

# Diagnosing Students' Ability in Distance Learning with the Winsteps

Safari\*

Center for Assessment and Learning, Research and Development  
Ministry of Education and Culture  
Jakarta, Indonesia

\*safari\_puspendik@yahoo.com

**Abstract**—The main purpose of this research is to answer the following question. Is Winsteps Software easy for teachers to diagnose the ability of students and questions in distance learning to anticipate the spread of the corona virus today, especially for elementary / MI, SMP / MTs, SMA / MA / SMK teachers in Indonesia. The method used in this study is the library method. The data in this study are in the form of information from 8 related software. After making software that is easy for teachers to use at school, relevant software immediately puts into practice to diagnose it. The data used were student answer data on the 50-item English Middle School test with 4 multiple choice questions, which were answered by 30 students. The analytical method used in this research is a descriptive analysis. Based on the results of the analysis of several software, the presentation of results with the most information is Winsteps. To run the software, most must first write syntax, such as Parscale, ConQuest, Mplus, IRTPRO 3, including Bilog-MG, except Winsteps. Of the 8 software that is easy to use by the teacher is Winsteps software. The conclusion is that the Winsteps program is very practical for teachers to diagnose students' abilities and questions about aspects of knowledge, skills and attitudes in distance learning to anticipate the spread of the corona virus.

**Keywords**—winsteps, diagnosis, ability, learning, corona

## I. INTRODUCTION

The condition of a pandemic, corona virus or covid-19, forces most students in all countries of the world to carry out learning activities at home because the health and safety of students, teachers, education personnel, families, and the community is a top priority. Current educational activities ranging from selection, learning process, for graduation are held online. Of course, things like this cannot completely replace educational activities, especially the physical learning process in the classroom, school and campus. Both are here to complement each other, not to replace each other. The sense of love and respect for the teacher grows when a student kisses the teacher's tag when entering and returning from school and the personal interactions that occur during the education process. New diversity is tested when students interact physically with school friends or lectures from various circles. Homeland love grows when students sing the national anthem

at official events at school. Character like this must be maintained even though there is a pandemic. Teachers and parents have to guard it closely so that the character doesn't just disappear because of Covid-19. Although the condition of each family in some areas is not the same level of spread of the covid virus, including the availability of learning facilities such as computers both at home and at school [1]. Therefore, this study will help teachers, especially in terms of diagnosing the ability of students who carry out learning activities from their homes. Hopefully this experience and study can be useful for other countries that have not yet found a way of transition [2].

In pandemic conditions, the teacher's task of diagnosing the ability of students from the results of the distance learning assessment is very important. Based on the results of the diagnosis, the teacher can immediately provide appropriate feedback [3]. Because the activity is a process of gathering, summarizing, and using information from students' answers to make decisions and follow-up [4]. The purpose of the diagnosis is to find the source of learning difficulties and formulate a remedial action plan or its improvement. In addition to diagnostic purposes, also to help analyze the measurement tool that is improving its quality through the revision of ineffective questions [5]. Quality questions are questions that can provide information in accordance with the objectives of which can determine which students have or have not mastered the material taught by the teacher.

Teachers in schools are used to diagnose students' abilities and analyzing items with classical theoretical approaches such as ITEMAN, EXCEL, SPSS [6]. For future development towards the 21st century the implementation of literacy and numeracy literacy, especially distance learning during this pandemic, the use of this theory needs to be complemented by a broader measurement of the benefits of using the Rasch model. Rasch is a mathematical process that knows no substantive causes and consequences developed by Georg Rasch. It converts sequence numbers to additives (linear, interval) steps on the mathematical defined "dimensions". For convenience, this dimension is labeled "latent variable". If the data generated by the simulation, the latent variable have no meaning at all, but it has no effect on the operation of the Rasch model [7]. The Rasch model is a psychometric analysis

technique that can be used by teachers to develop, test items that can provide information related to student assessments for learning. The advantages of the Rasch Model include being able to give rights according to students' abilities that are based on a linear scale with the same interval. Students who are right or wrong in answering the same number of questions will get a different score if the different item numbers are answered correctly or incorrectly. Through this analysis teachers can easily diagnose the ability level of their students or Indonesian students when compared to the ability of international level students such as the PISA test, TIMSS, PIRSL, or other international tests. Based on the results of the analysis, the teacher can easily determine the correct improvement. Because the results of applying the Rasch model are very good and precise, as well as the compatibility of items to the model [8].

Several computer programs have been developed to calibrate item and person parameters in the Rasch model, such

as Winsteps [9], Bilog-MG [10], Parscale [11], WinGen [12][13], Param [14], ConQuest [15]. In addition, there are still many other programs that complement them, including: X-Caliber [16], LTM package [17] and the irtoys package [18] in R [19], jMetrik [20], IRTPRO [21], flexMIRT [22], and Mplus [23]. In this study only a few programs will be discussed, then recommended a program that is very easy to implement by teachers in schools. The intended software is: (1) Winsteps, (2) Bilog-MG, (3) Parscale, (4) WinGen, (5) ConQuest, (6) Mplus, and (7) IRTPRO 3.

A. Winsteps

Winsteps is a Rasch model analysis program developed by John Mike Linacre in 2012 which is a development of the Bigsteps program. The results of the analysis of the Winsteps program provide 36 information, as in Figure 1 below.

