test Set 3 (Form B*C)	
	$RMSD_{IIF}$	$RMSD_{ICC}$	$RMSD_{IIF}$	$RMSD_{ICC}$	$RMSD_{IIF}$	$RMSD_{ICC}$
1	.03	.02	.04	.03	.10	.03
2	.08	.04	.11	.12	.03	.03
3	.11	.07	.12	.12	.01	.02
4	.08	.07	.01	.13	.08	.14
5	.00	.02	.04	.09	.02	.03
6	.08	.04	.03	.04	.06	.07
7	.04	.03	.13	.08	.01	.02
8	.12	.11	.02	.02	.12	.09
9	.14	.06	.06	.03	.15	.06
10	.01	.07	.03	.07	.00	.07
11	.05	.03	.03	.02	.03	.02
12	.11	.08	.11	.08	.10	.01
13	.00	.02	.04	.05	.01	.02
14	.04	.05	.20	.07	.12	.05
15	.02	.02	.08	.05	.08	.04
16	.02	.01	.11	.10	.03	.02
17	.00	.07	.11	.10	.07	.13
18	.01	.01	.02	.06	.05	.10
19	.01	.06	.02	.05	.01	.05
20	.02	.07	.02	.09	.00	.06

Zhan, 2020). Petersen et al. (2020) state that means and standard deviations are evidences of the symmetry of the parallel forms. This result could be confirmed with the statistical analysis of IRT approaches. The results presented no differences in terms of item discrimination and item difficulty estimates among three test sets. The findings also presented conformity in accordance with TCC and TIF graphically. This might indicate that the test sets have similarity in terms of test difficulty and examinees have equal chance in their performances, regardless of the tested versions they received (Bae & Lee, 2011; Petersen et al., 2020). As a results, test scores from parallel test sets could be used interchangeably.

With respect to parallelism of test forms and item pairs, characteristic curves and information functions had the potential to measure conformity using the RMSD index for the parallelism of test forms and item pairs. More specifically, TIF and TCC methods yielded acceptable RMSD values for parallelism of Test set 1 (between Form A and Form B) and Test set 2 (between Form A and Form C). However, Test set 3 (between Form B and Form C) might not provide evidence to support parallelism. These results showed that the first two test sets have conformity in regard to TCC and TIF, whereas test set 3 does not. The reason might be because item difficulty estimates of Form C had unusually high values, including Item 44 (b = 45.78), Item 58 (b =13.27), and Item 59 (b = 8.51), whereas the item difficulty estimates of Form B ranged from 0.48–3.96. The deletion or revision of those items could result in the parallelism of both forms in terms of TCC and TIF of Test set 3. Accordingly, this study confirmed the conformity regarding TCC and TIF to determine the parallelism of reading literacy tests. TCC and TIF provide similar results of parallelism. Even though the TIF approach performed better than the TCC approach in terms of RMSD, the differences were not substantial. This study is coherent with the study of Armstrong et al. (1998) that there was a minimized difference in either TIF or TCC. On the other hand, the TIF seems to be more commonly chosen as the statistical characteristic and performed well to create parallel forms (Boekkooi-Timminga, 1990; Kantiwong & Pasiphol, 2018; Sanders & Verschoor, 1998; Van Der Linden & Adema, 1998; Veldkamp et al., 2013) This might be because the TIF is easier to manipulate and does not require identical observed score distributions (Lin, 2008; Samejima, 1977). Moreover, the difficulty characterized as the TCC might be implicit when the TIF of the tests are equal (Ariel et al., 2006). Hence, the selection of statistical indices for parallelism between the TCC and the TIF might be based on the purposes and natures of tests.

Similar to the results of TCC and TIF, the results of item-by-item parallelism showed that the RMSD seems to favor the IIF approach because it yielded higher parallelism than the ICC. However, both statistical indices interpreted the same meanings, showing that all items are parallel. The outcomes of both indices lead to a conclusion that the ICC performed as well as the IIF approaches.

Conclusion

This study aimed to investigate the statistical parallelism of the reading literacy instruments. The findings provided evidence supporting the parallelism at the level of the overall test sets, test forms and individual items. We conclude that IRT approaches were applicable for measuring the parallelism of the reading literacy instruments. Both TCC and TIF yielded satisfactory degrees of test parallelism. Moreover, ICC and IIF represented the acceptable degrees of item parallelism. Consequently, the three parallel test sets could be used as interchangeably alternative instruments for reading literacy assessment. This research may be beneficial to test developers and teachers who are attempting to develop the parallel measures in order to preserve the internal validity and ensure the comparable scores of the test forms without concerning the retest effects. Moreover, they are encouraged to make use of parallel forms in repeated measurement to facilitate the longitudinal assessment and still maintain test fairness. This research also encourages further investigation of parallel test forms using IRT with different kinds of subjects. On the contrary, this study is limited to RMSD in measuring conformity. A different study may be conducted with the different types of statistical indices for test and item parallelism such as the mean square deviation (MSD), the mean relative deviation (MRD), and the maximum relative information gap (MRIG).

Conflict of Interest

The authors declare there are no potential conflicts of interest.

