Research Article https://doi.org/10.12973/eu-jer.12.2.1097

European Journal of Educational Research
Volume 12, Issue 2, 1097 - 1107.
ISSN: 2165-8714
http://www.eu-jer.com/

Study Item Parameters of Classical and Modern Theory of Differential Aptitude Test: Is it Comparable?

Farida Agus Setiawati* Universitas Negeri Yogyakarta, INDONESIA
Rizki Nor Amelia Universitas Negeri Semarang, INDONESIA
Bambang Sumintono Universitas Islam Internasional Indonesia, INDONESIA
Edi Purwanta Universitas Negeri Yogyakarta, INDONESIA

Received: November 27, 2022 ▪ Revised: February 16, 2023 ▪ Accepted: March 14, 2023

Abstract: This study aimed to find the Classical Test Theory (CTT) and Modern Test Theory (MTT) item parameters of the Differential Aptitude Test (DAT) and examined their comparability of them. The item parameters being studied are difficulty level and discrimination index. 5.024 data of the result sub-test DAT were documented by the Department of Psychology and Guidance and Counselling bureau. The parameter of classical and modern test items was estimated and correlated by examining the comparability between parameters. The results show that there is a significant correlation between item parameter estimates. The Rasch and IRT 1-PL models have the highest correlation toward CTT regarding the item difficulty level. In contrast, model 2-PL has the highest correlation toward CTT in the item discrimination index. Overall, the study concluded that CTT and MTT were comparable in estimating item parameters of DAT and thus could be used independently or complementary in developing DAT.

Keywords: *Classical test theory, differential aptitude test, item parameter, modern test theory.*

To cite this article: Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study item parameters of classical and modern theory of differential aptitude test: Is it comparable? *European Journal of Educational Research*, *12*(2), 1097-1107. https://doi.org/10.12973/eu-jer.12.2.1097

Introduction

The aptitude test is a psychological measurement instrument that measures the specific ability that involves a knowledge or skill domain (Hashmi et al., 2012; Marais, 2007; Shah & Raza, 2009). The application of the aptitude test is most widely used in measuring readiness for secondary, tertiary, to postgraduates school (Cohen & Swerdlik, 2018) as well as checking motivation, abilities, and skills in work (Avvannavar et al., 2013). An aptitude test must be fair and objective in identifying candidates with the greatest potential to succeed in their careers, regardless of their geographical, educational, or social background (Dewberry, 2011).

DAT was widely used in education, especially for ability identification and educational guidance for students choosing the study program. It is not easy to choose a study program to their abilities, making many Indonesian students select a different major from their potential (Masriah et al., 2018). This is in line with research findings that some Indonesian students choose study programs not based on their talents but more than other factors, such as family, individual personality, peers, campus image, job prospects, or school of origin (Saputro, 2017). According to Košir and Pečjak (2007), the students' mistakes in choosing college majors are a sign of indecision in selecting elections, which in the psychological construct is known as career doubts.

The Differential Aptitude Test (DAT) is the most popular, especially in the educational and vocational setting (Mahakud, 2013; Mankar & Chavan, 2013). This test is used to predict the academic performance of students (Muhid et al., 2020; Pyari et al., 2016). DAT was presented in the form of multiple-aptitude tests. They are numerical ability (NA), verbal reasoning (VR), clerical speed and accuracy (CSA), abstract reasoning (AR), language usage (consisting of sentences and spelling), space relations (SR), and mechanical reasoning (MR) (Bennet et al., 1956). Furthermore, the language usage sub-test was not applied because of the differences in Indonesian structure sentences.

The construct was based on the group factor of intelligence theory by Thurstone's Primary Mental Ability Model (D'Oliveira, 2004). Thurstone's Primary Mental Ability has been developed in various aptitude tests, one of which is the GATB (Hakstian & Bennet, 1978). Because many have been researched and used practically in the field, this DAT

* **Corresponding author:**
Farida Agus Setiawati, Department of Psychology, Universitas Negeri Yogyakarta, Indonesia. ✉ farida_as@uny.ac.id

instrument is considered good quality. However, items of DAT were also widely adapted and developed to detect talent and career selection, although efforts for re-study of item quality are often not carried out.

The quality of the instrument can be seen from the item, which is a sample of the overall attributes. The item quality is known as item parameters. There are two approaches to describing quality instruments: Classical Test Theory (CTT) and Modern Test Theory (MRT). The CTT was defined as a standard for test development. It has been the mainstay of psychological test development for more than the 20th (Embretson & Reise, 2000). The classical test theory is a simple approach that is easily understood and widely applied in empirical item analysis (Eleje et al., 2018). If the MTT model does not fit well due to a violation of the model assumption, then CTT will be a better choice (Mead & Meade, 2010). The simplicity is also related to the ease of analyzing and the few subjects or data needed (Qasem, 2013). Still, unfortunately, it does not involve true latent variables: even though the actual score is not empirically observable (Progar et al., 2008).

Rasch and IRT were the latest test analysis theories related to the latent variable models and the successor of the classical test theory in psychological assessment (Andrich, 2011; Boone & Scantlebury, 2006; Brennan, 2010; Thomas, 2011; Tractenberg, 2010). MTT describes the relationships between the latent traits, item characteristics in the scale, and the answers for each item (Bond & Fox, 2015; Yang & Kao, 2014). This method is based on two postulates. First, the performance or score of test items of the examinee can be predicted based on traits, latent traits, or abilities. Second, an Item Characteristic Curve (ICC) is a monotonically increasing function that describes the relationship between the test taker's item performance and a set of characteristics that underlie the item's performance (Hambleton & Swaminathan, 1985).