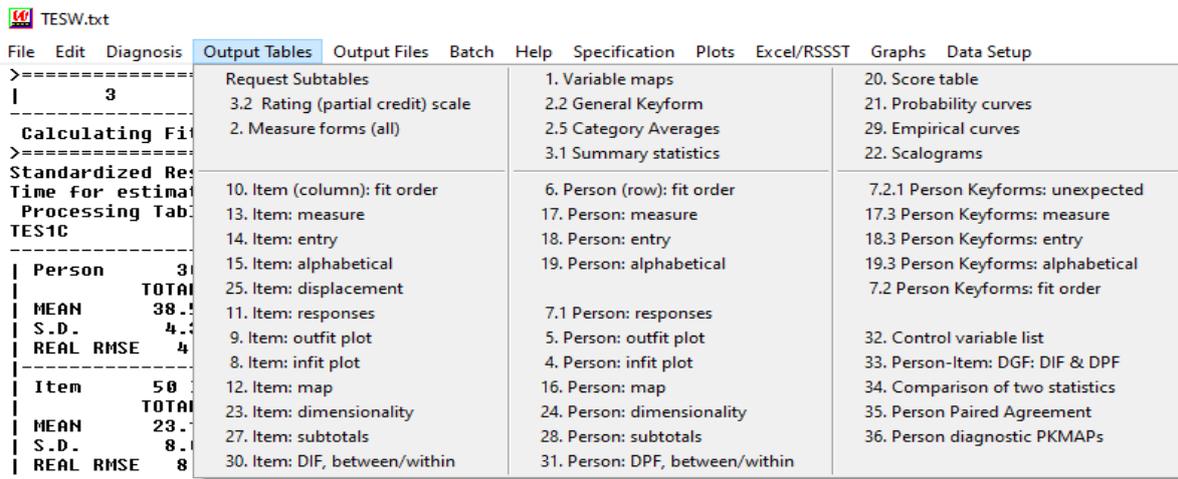


Fig. 1. Results of winsteps program analysis.

This result is very helpful for teachers in distance learning today, which is to provide complete information related to diagnosing student abilities and items.

B. Bilog-MG 3

Bilog-MG 3 was developed by Michele Zimowski, Eiji Muraki, Robert Mislevy, and Darrell Bock [10]. This program can display several important things, such as: (1) graphics; (2) binary item analysis including multiple choice forms, short answers, true / false / omit scores or blank / not displayed scores; (3) able to analyze large scale and handle several groups; (4) item analysis and scoring of a number of subjects or subscales; (5) non-statistic equating group, vertical equating group; (6) Differential item functioning (DIF); (7) detected and corrected for trend parameters over time or detection and correction for trends over time parameters (DRIFT); (8) Calibration and scoring of tests in a two-stage testing procedure; (9) estimation of latent ability or arability

distribution; (10) items are statistically, theoretically and empirically reliable; (11) information curves and test form reliability; (12) IRT graphics can be imported into Word, Access, etc. To use this program, writing syntax must be accurate.

Comparison of Winsteps and Bilog-MG software, Capability Estimation [24] in Table 1 shows the comparison of the average and standard deviations of simulated abilities and estimated the abilities of WINSTEPS and the two phases of BILOG-MG. The means of ability for the two phases of BILOG-MG experienced more setbacks to the overall average than those for WINSTEPS. The pattern of growth, with a smaller increase as class levels rise, is the same for all programs. The standard deviation for BILOG-MG Phase 2 is consistently smaller than BILOG-MG Phase 3 and WINSTEPS. The standard deviations range from 1.06 to 1.22 for WINSTEPS (0.16), 0.86 to 1.06 for Phase 2 (0.20), and 1.04 to 1.21 for Phase 3 (0.17).

TABLE I. SIMULATED, WINSTEPS AND BILOG-MG ABILITY ESTIMATES FOR SIMULATED DATA

Grade/ Level	Simulated Target		Simulated Sample		WINSTEPS		BILOG-MG Phase 2		BILOG-MG Phase 3	
	M	SD	M	SD	M	SD	M	SD	M	SD
Second	-0.27	0.92	-0.23	0.90	-0.12	1.06	0.00	0.86	0.04	1.04
Third	0.50	1.05	0.48	1.04	0.50	1.16	0.56	0.97	0.60	1.14
Fourth	1.09	1.12	1.08	1.13	1.07	1.22	1.02	1.06	1.07	1.21
Fifth	1.59	1.10	1.61	1.11	1.57	1.20	1.43	1.04	1.47	1.19
Sixth	1.90	1.01	1.90	1.01	1.90	1.15	1.70	0.99	1.72	1.13

Source: [24]

**C. Parscale 4.0**

Parscale 4.0 was developed by Eiji Muraki and Darrell Bock [11]. This program can display several important things, such as: (1) one, two, and three parameter logistics models; (2) the Samejima model for grade response; (3) partial credit models; (4) scale item analysis for the form of description, multiple choice; (5) Differential item functioning (DIF); (6) combined analysis of several types of items; (7) inter-rater effect analysis; (8) several groups of polytomous grain response models; (9) IRT graphics can be imported into Word, Access, etc.; (10) assistance is provided through the HELP menu for description, syntax, and examples.

number of computer programs to simulate IRT data have been developed since the early 1970s. However, most were developed under DOS (eg DATAGEN [25], GENIRV, [26], and RESGENT [27]). As a result, this program is limited due to DOS-related problems: (1) slowing down the performance speed (16-bit), (2) limited system resources, and (3) incompatibility with 32-bit Windows-based OS. Some computer simulation programs, such as WINIRT [28] and PARDSIM [29], were developed for Windows-based OS, but only provide a few options for generating data from the IRT model. Therefore, WinGen was developed to complete all the programs. Table 2 summarizes the features of WinGen and some other simulation software available.