Acknowledgments

This research was supported by the 100th Anniversary Chulalongkorn University Fund of Doctoral Scholarship and the 90th Anniversary Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund).

References

Ackerman, T. A. (1989). *An alternative methodology for creating parallel test forms using the IRT information function* [Paper presentation]. The Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*(3), 163–179. https://doi.org/10.1177/0146621615613308

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Appiled Psychological Measurement*, *30*(3), 204–215. https://doi.org/10.1177/0146621605284350

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*(3), 237–247. https://doi.org/10.1177/01466216980223004

Bae, J., & Lee, Y. S. (2011). The validation of parallel test forms: "Mountain" and "beach" picture series for assessment of language skills. *Language Testing*, *28*(2), 155–177. https://doi.org/10.1177/0265532210382446

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

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, *15*(2), 129–145. https://doi.org/10.3102/10769986015002129

Clause, C. S., Mulllns, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*(1), 193–208. https://doi.org/10.1111/j.1744-6570.1998.tb00722.x

Davey, T., & Hendrickson, A. (2010). *Classical versus IRT statistical test specifications for building test forms* [Paper presentation]. The Annual Conference of the National Council for Measurement in Education, Denver, CO.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Debeer, D., Ali, U. S., & Van Rijn, P. W. (2017). Evaluating statistical targets for assembling parallel mixed-format test forms. *Journal of Educational Measurement*, *54*(2), 218–242. https://doi.org/10.1111/jedm.12142

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *7*(4), 537–554. https://doi.org/10.1177/0265532209354771

Furtkamp, E., Giffort, D., & Schiers, W. (1982). In-class evaluation of behavior modification knowledge: Parallel tests for use in applied settings. *Journal of Behavior Therapy and Experimental Psychiatry*, *13*(2), 131–134. https://doi.org/10.1016/0005-7916(82)90054-4

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, *46*(1), 11–21. https://doi.org/10.1177/0013164486461002

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. https://doi.org/10.1177/014662168200600301

Kantiwong, A., & Pasiphol, S. (2018). Comparison of item and test parallel index between test forms among different characteristics of experts: Application of formal cultural consensus theory. *An Online Journal of Education*, *13*(4), 286–300.

Lin, C. J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *Journal of Technology, Learning, and Assessment*, *6*(8), 1–43. https://files.eric.ed.gov/fulltext/EJ838620.pdf

Liu, Y., & Yang, J. S. (2018). Interval estimation of latent variable scores in item response theory. *Journal of Educational and Behavioral Statistics*, *43*(3), 259–285. https://doi.org/10.3102/1076998617732764

Morin, A. J. S., Scalas, L. F., & Vispoel, W. (2017). The music self-perception inventory: Development of parallel forms A and B. *Psychology of Music*, *45*(4), 530–549. https://doi.org/10.1177/0305735616669426

Organisation for Economic Co-operation and Development [OECD]. (2016). *PISA 2015 Results: Excellence and equity in education (Volume 1)*. OECD Publishing. https://doi.org/10.1787/9789264266490-5-en

Organisation for Economic Co-operation and Development [OECD]. (2019a). *PISA 2018 Assessment and analytical framework, PISA*. OECD Publishing.

Organisation for Economic Co-operation and Development [OECD]. (2019b). *Thailand-country note - PISA 2018 results*. https://www.oecd.org/pisa/publications/PISA2018_CN_THA.pdf

Office of the Education Council. (2017). *National education plan B.E. 2560–2579 (2017–2036)*. http://www.onec.go.th/index.php/book/BookView/1540

Petersen, D. B., Spencer, T. D., Konishi, A., Sellars, T. P., Foster, M. E., & Robertson, D. (2020). Using parallel, narrative-based measures to examine the relationship between listening and reading comprehension: A pilot study. *Language, Speech, and Hearing Services in Schools*, *51*(4), 1097–1111. https://doi.org/10.1044/2020_LSHSS-19-00036

Quansah, F., & Cobbinah, A. (2021). Equivalence of parallel tests in a Basic Statistics course in higher education using classical measurement theory. *Canadian Journal of Educational and Social Studies*, *1*(2), 13–28. https://doi.org/10.53103/cjess.v1i2.11

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology*, *60*(2), 367–396. https://doi.org/10.1111/j.1744-6570.2007.00077.x

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*(2), 193–198. https://doi.org/10.1007/BF02294048

Sanders, P. F., & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement*, *22*(3), 212–223. https://doi.org/10.1177/01466216980223002

Schmucker, A., Abberger, B., Boecker, M., & Baumeister, H. (2019). Parallel short forms for the assessment of activities of daily living in cardiovascular rehabilitation patients (PADL-cardio): Development and validation. *Disability and Rehabilitation*, *41*(7), 826–832. https://doi.org/10.1080/09638288.2017.1407967

Tang, F., & Zhan, P. (2020). The development of an instrument for longitudinal learning diagnosis of rational number operations based on parallel tests. *Frontiers in Psychology*, *11*, 1–10. https://doi.org/10.3389/fpsyg.2020.02246

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

Van Der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*(3), 185–198. https://doi.org/10.1111/j.1745-3984.1998.tb00533.x

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*(2), 123–139. https://doi.org/10.1177/0146621612469825