The differences between the approaches are that CTT and IRT typically describe data properties. In contrast, RMT aims to describe the data, item characteristic curve, and fit items (Petrillo et al., 2015). There are three logistics models for MTT based on the item properties examined for dichotomous data. 1-PL has only one item parameter: difficulty level (b), 2-PL adds discrimination index (a) as the second parameter, and 3-PL model adds pseudo guessing (c) as the third parameter (de Ayala, 2009; DeMars, 2010; Thorpe & Favia, 2012). The other MTT model was Rasch Model, developed by George Rasch, providing unbiased, efficient, and consistent estimation results on item and person calibrations (Dardick & Mislevy, 2016; de Ayala, 2009; DeMars, 2010). Even though Rasch Model looks similar to IRT 1-PL, they are different in assumption and assessment theory (Hu et al., 2021). The IRT 1-PL model focused on fitting the data as well as possible, given the model's constraints. In contrast, the Rasch model constructs the variable of interest (de Ayala, 2009). IRT was a statistical model that aimed to create a model that explained as much as possible the variance observed in the data, but Rasch showed invariant across participants and tested the data fit on a measurement scale (Stemler & Naples, 2021).

Analyzing item parameters using MTT will make it easier for researchers or test developers to equate items and make computer adaptive testing (Ekpo et al., 2016). This is because MTT provides researchers with various statistical tools to assess measurement characteristics. In CTT, item parameters depend highly on the subject or sample being measured (sample dependent). Whereas in MTT, the sample is invariant; that is, the item properties do not depend on the sample's ability level (Adedoyin et al., 2020; AL-khadher & Albursan, 2017; Kohli et al., 2015). In CTT, the ability estimation refers to the test takers' average scores and the reliability index. MTT, on the other hand, bases the probability of correctly answering a question on the ability (θ), item characteristics, and the model used (Baker, 2001). Therefore, the results of the analysis with MTT offered sophisticated information and comprehensive or robust, especially in terms of assessing the attributes of the instrument (Ekpo et al., 2016; Pollard et al., 2009).

Relevant studies about CTT and MTT applications (especially IRT) have been studied to describe the comparison and correlation item parameters. Still, local literature was limited in replicating the studies and results, and none used Rasch Model. The items' level of difficulty and discrimination index was the most popular studies of item parameters. This popularity was caused by the fact that both item parameters were usually used to evaluate the items under particular test conditions. Description of level difficulty and discrimination index on CTT and IRT studied by many researchers. The result of the study found that a lot of items were good item parameters from analyzed CTT and also suitable by IRT unless in the small item that could not be parallel (Abed et al., 2016; AL-khadher & Albursan, 2017; Bichi et al., 2019; Courville, 2004; Hashmi et al., 2012; Magno, 2009).

The item parameter analyzed in this study is the difficulty of the item and the discrimination index. In CTT, the difficulty was defined as the percentage of the examinee answering particular items correctly that scores between 0 to 1. Some experts suggest that the accepted item difficulty index is from .30 to .70 (Mehta & Mokhasi, 2014; Sayyah et al., 2012). Items that have a difficulty level below .30 can be categorized as difficult or hard, while items with a difficulty level above 0.70 are considered as easy. In the MTT, an item difficulty level was a point or location at which the S-shaped curve has the steepest slope in the ability scale. The degree of the item difficulty level based on modern test theory ranged from $-\infty$ to $+\infty$, although it was generally -2 until +2. It will be neither easy nor difficult for the intended subject (DeMars, 2010; Hambleton & Swaminathan, 1985). The item difficulty index below -2 was categorized as easy, between -2 and +2 was categorized as moderate, and more than +2 was hard.

The item discrimination index was defined as the ability to distinguish between high and low-performing students. This definition is synonymous with MTT. From a CTT perspective, an item discrimination index is calculated using a biserial point statistic with a coefficient ranging from -1 to +1. The items with an index below .2 are poor or revised, between .2 to .3 are acceptable, and above .3 are good items discrimination index (Boopathiraj & Chellamani, 2013; Mitra et al., 2009; Philip & Odunayo, 2017; Sayyah et al., 2012). Item discrimination of the MTT was scored as the range of the differences, with the score ranging from $-\infty$ to $+\infty$. However, the discrimination index usually varies between 0 to 2 and rarely surpasses 2 (Ahmadi & Thompson, 2012; Hambleton & Swaminathan, 1985). Therefore, the accepted discrimination index above 0 and the item discrimination index below 0 are not acceptable.

Fan (1998) conducted a correlational study using the Texas Assessment of Academic Skills (TAAS) instrument, a combination of multiple-choice and essay forms. The study found a correlation of difficulty between CTT and 1-PL IRT. The coefficient correlation for all groups is above .90 and significantly less than .05. The coefficient correlation of CTT was negative, so the higher the item was more difficult. Meanwhile, for CTT and IRT 2-PL, the coefficient was lower (below .90) than all groups compared to the correlation between CTT and IRT 1-PL. Moreover, this is the correlation with 3-PL. (2) Compared to the difficulty level, the correlation on the discrimination index was relatively lower (.60 to .90) in all conditions. This relationship shows significant differences between tests (mathematics vs. reads), sampling conditions, and IRT models (2 vs. 3-parameter logistic).