**D. WINGEN**

WINGEN was developed by Kyung T. Han and Ronald K. Hambleton of the University of Massachusetts Amherst [13]. A

TABLE II. COMPARISON OF THE WINGEN SIMULATION PROGRAM AND OTHER PROGRAMS

	DATAGEN	GENIRV	RESGEN	WINIRT	PARDSIM	WinGen2
Authors	Hambleton & Rovinelli (1973)	Baker (1989)	Muraki (1992)	Fang & Johanson (2005)	Yoes (1997)	Han (2007)
Examinees	4000	4000	1000	4000	N/A	100.000.000
Items	400	100	100	400	N/A	100.000.000
OS	DOS	DOS	DOS	Windows	Windows (16,32 bit)	Windows (32,64 bit)
IRT Models	1 Dichotomous	1 Dichotomous 2 Polytomous	1 Dichotomous 2 Polytomous 3 Multidimensional	1 Dichotomous	1 Dichotomous	1 Dichotomous 2 Polytomous 3 Nonparametrik 4 Multidimensional
Distribution	1 Uniform 2 Normal	1 Uniform 2 Normal	1 Uniform 2 Normal 3 Lognormal 4 Gamma	1 Uniform 2 Normal	1 Uniform 2 Normal	1 Uniform 2 Normal 3 Lognormal 4 Two-Parameter Beta
Options	1 Replicating		1 Replicating 2 Multiple Blocks, subtests, groups			1 Replicating 2 Introducing DIF/ Drift 3 Communicating with other IRT SOFTWARE 4 Analyzing outputs of other IRT software 5 Plotting ICF, TCC, IIF, TIF, and histogram of thetas 6 Multiple tests, groups

Source: [12]

**E. ConQuest**

ConQuest is a computer program to adjust the response of items and latent regression models. ConQuest was developed by the Australian Council for Educational Research (ACER) and the University of California, Berkeley, February 25, 2007.

The programmers are Margaret Wu, Ray Adams, and Sam Haldane. Professor Ray Adams is Director of the ACER Global Education Monitoring Center and is a member of the University of Melbourne Professor, specializing in psychometrics, educational statistics, large-scale testing, and international comparative studies. The program provides a

comprehensive and flexible factor response model to performance appraisal analysis, traditional valuation, and rating scale. ConQuest also provides extensive research, the most recent psychometric methods of the multifaceted item response model, the multidimensional item response model, the latent regression model and the theta estimate with its error distribution.

ConQuest unites in one program a variety of item response models (including multidimensional models) and provides integration of item responses and regression analysis: (1) Rasch Simple Logistic Model, (2) Rating Scale Model, (3) Partial Credit Model, (4) Ordering Partition Model, (5) Linear Logistic Test Model, (6) Multifaceted Model, (7) Generalized Unidimensional Model, (8) Multidimensional Item Response Model, (9) Latent Regression Model. ConQuest also generates an estimated MMLE maximum likelihood for model parameters. The estimation algorithm used is an adaptation of the quadrature method described by Bock and Aitken [30] and the Monte Carlo method of Volodin and Adams [31]. The suitability of the model is ensured by the generalization of the Wright and Masters residue-based method (1982) developed by Wu [32] which includes: (1) several ConQuest applications, (2) conducting item analysis, (3) examining item differential functions, (4) exploring effect assessors, (5) estimate latent correlations, (6) test dimensions, and (7) draw values that make sense. The demo version can be downloaded at [www.acer.edu.au/quest](http://www.acer.edu.au/quest).

#### F. The latest software

The latest software, Mplus 7 [23] and IRTPRO 3 [21], which can be used to analyze various kinds of data and variables. Estimating the parameters of multidimensional items and the ability of examinees. IRTPRO 3 software has the ability to run unidimensional and multidimensional models in 1PL, 2PL, 3PL, and partial credit models. Mplus 7 can run unidimensional and multidimensional models, but several studies have shown that it has difficulty estimating guessing parameters (c) [33]. In addition, the newer version of Mplus produces factor loads and thresholds as well as item difficulties and discrimination and therefore no conversion is needed. Both software programs can run simulations and can produce charts for items and tests. Mplus 7 can also be used to run both confirmatory factor analysis and exploratory factor analysis. IRTPRO 3 cannot carry out this analysis and hence the user must know which a priori items are loaded into which dimensions when modeling a multidimensional test. Unable to run EFA indicates that IRTPRO 3 must be used in conjunction with other software tools that have the ability to do so. The ability of IRTPRO 3 to run various models and even different models in different parts of the test makes it far superior to Mplus 7 in modeling capabilities. Mplus 7's ability to carry out exploratory factor analysis and confirmatory factor analysis provides benefits for practitioners who want to use or can only afford one software application. Table 3 provides a visual comparison of each software program as well as other information.

TABLE III. COMPARISON OF MPLUS AND IRTPRO

Topic	MPLUS 7	IRTPRO 3
Ability to write syntax	Yes	Yes
Point and click user interface	No	yes
1PL models	Yes	Yes
2PL models	Yes	Yes
3PL models	Yes	No
Exploratory factor analysis	Yes	no
Confirmatory factor analysis	yes	Yes
Provides item parameters	Yes, but must convert	Yes
Provides examinee ability	Yes	Yes
Graphics	yes	Yes

Source: [34]

From the various descriptions above, the problem with this study is which software is very appropriate and easy for teachers to use in distance learning to diagnose student abilities and items? Is Winsteps Software easy to use by teachers in schools, especially teachers in Indonesia, namely elementary / MI, SMP / MTs, SMA / MA / SMK teachers?

## II. RESEARCH METHODS

The method used in this study is the library method. Researchers obtained data from the results of literature studies from various literature books, journals, scientific works, and also other documents that study related to the Rasch model. The basis for using this method is adapted to the main objective of this research, among others, is to obtain a computer program that is easily used by teachers in distance learning at this time, especially teachers in Indonesia, namely elementary / MI, SMP / MTs, SMA / MA / SMK. The recommended software in this study is 8 software: Winsteps, Bilog-MG, Parscale, WinGen, Param, ConQuest, Mplus, IRTpro3. The data in this study are in the form of information from the 8 software concerned. After establishing a software that is easy for teachers to use in school, the relevant software is immediately practiced to analyze the items. The data used are the data of students' answers in the English Middle School test 50 items with 4 multiple choice questions, which were answered by 30 students. The analytical method used in this study is a descriptive analysis.

## III. RESULTS AND DISCUSSION

Based on the analysis of all software, most of the programs must first write the syntax, such as Parscale, ConQuest, Mplus, IRTPRO 3, including Bilog-MG, except Winsteps. The writing syntax for teachers requires individual understanding and must take time for it. Therefore, for analysis activities, practical software is really needed. From all software, the presentation of the results of the analysts with the most information is Winstep. This information is seen in the summary of the analysis results in Figure 1, Table 4 and 5.

TABLE IV. SOFTWARE OPERATION TECHNIQUES

Software Name	Kind of Data	Data Import	Making Syntax	Fill in the Column	IRT
Winsteps	Dicotomus and politomus	Yes	No	No	Rasch Model
Bilog-MG	Dicotomus and politomus	Yes	No	No	1,2,3 PL
Parscale	Politomus	Yes	Yes	No	1,2,3, PL
ConQuest	Dicotomus and politomus	Yes	Yes	Yes	1,2,3, PL
Mplus	Dicotomus and politomus	Yes	Yes	No	1,2,3, PL
IRTPLUS	Dicotomus and politomus	Yes	Yes	No	1,2,3, PL

TABLE V. RESULTS OF EACH SOFTWARE

Name Software	Reliability	Item Fit	Wright Map	Item Measure	Option/ Distractor Analysis	Guttman Scalogram	Transform Score
Winsteps	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bilog-MG	Yes	Yes	No	Yes	No	No	No
Parscale	Yes	Yes	No	Yes	No	No	No
ConQuest	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mplus	Yes	Yes	No	Yes	No	No	No
IRTPLUS	Yes	Yes	No	Yes	Np	No	No

Based on the analysis results in Table 4 and 5, and Figure 1, the software that is most easily used by teachers to diagnose student abilities and items is Winsteps. Winsteps program is very practical to be used by teachers in schools (SD / MI, SMP / MTs, SMA / MA / SMK) to diagnose student abilities and items on aspects of knowledge, skills, and attitudes because this program can analyze various types of data. Like dichotomous data (true-false), that is data from answers to questions in the form of multiple choice and politomus data, namely data from answers to questions about description, practice (performance, project, product), and attitude / character scale. Student version of Wisteps Program / demo with the name Minitap can be downloaded for free to analyze a maximum of 25 items and 75 students. The website (website) <http://www.winsteps.com/ministep.htm>. The Winsteps program to analyze data that is larger than that is up to 9,999,999 students and 60,000 teachers have to buy the program with licenses that cost around \$ 149 with a version that is the latest June 2020 is Winsteps 4.5.4.

In distance learning at this time, the results of the Winsteps analysis are very useful for teachers, including: (1) diagnosing students' abilities and follow-up, (2) diagnosing students' ability to items, (3) diagnosing student answer patterns, (4) item and person reliability, (5) item fit, (6) item difficulty level, (7) analyzing deception / option, (8) analyzing rubric scale, and (9) score transformation.

A. Diagnose Students' Abilities and Follow-up

All software displays this result even though the column display varies. To diagnose the ability of students in the Winsteps program can be seen in the "output tables number 17 above. Person: measure". Its form is as follows figure 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT [MNSQ]	OUTFIT [ZSTD]	PT-MEASURE [CORR.]	EXACT MATCH [EXP%]	Person				
1	45	50	74.53	4.98	.85	-.3	.34	.58	.53	92.3	90.7	0	
2	44	50	71.74	4.61	.77	-.7	.32	.62	.55	89.7	88.5	0	
15	43	50	69.31	4.34	.81	-.6	.40	.63	.57	87.2	86.5	1	
22	43	50	69.31	4.34	.81	-.6	.40	.63	.57	87.2	86.5	2	
30	43	50	69.31	4.34	.81	-.6	.40	.63	.57	87.2	86.5	3	
3	42	50	67.13	4.15	.69	-1.2	.34	.67	.59	85.7	85.5	0	
9	42	50	67.13	4.15	.85	-.5	.43	.64	.59	84.6	85.5	0	
11	42	50	67.13	4.15	.85	-.5	.43	.64	.59	84.6	85.5	1	
14	42	50	67.13	4.15	.94	-.1	.96	.64	.59	84.6	85.5	1	
16	42	50	67.13	4.15	.94	-.1	.96	.64	.59	84.6	85.5	1	
21	41	50	65.13	3.99	.85	-.5	.82	.63	.60	89.7	84.3	2	
26	41	50	65.13	3.99	1.15	-.6	1.07	.56	.60	79.5	84.3	2	
4	40	50	63.27	3.87	1.26	1.0	.94	.56	.61	74.4	83.3	0	
13	40	50	63.27	3.87	1.21	-.9	2.05	1.5	.54	.61	79.5	83.3	1
5	39	50	61.51	3.77	.78	-.9	.56	.69	.62	87.2	82.7	0	
6	39	50	61.51	3.77	1.08	-.4	1.68	1.2	.59	.62	82.1	82.7	0
25	39	50	61.51	3.77	.90	-.3	.94	.65	.62	87.2	82.7	2	
17	38	50	59.84	3.68	.82	-.7	.67	.69	.64	84.6	82.0	1	
27	38	50	59.84	3.68	1.05	-.3	1.72	1.3	.59	.64	84.6	82.0	2
7	37	50	58.23	3.61	.82	-.7	.57	.70	.64	87.2	81.3	0	
10	37	50	58.23	3.61	1.03	-.2	1.03	.63	.64	76.9	81.3	1	
12	37	50	58.23	3.61	1.13	-.6	2.18	2.1	.56	.64	82.1	81.3	1
29	37	50	58.23	3.61	1.46	1.8	1.47	1.1	.52	.64	71.8	81.3	2
28	36	50	56.69	3.55	1.48	1.9	1.98	1.9	.50	.65	74.4	80.6	2
18	35	50	55.20	3.50	.82	-.8	.64	.71	.66	82.1	80.2	1	
23	33	50	52.34	3.41	1.12	-.6	1.09	.4	.64	.67	71.8	78.8	2
8	31	50	49.59	3.35	1.38	1.7	1.30	.9	.58	.67	61.5	77.1	0
20	31	50	49.59	3.35	1.03	-.2	1.04	.2	.67	.67	71.8	77.1	2
19	30	50	48.25	3.33	1.10	-.5	1.21	.7	.64	.68	74.4	76.7	1
24	29	50	46.92	3.32	.70	-1.6	.62	-1.1	.76	.68	87.2	76.5	2
MEAN	38.5	50.0	61.41	3.87	.98	.0	.95	.1		82.2	82.9		
S.D.	4.3	.0	7.11	.40	.22	.9	.54	.9		7.2	3.4		

Fig. 2. Diagnosing students' capabilities.