A more comprehensive study was conducted by Courville (2004). The study also carried out an analysis based on sample size grouped into smaller samples (n=100 students per sub-test) and more extensive samples (n=1000 students per sub-test). Item parameter characteristic data were analyzed using the American College Testing (ACT), consisting of 75 items for English, 60 for Mathematics, 40 for Reading, and 40 for Science. The facts gathered from the research were there was a significant correlation in the items' level of difficulty between CTT and IRT in all models. The coefficient of the correlation ranges from .553-1.000 (small sample) and .665-1.00 (significant sample). (2) There was a significant correlation in the discrimination index between CTT and IRT in all models, with the coefficient of the correlation ranging from .229-.957 (small sample) and .613-.930 (significant sample). Meanwhile, Progar et al., (2008) carried out a slightly different study. The study only investigated the correlation between CTT and 2-PL IRT. 2-PL IRT model was used because, among the three models of IRT, this model was proven to be the fittest with the data gathered from the International Mathematics and Science Study (TIMSS) instrument. The findings showed that in the mathematics sub-test, the correlation coefficients for the level of difficulty and discrimination index between CTT and the 2-PL IRT were .922 and .831, respectively.

Research about item property of DAT using CTT has been done by Setiawati et al. (2018a), also the analysis of item parameters using IRT sub-tests of verbal and numerical (Setiawati et al., 2018b) and space relations of DAT (Setiawati et al., 2018c). The completed information about items DAT using CTT and MTT needs to describe the property psychometric of items, also the relationship between two test theories. The correlation of parameter item property of DAT is based on CTT, IRT, and Rasch model. Analysis of all sub-tests needs to be carried out to check if both methods are comparable.

The research focuses on evaluating the property psychometric of a DAT using classical and modern test theories. It is essential to find out the parameters property all of the items. Whether there is any inconsistency in item parameter characteristics from the analysis results based on both classical and modern test theories, this technique is used to assess if items categorized as difficult based on CTT analysis are also considered difficult based on MTT analysis. If items have a satisfying discrimination index in CTT, they also have a satisfying index in IRT. Therefore, the study aims to find the DAT parameters item and examine its correlation using two test theories. The major hypothesis of this research is there were significant correlations between parameters item CTT and MTT's approaches. The minor hypotheses are: (1) There were correlations between item difficulty levels between CTT and MTT's approaches (Rasch, IRT 1-PL, IRT 2-PL, and IRT 3-PL), and (2) There were correlations in item discrimination index between CTT and MTT's approaches (IRT 2-PL, and IRT 3-PL).

Methodology

Research Design

This research used a quantitative approach consisting of interrelated parts—i.e., a study of instrument item parameters analysis using classical and modern analysis. The research focus describes the difficulty and discriminant of the items that they are the same parameter psychometric CTT and MTT.

Sample and Data Collection

Data of the study were documentation of the result of the psychological testing of students. These were collected from 2014 to 2017. Documentation of the test results was conducted in the laboratory of the Department of Psychology and Guidance and Counselling Center in the state university in Indonesia. The data were classified based on gender and educational qualification, displayed in Table 1. Because of data documentation, the number of data was different in each classification and the sub-test.

Table 1. Sample of Research

Sample Characteristic	Verbal reasoning	Numerical ability	Abstract reasoning	Space-relation	Mechanical reasoning
Gender					
Male	588	621	289	747	788
Female	420	426	451	299	395
Educational qualifications					
High School	171	147	42	42	146
Under Graduate	837	900	698	1004	1037
Total	1008	1047	740	1046	1183

The data collected from DAT included five subtests—i.e., verbal reasoning, numerical ability, abstract reasoning, space relations, and mechanical reasoning. These instruments were in the form of multiple-choice items consisting of five sub-tests with the 50 items of verbal reasoning, 40 items of numerical reasoning, 50 items of abstract reasoning, 60 items of space relations, and 68 items of mechanical reasoning. The reliability of the sub-test verbal reasoning is .809, the numerical ability is .850, abstract reasoning is .858, space relations is .862, and mechanical reasoning is .737.

Analyzing of Data

The data were analyzed quantitatively using the CTT and MTT to see the item parameters of the DAT, which included the difficulty (b) and discrimination index (a) of the items. Item parameter analysis based on the CTT was done using the ITEMAN software. In contrast, item parameter analysis based on RMT was done using Winsteps (Linacre, 2012) and IRT was conducted using the BILOG software for all logistic models (du Toit, 2003).

After the fit and parameters items ware estimated, the correlations between parameters were analyzed to examine the hypothesis research. The correlation technique used was the Pearson Product Moment. It provided information related to correlational statistics with its significance (the significant level this study was accepted if less than 0.05) that interpreted the relationship between item difficulty level of CTT and Rasch, 1-PL, 2-PL, as well, as 3-PL. Also, the correlation discrimination between CTT and IRT (2-PL and 3-PL) could be analyzed. The correlational item discrimination of the Rasch and 1-PL IRT models were not analyzed because this model couldn't estimate the item discrimination index.