TABLE 3.1 TES1C  
 INPUT: 30 Person 50 Item REPORTED: 30 Person ZOU366WS.TXT Jun 23 13:54 2021  
 50 Item 2 CATS WINSTEPS 3.7

SUMMARY OF 30 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFINIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	38.5	50.0	61.41	3.87	.98	.0	.95	.1
S.D.	4.3	.0	7.11	1.40	.22	.9	.54	.9
MAX.	45.0	50.0	74.53	4.98	1.48	1.9	2.18	2.1
MIN.	29.0	50.0	46.92	3.32	.67	-1.6	.30	-1.1
REAL RMSE	4.03	TRUE SD	5.86	SEPARATION	1.46	Person	RELIABILITY	.68
MODEL RMSE	3.89	TRUE SD	5.95	SEPARATION	1.53	Person	RELIABILITY	.70
S.E. OF Person MEAN = 1.32								

Person RAW SCORE-TO-MEASURE CORRELATION = .99  
 CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .72

SUMMARY OF 39 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFINIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	21.9	30.0	48.31	5.14	.99	.1	.95	.1
S.D.	7.4	.0	13.96	1.92	.22	1.0	.40	1.1
MAX.	29.0	30.0	81.69	8.47	1.58	3.2	1.90	3.5
MIN.	3.0	30.0	30.78	3.26	.59	-1.6	.36	-1.6
REAL RMSE	5.65	TRUE SD	12.76	SEPARATION	2.26	Item	RELIABILITY	.84
MODEL RMSE	5.48	TRUE SD	12.83	SEPARATION	2.34	Item	RELIABILITY	.85
S.E. OF Item MEAN = 2.26								

MAXIMUM EXTREME SCORE: 1 Item  
 MINIMUM EXTREME SCORE: 10 Item  
 UMEAN=48.3059 USCALE=8.2534

Fig. 5. Summary statistics.

The information is as follows. (1) Mean Measure Person = 61.41 logit, whereas Mean Measure Item = 48.31 logit. This shows the average value of students in working on all the items given. Value 61.41 > 48.31 The average value of the item. This shows that the value of student ability is greater than the level of difficulty of the questions. (2) The reliability of the test is chosen on the Cronbach Alpha (KR-20) Person raw Score "Test" Reliability = 0.72. This reliability is an interaction between person and item as a whole. According to Sumintono and Widhiarso [35] the criteria are <0.5 = Poor; 0.5-0.6 = Bad; 0.6-0.7 = Enough; 0.7-0.8 = Good; > 0.8 = Very good. The reliability of the person above is 0.72 which means Good. (3) Person reliability and item reliability values are 0.68 and 0.84. According to Sumintono and Widhiarso [35] the criteria are <0.67 = Weak; 0.67-0.80 = Enough; 0.80-0.90 = Good; 0.91-0.94 = Very good; > 0.84 = Special. So the consistency of answers from students is "Enough", while the quality of the items in the test aspect of reliability is "Good". (4) MNSQ INFIT and MNSQ OUTFIT for table person = 0.98 and 0.95

while for the item table = 0.99 and 0.95. The analysis is the ideal value is 1.00. The closer to the value of 1.00, the better the quality. For ZSTD INFIT and ZSTD OUTFIT for person table = 0,0 and 0,1 for item table = 0,1 and 0,1. The analysis is the ideal value is 0.0. The closer to the value of 0.0, the better the quality.

E. Fit Items

The level of suitability of the item is used to see the accuracy of the item with the model or item fit. Item fit explains whether the item functions normally to take measurements or not. If there are items that are not fit, this indicates a subject misconception in answering the problem. All software displays these results with different views. For Winsteps click output tables, then select 10. Item: fit order the results are as in Figure 6 below.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFINIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXACT MATCH EXP.	OBS%	EXP%	Item
14	27	30	-.89	.63	1.21	.6	1.90	1.3	A-.08	.26	90.0	89.9	S14
38	29	30	-2.12	1.03	1.08	.4	1.74	.9	B-.05	.16	96.7	96.6	S38
31	14	30	1.76	.39	1.48	3.2	1.73	3.5	C-.19	.37	46.7	66.1	S31
48	8	30	2.76	.44	1.11	.6	1.66	1.8	D .09	.32	80.0	74.1	S48
18	19	30	.97	.41	1.58	3.0	1.62	2.5	E-.23	.38	43.3	70.1	S18
11	29	30	-2.12	1.03	1.07	.4	1.42	.7	F .00	.16	96.7	96.6	S11
40	27	30	-.89	.63	1.18	.5	1.29	.6	G .04	.26	90.0	89.9	S40
22	28	30	-1.36	.75	1.13	.4	1.25	.6	H .03	.22	93.3	93.3	S22
20	14	30	1.76	.39	1.17	1.2	1.21	1.2	I .19	.37	60.0	66.1	S20
26	29	30	-2.12	1.03	1.06	.4	1.16	.6	J .04	.16	96.7	96.6	S26
4	17	30	1.29	.40	1.15	1.1	1.16	.9	K .22	.38	60.0	67.1	S4
5	12	30	2.07	.40	1.14	1.0	1.12	.6	L .22	.36	53.3	66.9	S5

Fig. 6. Item statistics: Misfit order.

The results of this analysis are arranged based on misfit order. Winstep has sorted items by unsuitable items. Items that are not fit are placed at the top. As shown in the example above, the items displayed have low Point Measure Correlation values even if other criteria (outfit means-square and z-standard outfit) still meet the requirements.

According to Boone, Staver and Yale [36], the value of means-square outfit, z-standard outfit, and point measure correlation are the criteria used to see the level of appropriateness of items. If the item does not meet the criteria, it is better if the item is repaired or replaced. Guidelines for assessing item conformity criteria according to Boone, et al [36] are: (1) Outfit Mean Square (MNSQ) values received: 0.5 <MNSQ <1.5; (2) Z-standard outfit values (ZSTD) received: -

2.0 <ZSTD <+2.0; (3) Point Measure Correlation value received: 0.4 <pt measure corr <0.85. Because the point measure correlation is in principle the same as the point-biserial correlation in classical test theory, Alagumalai, Curtis, and Hungi [37] classify the value of Point Measure Correlation to be very good (> 0.40), good (0.30-0, 39), sufficient (0.20-0.29), unable to discriminate (0.00-0.19), and requires examination of items (<0.00).

F. The Difficulty Level of the Items

All software displays this result even though the column display varies. For Winsteps the level of difficulty of the bullet can be seen in "output tables number 13 from the table above. Item: measure". Its form is as follows Figure 7.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
46	0	30	6.49	1.83	MAXIMUM MEASURE			.00	.00	100.0	100.0	S46	
13	3	30	4.05	.62	1.09	.4	.98	-.2	.12	.21	90.0	90.0	S13
48	8	30	2.76	.44	1.11	.6	1.66	1.8	.09	.32	80.0	74.1	S48
8	9	30	2.58	.42	.95	-.2	.89	-.3	.39	.33	70.0	71.3	S8
27	9	30	2.58	.42	.95	-.2	.81	-.6	.42	.33	63.3	71.3	S27
32	9	30	2.58	.42	1.08	.5	1.05	.3	.25	.33	76.7	71.3	S32
5	12	30	2.07	.40	1.14	1.0	1.12	.6	.22	.36	53.3	66.9	S5
28	13	30	1.92	.40	1.07	.6	1.04	.3	.30	.37	63.3	66.4	S28
20	14	30	1.76	.39	1.17	1.2	1.21	1.2	.19	.37	60.0	66.1	S20
31	14	30	1.76	.39	1.48	3.2	1.73	3.5	-.19	.37	46.7	66.1	S31
4	17	30	1.29	.40	1.15	1.1	1.16	.9	.22	.38	60.0	67.1	S4
9	17	30	1.29	.40	1.00	.1	.97	-.1	.39	.38	73.3	67.1	S9
18	19	30	.97	.41	1.58	3.0	1.62	2.5	-.23	.38	43.3	70.1	S18
25	19	30	.97	.41	.75	-1.6	.68	-1.6	.66	.38	76.7	70.1	S25
6	23	30	.23	.46	.82	-.7	.68	-.9	.56	.35	83.3	77.8	S6

Fig. 7. Item statistics: Measure order.

This program has ranked the difficulty level of grain of the most difficult grains with the easiest grains. A high logit (measure) value indicates that the item in question has a high degree of difficulty. This correlates with the total score (number of correct answerers) with a higher measure value.

The Number column shows the easy-to-date ballot number sequence. The column is the rightmost column. The Total Score column is the number of correct answerers on the question in question. The Total Count column is the number of answers to the question in question. The Measure column is the level of difficulty of items in the relevant numbers. The MNSQ and ZSTD Infit-outfit columns indicate whether the items are fit or in accordance with the Rasch model. Column PT-Measure Corr or Point Measure Correlation, which is point-biserial correlation in classical test theory. This parameter shows the power of item discrimination.

This measure value is also arranged as a Z value where the value will usually range from -3 to +3. But the value in this

case logit values above 2 or below -2 can already be considered as extreme values. Sumintono and Widhiarso [35] provide guidance in assessing these items into four categories, namely: Measure value <-1 = very easy item; Measure value -1 to.d. 0 = easy items; Measure value 0 to 1 = difficult item; Measure value > 1 = very difficult item.

G. Analyzing Deception /Options

All software displays this result even though the column display varies. For Winstep leveling the deception function or the option for each bullet can be seen in "output tables number 13 from the Figure 7. Item: measure". Its form is as in the following Figure 8. So that the analyzed data can inform each option, the analyzed data is typed completely in the form of a politomus (ABCD, ABCD, or 1234) not dichotomous data (True-False / 0-1).

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE ABILITY	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	Item
46	D	0	30	100	1.59	.16		.00	S46
13	D	0	3	10	-.01	.09	.2	-.62	S13
	B	0	24	80	1.75	.16	1.2	.38	
	A	1	3	10	1.89	.08	1.0	.12	
48	D	0	1	3	-.17		.1	-.38	S48
	0	0	1	3	.49		.3	-.24	
	A	0	20	67	1.68	.14	1.1	.15	
	B	1	8	27	1.72	.42	1.9	.09	
8	D	0	21	70	1.37	.19	1.0	-.39	S8
	C	1	9	30	2.10	.23	.9	.39	
27	B	0	21	70	1.35	.20	1.0	-.42	S27
	D	1	9	30	2.14	.13	.7	.42	

Fig. 8. Item category/option/distractor frequencies: Measure order.

Item number 46, all students chose Option D, for choices A, B, C, none of the students chose it. How to fix it is checked its lattice. Are KD demands in terms of everything in accordance with the relevant grain items? Is the material being asked is material / important? If the case is compatible with the problems and demands of the KD, the points are corrected in accordance with the KD demands in the grid.

For items number 13, answer D is only selected by 3 students (10%). Many students chose B (23 students / 79%). Students who choose Option A have 3 students (10%). Students who choose C are not available. From data such as this for the number 13 item, the revision must be C and check the answer. The correct key is D or B? Etc.

H. Analyzing Rubric Scale

The software that displays rubric analysis is only the Winsteps program. To analyze the rubric, click on "Output

tables number 3.2 Rating Scaledarabables above." Its form is as in Figure 9 below.