Findings / Results

The study results described the number of items DAT from many criteria: item fits from MTT analysis, difficulty and discrimination index, the average, and the correlation of difficulty and discrimination index between two test theories. Table 2 presents the fit items, and table 3 displays the information about the parameters of item difficulty in DAT based on CTT and MTT. Most items are fit models. Two items of verbal (item number 23 for IRT 2-PL; item number 46 for IRT 3-PL) and many items of mechanical sub-test (item numbers 5, 22, 52, 55, 61 for IRT 2-PL and 3-PL; item numbers 41, 45, 48, 53 for IRT 3-PL) could not be analyzed. Most items are categorized as moderate, followed by a few numbers easy and hard items.

Table 2. The Number of Items fit model of DAT

Sub-test	1PL		2-PL		3-PL	
	Fit	Unfit	Fit	Unfit	Fit	Unfit
Verbal reasoning	23	27	47	3	47	2
Numerical ability	17	23	36	4	32	8
Abstract reasoning	36	14	50	-	48	2
Spatial relation	30	30	59	1	59	1
Mechanic reasoning	36	32	58	5	57	2

Table 3. The Item Difficulty Distribution of DAT

Sub-test	Difficulty Level	CTT	Rasch	1-PL	2-PL	3-PL
Verbal reasoning	Easy	12	4	4	7	5
	Moderate	24	43	43	32	33
	Hard	14	3	3	10	11
Numerical ability	Easy	8	3	3	4	2
	Moderate	27	35	35	33	34
	Hard	5	2	2	3	4

Table 3. Continued

Sub-test	Difficulty Level	CTT	Rasch	1-PL	2-PL	3-PL
Abstract Reasoning	Easy	31	2	21	19	13
	Moderate	18	45	28	30	35
	Hard	1	3	1	1	2
Space relations	Easy	30	2	2	11	3
	Moderate	19	51	52	41	49
	Hard	11	7	6	8	8
Mechanical reasoning	Easy	19	1	0	16	4
	Moderate	38	61	62	37	47
	Hard	11	6	6	10	8

Table 4 summarizes the item discrimination indexes in DAT based on CTT and IRT analysis. It can be seen that all sub-tests analyzed by IRT 2-PL and 3-PL indicate that all items have a good item discrimination index. The item discrimination index of IRT provides a higher number of items with an accepted discrimination index than CTT. Based on CTT analyses, mechanical reasoning has the lowest number of accepted items of the five sub-tests.

Table 4. Item Discrimination Index Distribution on DAT Instrument

Sub-test	Item Discrimination Index	CTT	IRT 2-PL	IRT 3-PL
Verbal reasoning	Good	27	49*	49*
	Acceptable	12	-	-
	Poor/to be revised	11	-	-
Numerical ability	Good	30	40	40
	Acceptable	8	-	-
	Poor/to be revised	2	-	-
Abstract reasoning	Good	38	50	50
	Acceptable	8		
	Poor/to be revised	4	-	-
Space relations	Good	41	60	60
	Acceptable	14	-	-
	Poor/to be revised	5	-	-
Mechanical reasoning	Good	20	63*	59*
	Acceptable	23	-	-
	Poor/to be revised	25	-	-

* Some items of verbal and mechanical sub-test cannot be calibrated on IRT 2-PL and 3-PL

Table 5 shows descriptive statistics on average item difficulty levels and item discrimination indexes on DAT based on CTT and MTT analyses. Based on the difficulty level, the verbal reasoning sub-test is the most difficult one, because it has the lowest average item difficulty on CTT and highest average item difficulty on MTT. Meanwhile, abstract reasoning subtest is the easiest sub-test based on the analysis of CTT and MTT because it has the highest mean of item difficulty in CTT and the lowest mean of item difficulty in MTT. Then, based on the item discrimination index, the numerical ability is the sub-test with the best item discrimination index based on CTT analysis. Meanwhile, abstract reasoning is the sub-test with the best item discrimination index based on 2-PL IRT analysis. The space relation is the sub-test with the best item discrimination index based on 3-PL IRT analysis.

Table 5. The Average of Difficulty and Discrimination Items of DAT on CTT and MTT

Subtests	CTT		MTT					
			Rasch	IRT 1-PL	IRT 2-PL		IRT 3-PL	
	b	a	b	b	b	a	b	a
Verbal reasoning	.489	.306	.000	.000	.352	.777	.551	1.108
Numerical ability	.531	.377	.000	.000	-.211	.883	.255	1.157
Abstract reasoning	.744	.361	.000	-1.730	-1.448	.989	-1.071	1.131
Space relation	.596	.331	.000	.000	-.363	.934	.019	1.387
Mechanical reasoning	.542	.226	.000	.000	-.550	.564	.290	1.025

Note: *b* = difficulty of item *a* = discrimination index of item

The research hypothesis tested the correlations between item parameters in CTT and MTT. Thus, the data analysis was described from the correlations of item difficulty between CTT and MTT (Rasch, 1-PL, 2-PL, and 3-PL) and the

correlations on item discrimination index between CTT and MTT's approaches (2-PL vs 3-PL). Table 6 presented the Pearson correlation of item difficulty in the results of CTT and MTT analyses. Numerical ability, mechanical reasoning, verbal reasoning, and space relation correlate significantly with high and negative correlation coefficients. The item difficulty correlation between CTT and Rasch gives the highest correlation value compared to CTT and 1-PL, CTT and 2-PL, and CTT and 3-PL.

Table 6. Correlation of Item Difficulty Level (b) between CTT and MTT

Sub-test		MTT			
		Rasch	**1-PL**	**2-PL**	**3-PL**
Verbal reasoning	**CTT**	-.987**	-.985**	-.540**	-.918**
Numerical ability		-.994**	-.992**	-.895**	-.824**
Abstract reasoning		-.981**	-.980**	-.903**	-.962**
Space relation		-.995**	-.990**	-.917**	-.898**
Mechanical reasoning		-.995**	-.987**	-.858**	-.903**

Note: ** = significant correlation of < .01

Table 7 presents the Pearson correlation of discrimination indexes based on two test theories. There was positive correlations coefficient between CTT and MTT on the whole sub-test. Because the Rasch and IRT 1-PL assume fixed (a constant value) item discrimination for all items, it could not be computed; hence N/A (not applicable) is entered under both columns for Rasch and IRT 1-PL in the table. The correlation coefficient of CTT and 2-PL is higher than CTT and 3-PL, except on the verbal sub-test. The numerical sub-test gives most significant correlation coefficient, while the space relation sub-test provides the smallest correlation coefficient.

Table 7. Correlation of Item Discrimination Index (a) Between CTT and MTT

Sub-test		MTT			
		Rasch	**1-PL**	**2-PL**	**3-PL**
Verbal reasoning	**CTT**	N/A	N/A	.373**	.766**
Numerical ability		N/A	N/A	.848**	.758**
Abstract reasoning		N/A	N/A	.744**	.734**
Space relation		N/A	N/A	.824**	.280*
Mechanical reasoning		N/A	N/A	.749**	.328*

Note: * = significant correlation of < .05. ** = significant correlation of < .01

Discussion

Table 3 shows the difficulty item calculated based on two test theories, and the number of items categorized as moderate is higher analyzed by MTT than CTT. This difference is due to the way of classifying. In CTT, the category of moderate items is in the range of 0.3 to 0.7. It has a smaller limit than the moderate item category in MTT (-2 to 2). The discrimination indexes are also shown in Table 4. Discrimination index was an item parameter used to categorize an item that is good, acceptable, or needs improvement. This study indicates that more items must be corrected when analyzed by CTT than MTT. By MTT analysis, all items could be used because it has acceptable item discrimination (a > 0 or positive value), while in CTT, some items needed to be better. Of course, that condition influenced the limit made by CTT, with a minimum item discrimination index of 0.2 compared to a logit scale of positive value on MTT. It means that the quality category of MTT parameters is broader than CTT, thus allowing more items to pass the selection. This finding was consistent with (Eleje et al., 2018), who showed that many item of DQEST (Diagnostic Quantitative Economics Skill Test) were rejected by CTT rather than IRT. It means that the criteria of item parameters determine the basis for categorizing good items or not, especially in the results in the area around the border. The different criteria used in analyzed item parameters allow different results that would influence the decision-making in interpreting the items' quality from the analysis results.

The IRT analysis indicates that across all parameter logistic models (see Table 5). The item difficulty level of the abstract reasoning sub-test shows the same interpretation as the CTT analysis. Therefore, it can be concluded that in all the tested logistic models and CTT analyses. The abstract reasoning sub-test is the easiest compared to the other four sub-tests. Although considered the easiest subtest, abstract reasoning has the highest item discrimination by the 2-PL. For details, the results of the analyses towards the DAT show that: (1) the average item difficulty level based on CTT ranges from .489 to .774; meanwhile the average item difficulty level based on the MTT for Rasch, 1-PL, 2-PL, and 3-PL models ranges from .000, -1.730 to .000, -1.448 to .352, and -1.071 to .551 respectively; and (2) the average item discrimination index based on CTT ranges from .226 to .377 and the item discrimination index based on IRT for 2-PL and 3-PL models ranges from .564 to .989 and 1.025 to 1.387. The Rasch analysis showed that the average item difficulty level is consistent at 0 for all sub-tests. For Rasch, the average item difficulty level is always set at 0, which

indicates the initial reference point of the scale (Sumintono & Widhiarso, 2015). Furthermore, Rasch model, 1-PL, and 2-PL have a lower index of difficulty item than the 3-PL model in all tested sub-tests. Likewise, the item discrimination of 2-PL IRT that also lower than the 3-PL in all the sub-tests.

A significant, linear, and negative correlation could be found between CTT and MTT (on Rasch and IRT across all logistic models) in terms of item difficulty (see Table 6). The result of a negative correlation is because the difficulty of item in CTT has the opposite meaning to MTT. As mentioned above, the difficulty in CTT is the percentage of the examinee answering particular items correctly. This means that the higher the proportion, the item is rated as an easy item, and vice versa. Refers to the significance, the all-correlation coefficient of the item difficulty levels between CTT and MTT is below 0.01. Table 5 also shows that there are variations in correlation coefficients. Rasch model has the highest correlation compared to the other models in all sub-tests, followed by 1-PL. In contrast, 2-PL and 3-PL models could not be told that one had a higher correlation coefficient than the other. The correlation coefficient of CTT-IRT 2-PL and CTT-IRT 3-PL appear to be slightly weaker, although still reasonably strong. This is because most of the coefficients were above .80 for most conditions, except for the verbal sub-test (r = -.540 based on CTT-IRT 2-PL).