TABLE VI. EXAMPLE OF SCORING GUIDELINES CONCOCTED

Score	1	2	3	4	5
Spelling	There are more than 10 spelling mistakes.	There are 6-10 spelling mistakes.	There are 3-5 spelling mistakes.	There are 1-2 spelling mistakes.	No spelling errors found.

Criteria: Andrich Threshold between score = 1.4 - 5.0 logit.  
 Less than 1.4 logit = rubric combined.  
 More than 5.0 logit = rubric separated.

SUMMARY OF CATEGORY STRUCTURE, Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRG	SAMPLE EXPECT	INFIIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE
1	1	67	10	-.08	-.07	1.00	1.05	NONE	( -1.70)
2	2	80	12	.14	.13	.96	.95	-.14	-.65
3	3	84	12	.27	.30	.88	.84	-.17	-.06
4	4	225	33	.52	.48	.95	1.08	-.60	.58
5	5	229	33	.70	.72	1.05	1.03	.58	( 1.90)
MISSING		15	2	.71					

Fig. 9. Results of Andrich Threshold analysis.

Note:  
 Ideal OBSERVED AVERAGE - increase value  
 NONE = 0.0 logit  
 Difference in scores 1 and 2 = (0.00) - (-0.14) = -0.14 (<1.4)

Difference between scores 2 and 3 = (-0.14) - (+0.17) = 0.31 (<1.4)  
 Etc. ...

I. Transformation of Scores

Software that displays score transformation is only two programs namely Winsteps and ConQuest. For Winsteps, you

can click on "output tables number 20. Tabled score above." Its form is as in Figure 10 below.

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-5.85E	1.84	14	-.98	.40	28	1.40	.45
1	-4.61	1.03	15	-.82	.40	29	1.61	.46
2	-3.85	.75	16	-.66	.40	30	1.82	.47
3	-3.38	.63	17	-.50	.40	31	2.05	.48
4	-3.03	.56	18	-.34	.40	32	2.29	.50
5	-2.74	.52	19	-.18	.40	33	2.56	.53
6	-2.48	.49	20	-.01	.40	34	2.85	.56
7	-2.26	.47	21	.15	.41	35	3.19	.61
8	-2.05	.45	22	.32	.41	36	3.60	.67
9	-1.85	.44	23	.49	.41	37	4.13	.79
10	-1.67	.43	24	.66	.42	38	4.96	1.07
11	-1.49	.42	25	.84	.42	39	6.26E	1.87
12	-1.31	.41	26	1.02	.43			
13	-1.15	.41	27	1.21	.44			

CURRENT VALUES, UMEAN=.0000 USCALE=1.0000  
 TO SET MEASURE RANGE AS [0-100, UMEAN=48.3059 USCALE=8.2534]  
 TO SET MEASURE RANGE TO MATCH RAW SCORE RANGE, UMEAN=18.8393 USCALE=3.2188  
 Predicting Score from Measure: Score = Measure \* 4.3636 + 19.4783  
 Predicting Measure from Score: Measure = Score \* .2191 + -4.2683

Fig. 10. Measure on test.

The figure above informs about the score, if wrong all the scores are -5.85, Correct 1 score -4.61, correct 2 score -3.38, correct 3 score -3.38, correct 4 score -3.03, correct 5 the score is -2.74, and so on. If true, all the scores are 6.26. The value of logit like this is not widely understood by parents of students.

The winsteps program provides a score scale of 0-100 as shown above. The code is copied (CTRL + C) then transferred to the first display code as in the following check box, Figure 11.

```

TESW.txt - Notepad
File Edit Format View Help
&INST
  TITLE = TES1C
  PERSON = Person ; persons are ...
  ITEM = Item ; items are ...
  ITEM1 = 13 ; column of response to first item in data record
  NI = 50 ; number of items
  NAME1 = 1 ; column of first character of person identifying label
  NAMELEN = 1 ; length of person label
  XWIDE = 1 ; number of columns per item response
  CODES = 0ABCD ; valid codes in data file
  UIMEAN = 0 ; item mean for local origin
  USCALE = 1 ; user scaling for logits
  UDECIM = 2 ; reported decimal places for user scaling
  UMEAN=48.3059;
  USCALE=8.2534;
  KEY1 = CABDDBCBCDABAACABCDABCDDBCDABDABACCCBCABABDBCBABCD ; key for MCQ scoring

&END
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
    
```

Fig. 11. Change in score range.

Click the arrow, the results are as follows Figure 12.

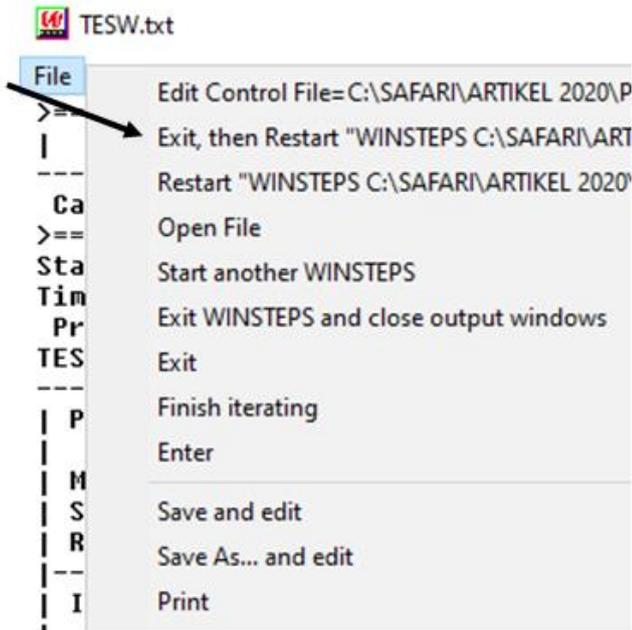


Fig. 12. Next step Change in score range.