Regarding the item discrimination index, the results of this study revealed a significant linear and positive correlation between two test theories (see Table 7). The correlation coefficient of the item discrimination index between CTT and IRT ranges from .373 to .848 (2-PL) and .280 to .766 (3-PL). For CTT-IRT 2-PL, the numerical sub-test gave the highest correlation coefficient (r = .848), while the verbal sub-test gave the lowest correlation coefficient (r = .373). Meanwhile, for CTT-IRT 3-PL, the verbal sub-test gave the highest correlation coefficient (r = .766), while the space relation sub-test gave the lowest correlation coefficient (r = .280). Similar to item difficulty level, the correlation coefficient of the item discrimination index in CTT with IRT 2-PL and 3-PL were inconsistent; the correlation CTT to 2-PL model's correlation coefficient was higher than CTT to the 3-PL model, except for the verbal test. But in sum, this finding was consistent with previous studies (Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; MacDonald & Paunonen, 2002). The results further showed that the item difficulty categorization (on the CTT) also corresponded to MTT.

Generally, this study found that CTT and MTT can be used separately or together to evaluate the item's parameter properties. Even though this study didn't in line with the research that found which 3-PL gives a better coefficient of correlation towards CTT than the 1-PL and 2-PL (Çıkrıkçı, 2002) It supported the researcher's conclusion that the item characteristic of 1-PL and 2-PL are more suitable for CTT compared to the 3-PL (Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; MacDonald & Paunonen, 2002). Moreover, this study provides the findings that the CTT and Rasch Model shows the best coefficient on item difficulty level. However, numerically it is not too different from CTT-IRT 1-PL. The 3-PL model does not give a better result because this model accommodates a pseudo-guessing parameter that significantly decreases variants between IRT and CTT by penalizing the low performer with advanced specialized knowledge and also the non-guesser (Hernandez, 2009).

Although CTT and MTT offer comparable parametric terms, MTT can display the ICC and more sophisticated mechanisms in conceptualizing measurement errors. In addition, MTT also provides analytical tools for development, parameter invariants for both items and persons, and suitable statistical models. Areas where the MTT advantage is not apparent, are in small samples. The data are inconsistent with the MTT model used, possibly in the area of usability.

The choice of which psychometric approach to use depends on some factors. Researchers must develop appropriate assessment methods and consider the audience. A cursory survey of psychometric properties using CTT is acceptable if the tool development aims to explain phenomena with a small sample and a limited budget. However, for high-risk tests (e.g., the college entrance exams, the high school or secondary school exit exams, and the professional licensure exams), psychometric properties are explored and presented as RMT, IRT, or both quantitative and qualitative desirable.

Conclusion

This study depicts item parameters, item difficulty, and discrimination indexes of DAT items analyzed using CTT and MTT and their categories. The results accept the major and minor hypotheses. There were correlations between parameters item CTT and MTT's approaches; there is a correlation between item difficulty between CTT to RMT, IRT 1, 2, and 3-P, and there is a correlation in item discrimination index between CTT to IRT 2-PL and IRT 3-PL. Based on the coefficient correlation, the Rasch Model with IRT 1-PL has the highest correlation toward CTT in the item difficulty level estimates. The item discrimination indicates the coefficient correlation between CTT and IRT 2-PL was more elevated than between CTT and IRT 3-PL.

Recommendations

Generally, CTT and MTT are comparable in evaluating the characteristics of the item. Thus, the research implications are that if the good item parameters of CTT indicated the good parameters of MTT and vice versa, the bad item parameters of CTT indicated the bad parameters of MTT. The analysis results with one theory can be a good prediction with another theory so that the use of one study can represent how well the item analysis is in the field. This research shows that the Differential Aptitude Test parameter items have various properties. Other researchers can use these results or DAT in selecting or selecting items. Good and fit items can be used for collections or question banks for

practical needs and question development. Unfit and bad items can be corrected or not used in practical measurements in the field.

Limitations

The results of this study demonstrate the psychometric properties of the field data. These psychometric properties may change if they are applied to different data. This is following measurements using classical theory, which are not invariant, so re-analysis of psychometric properties with the classical approach should always be carried out in new research. In contrast to modern analysis, which is invariant, it is possible to get the same results even though the repeated study is carried out with different samples or targets. Of course, this applies to items that have a fit model. For this reason, it is necessary to evaluate the unfit model and delete or impair the DAT items so that further development and utilization of the item analysis results are needed to obtain relevant measurement results.

This research is analyzed from field data which sometimes gets results that could be under better conditions. Further data analysis with simulation data, generating row scores with various situations, and analyzing them with classical and modern theories will allow for more accurate results.

Authorship Contribution Statement

Setiawati: Conceptualization, design, analysis, writing. Amelia: Data analysis, drafting manuscript. Sumintono: Critical revision of manuscript, statistical analysis. Purwanta: Reviewing, supervision.

References

Abed, E. R., Al-Absi, M. M., & Abu shindi, Y. A. (2016). Developing a numerical ability test for students of education in Jordan: An application of Item Response Theory. *International Education Studies*, *9*(1), 161. https://doi.org/10.5539/ies.v9n1p161

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2020). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *International Journal of Educational Research and Reviews*, *7*(11). https://bit.ly/3JFG5T3

Ahmadi, A., & Thompson, N. A. (2012). Issues affecting Item Response Theory fit in language assessment: A study of differential item functioning in the Iranian National University entrance exam. *Journal of Language Teaching and Research*, *3*(3), 401–412. https://doi.org/10.4304/jltr.3.3.401-412

AL-khadher, M. M. A., & Albursan, I. S. (2017). Accuracy of measurement in the classical and the modern test theory: An empirical study on a children intelligence test. *International Journal of Psychological Studies*, *9*(1), 71–80. https://doi.org/10.5539/ijps.v9n1p71

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 571–585. https://doi.org/10.1586/erp.11.59

Avvannavar, S. M., Ambrose, B., & Chandavarkar, M. (2013). Determination of effectiveness of aptitude test to improve sincerity in the recruitment process. *International Journal of Management*, *4*(4), 75–81.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. https://eric.ed.gov/?id=ED458219

Bennet, G. K., Seashore, H. G., & Wesman, A. G. (1956). The Differential Aptitude Test: An overview. *The Personnel and Guidance Journal*, *35*(2), 81–91. https://doi.org/10.1002/j.2164-4918.1956.tb01710.x

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Bin Ibrahim, A. (2019). Comparative analysis of classical test theory and item response theory using chemistry test data. *International Journal of Engineering and Advanced Technology*, *8*(5C), 1260–1266. https://bit.ly/40Ifc8k

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental mesurement in the human sciences. In *Applying the Rasch Model* (Third ed.). Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2), 253–269. https://doi.org/10.1002/sce.20106

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, *2*(2), 189–193. https://bit.ly/3lzCVYQ

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Cohen, R. J., & Swerdlik, M. E. (2018). Psychological testing and assessment. In G. K. Zammit & J. W. Hull (Eds.), *Guidebook for clinical psychology interns* (pp. 121-134). Springer. https://doi.org/10.1007/978-1-4899-0222-1_8

Courville, T. G. (2004). *An empirical comparison of Item Response Theory and Classical Test Theory item/person statistics*. Texas A&M University. https://core.ac.uk/download/pdf/147123147.pdf

Çıkrıkçı, N. (2002). A study of Raven Standard Progressive Matrices Test's item measure under classic and item response models: An empirical comparison. *Journal of Faculty of Education Sciences*, *35*(1), 71-80. https://doi.org/10.1501/Egifak_0000000055

D'Oliveira, T. C. (2004). Dynamic spatial ability: An Exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology*, *14*(1), 19–38. https://doi.org/10.1207/s15327108ijap1401_2

Dardick, W. R., & Mislevy, R. J. (2016). Reweighting data in the spirit of Tukey: Using Bayesian Posterior Probabilities as Rasch Residuals for studying misfit. *Educational and Psychological Measurement*, *76*(1), 88–113. https://doi.org/10.1177/0013164415583351

de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guildford Press.

DeMars, C. (2010). *Item Response Theory: Understanding statistics measurement*. Oxford University Press. https://doi.org/j3rq

Dewberry, C. (2011). Aptitude testing and the legal profession. In *Aptitude tests currently used in the professional services sector* (Issue June). https://bit.ly/3JAUaAT

du Toit, M. (Ed.). (2003). *IRT from SSI*. Scientific Software International, Inc.

Ekpo, E. O., Egbonyi, I. C., & Bassey, S. W. (2016). Transformation from classical test theory (CTT) to item response theory (IRT) in research instrument validation. *Journal of Educational Research*, *1*(4), 104–117. https://bit.ly/40wyFZ3

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of Classical Test Theory and Item Response Theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences*, *3*(1), 57–75. https://bit.ly/3TQjKa5

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Hakstian, A. R., & Bennet, R. W. (1978). Validity studies using the comprehensive ability battery (CAB): II. Relationship with the DAT and GATB. *Educational and Psychological Measurement*, *38*(7), 1003–1015. https://doi.org/10.1177/001316447803800419

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and appliction* (Vol. 1). Springer.

Hashmi, M., Zeeshan, A., Saleem, M., & Akbar, R. (2012). Development and validation of an aptitude test for secondary school mathematics students. *Bulletin of Education and Research*, *34*(1), 65-76. https://bit.ly/3ZGmmZc

Hernandez, R. (2009). Comparison of the item discrimination and item difficulty of the Quick-Mental Aptitude Test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment*, *1*(1), 12–18. https://bit.ly/3JCh9Mb

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of Classical Testing Theory and Item Response Theory. *Psychology*, *12*(9), 1397–1409. https://doi.org/10.4236/psych.2021.129088

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among Classical Test Theory and Item Response Theory frameworks via Factor Analytic Models. *Educational and Psychological Measurement*, *75*(3), 389–405. https://doi.org/10.1177/0013164414559071

Košir, K., & Pečjak, S. (2007). Personality, motivational factors and difficulties in career decision-making in secondary school students. *Psychological Topics*, *16*(1), 141–158. https://hrcak.srce.hr/file/32320

Linacre, J. M. (2012). *Winsteps help for Rasch analysis*. Winsteps. https://www.winsteps.com/winman/copyright.htm

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement*, *62*(6), 921–943. https://doi.org/10.1177/0013164402238082

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. In *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11. https://bit.ly/3Z9TSXE

Mahakud, G. C. (2013). Is it essential to measure intelligence along with aptitude test for career guidance. *Research World-Journal of Arts, Science & Commerce*, *4*(1), 92–102. https://bit.ly/3Z9TiZY

Mankar, J., & Chavan, D. (2013). Differential aptitude testing of youth. *International Journal of Scientific and Research Publications*, *3*(7), 1–6. https://bit.ly/3K0BImK

Marais, A. C. (2007). *Using the differential aptitude test to estimate intelligence and scholastic achievement at grade nine level* (Issue June) [University South Africa]. https://bit.ly/3lBZ5d1

Masriah, Z., Nursalim, M. M., & Fitriani, A. (2018). Persepsi mahasiswa terhadap jurusan perguruan tinggi dan konsep diri dengan kesesuaian minat memilih [Effect of student perceptions of college majors and self-concept on the suitability of interest in choosing]. *Anfusina: Journal of Psychology*, *1*(1), 61–76. https://doi.org/10.24042/ajp.v1i1.3639

Mead, A. D., & Meade, A. W. (2010). CTT and IRT 1Test Construction using CTT and IRT with unrepresentative samples. *Paper Presented at the Annual Meeting of the Society for Industrial and Organizational Psychology in Atlanta GA*, 56. https://bit.ly/3naYBeg

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple choice questions-An assessment of the assessment tool. *International Journal of Health Sciences and Research*, *4*(7), 197–202. https://bit.ly/42KzWxF

Mitra, N. K., Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2009). The levels of difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *International EJournal of Science Medicine Education*, *3*(1), 2–7. https://doi.org/10.56026/imu.3.1.2

Muhid, A., Yusuf, A., Kusaeri, Novitasari, D. C. R., Asyhar, A. H., & Ridho, A. (2020). Determining scholastic aptitude test as predictors of academic achievement on students of Islamic School in Indonesia. *New Educational Review*, *61*, 211–221. https://bit.ly/3ZJOafu

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Philip, A., & Odunayo, O. B. (2017). Application of Item Characteristic Curve (ICC) in the selection of test items. *British Journal of Education*, *5*(2), 21–41. https://bit.ly/409wvif

Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: An item analysis using Classical Test Theory and Item Response Theory. *Health and Quality of Life Outcomes*, *7*, Article 41. https://doi.org/10.1186/1477-7525-7-41

Progar, S., Socan, G., & Slovenija, M. P. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology*, *17*(3), 5–24. https://bit.ly/3JGdr3Y

Pyari, P., Mishra, K., & Dua, B. (2016). A study of impact of aptitude in mathematics as stream selection at higher secondary level. *Issues and Ideas in Education*, *4*(2), 141–149. https://doi.org/10.15415/iie.2016.42011

Qasem, M. A. N. (2013). A comparative study of Classical Theory (CT) and Item Response Theory (IRT) in relation to various approaches of evaluating the validity and reliability of research tools. *IOSR Journal of Research & Method in Education*, *3*(5), 77–81. https://doi.org/10.9790/7388-0357781

Saputro, M. (2017). Analisis faktor-faktor yang mempengaruhi keputusan mahasiswa dalam memilih program studi [Analysis of the factors that influence student decisions in choosing a study program]. *Jurnal Pendidikan Informatika Dan Sains*, *6*(1), 83–94. http://bit.ly/3MAr8or

Sayyah, M., Vakili, Z., Alavi, N. M., Bigdeli, M., Soleymani, A., Assarian, M., & Azarbad, Z. (2012). An item analysis of written multiple-choice questions: Kashan University of Medical Sciences. *Nursing and Midwifery Studies*, *1*(2), 83–87. https://core.ac.uk/download/pdf/143838611.pdf

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018a). Evaluasi karakteristik psikometrik tes bakat [Evaluation of psychometric property of the aptitude test]. *Humanitas*, *15*(1), 46–61. https://bit.ly/3ZC7laV

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018b). Analisis respon butir pada tes bakat skolastik [The item response theory analysis of scholastic test]. *Jurnal Psikologi/Psychology Journal*, *17*(1), 1–17. https://doi.org/10.14710/jp.17.1.1-17

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018c). Items parameters of the space-relations subtest using Item Response Theory. *Data in Brief*, *19*, 1785–1793. https://doi.org/10.1016/j.dib.2018.06.061

Shah, A. F., & Raza, M. A. (2009). *The impact of parents ' education towards the science aptitude of the students at elementary level in Southern Punjab*. *Pakistan Journal of Social Sciences*, *29*(1), 117–125. https://bit.ly/3n9qGme

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item Response Theory: Knowing when to cross the line. *Practical Assessment, Research and Evaluation*, *26*, Article 11. https://doi.org/10.7275/v2gd-4441

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi permodelan Rasch pada assessment pendidikan* [Rasch modeling application in educational assessment]. Trim Komunikata.

Thomas, M. L. (2011). The value of Item Response Theory in clinical assessment: A review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797

Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology*, *20*, 1–34.

Tractenberg, R. E. (2010). Classical and modern measurement theories, patient reports, and clinical outcomes. *Contemporary Clinical Trials*, *31*(1), 1–3. https://doi.org/10.1016/S1551-7144(09)00212-2

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177.