SCORE	MEASURE	S. E.	SCORE	MEASURE	S. E.	SCORE	MEASURE	S. E.
0	.00E	15.20	14	40.21	3.34	28	59.88	3.69
1	10.28	8.50	15	41.55	3.32	29	61.56	3.77
2	16.52	6.19	16	42.88	3.31	30	63.33	3.87
3	20.39	5.21	17	44.21	3.31	31	65.21	4.00
4	23.31	4.65	18	45.53	3.31	32	67.22	4.15
5	25.72	4.29	19	46.86	3.32	33	69.41	4.35
6	27.81	4.03	20	48.20	3.33	34	71.84	4.62
7	29.69	3.85	21	49.55	3.35	35	74.63	5.00
8	31.41	3.71	22	50.92	3.38	36	77.99	5.57
9	33.03	3.60	23	52.32	3.41	37	82.37	6.55
10	34.56	3.52	24	53.75	3.45	38	89.21	8.82
11	36.03	3.45	25	55.21	3.50	39	100.00E	15.41
12	37.45	3.40	26	56.71	3.55			
13	38.84	3.37	27	58.27	3.61			

CURRENT VALUES, UMEAN=48.3059 USCALE=8.2534  
 TO SET MEASURE RANGE AS 0-100, UMEAN=48.3059 USCALE=8.2534  
 TO SET MEASURE RANGE TO MATCH RAW SCORE RANGE, UMEAN=18.8393 USCALE=3.2188  
 Predicting Score from Measure: Score = Measure \* .5287 + -6.0613

Fig. 13. Changes in the Score Range from 0-100.

IV. CONCLUSION

Based on the analysis of all the software above, the results can be concluded as follows. Most of the above software must write the first syntax, such as Parscale, ConQuest, Mplus, IRTPRO 3, including Bilog-MG, except Winsteps. Of all the software, the presentation of the results of its analysts with the most information is Winsteps. So, the Winsteps program is very practical to be used by teachers in schools (SD / MI, SMP / MTs, SMA / MA / SMK) to diagnose students' abilities and analyze items on aspects of: (1) knowledge that includes data from answers to form questions (descriptions, multiple choice, short answer, answer, match, True-False); (2) skills that include

data from observations / observations on performance appraisals, assignments / projects, work / products, and portfolios; and (3) attitude / character which includes data from direct and indirect observations using questionnaires. One computer program that is easily used by teachers in schools is the Winsteps program. Student version of the Wisteps Program / demo with the name Minitap can be downloaded for free to analyze a maximum of 25 items and 75 students. The website (website) is <http://www.winsteps.com/ministep.htm>.

## V. SUGGESTION

Based on the results of the study above, there are important suggestions like the following. Especially for teachers who are conducting distance learning at this time, in addition to using classical test theory, it is time now to diagnose student abilities and items used in modern test theory, the Rasch model using Winsteps software. In order to maximize the results of the analysis there are a number of suggestions: (1) understand the operational manual correctly; (2) type the data correctly; (3) for multiple choice form data, data should be typed complete not 0-1 (0 = false 1 = true), for example, ABCD, abcd, 1234 or ABCDE, abcde, 12345, the Winsteps program will help analyze each option / distractor. This greatly facilitates the teacher to revise options that do not work on each item; (4) to obtain a large sample, the teacher can combine the same data with student data between classes, between semesters, etc., which is very useful for developing question banks in schools; (5) so that the data analyzed is more accurate, (like an ill body, it is easy to know the disease to determine the right medicine so that it can heal quickly and be healthy), the items analyzed must be items that are arranged according to the grid correctly [38].

## ACKNOWLEDGMENTS

The author would like to thank Mr. Linacre and Wright for creating the Winsteps program whose software was used for this study. In addition, the authors would like to thank friends at the Center for Assessment and Learning Balitbang Kemdikbud who have provided input, suggestions, and comments, so that the realization of this article.

## REFERENCES

- [1] Safari, "The Influence of Computer Availability at Home and at School for Students on 2019 UNBK Results," ICEAP Proceeding B. 2019, pp. 284–291, 2019.
- [2] G. Basilaia and D. Kvavadze, "Transition to online education in schools during a SARS-CoV-2 coronavirus (COVID-19) pandemic in Georgia," *Pedagog. Res.*, vol. 5, no. 4, 2020.
- [3] Safari, "Students' Perception of Teacher Guidance on Reading Learning Based on Results of PISA 2018," *IJEA Indones. J. Educ. Assess.*, vol. 3, no. 1, pp. 32–41, 2020.
- [4] A.J. Nitko, *Educational assessment of students*. ERIC, 1996.
- [5] L.R. Aiken, *Psychological testing and assessment*. Pearson Education India, 2009.
- [6] Safari, *Analisis Data Dengan Program Komputer: Berdasarkan Teori Tes Klasik (EXCEL, SPSS, ITEMAN) dan Modern (WINSTEPS, QUEST, BILOG, FACETS, LISREL, AMOS, MPLUS)*. Jakarta: Universitas Negeri Jakarta, 2020.
- [7] J.M. Linacre, "What do infit and outfit, mean-square and standardized mean," *Rasch Meas. Trans.*, vol. 16, no. 2, p. 878, 2002.
- [8] T. I Deviana, B. Hayat, and B. Suryadi, "Validasi instrumen social provision scale pada siswa Sekolah Menengah Atas: studi dengan penerapan rasch model," 2020.
- [9] J.M. Linacre and B.D. Wright, "Winsteps," URL <http://www.winsteps.com/index.htm> [accessed 2013-06-27][WebCite Cache], 2000.
- [10] M.Z. F. E. Muraki, R. Mislevy, and D. Bock, *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc, 2003.
- [11] E. Muraki and R.D. Bock, "PARSCALE (Version 4.1)[Computer software]," Lincolnwood, Sci. Softw. Int., 2002.
- [12] K.T. Han, "WinGen3: Windows software that generates IRT parameters and item responses [computer program]," Amherst, MA Cent. Educ. Assessment, Univ. Massachusetts Amherst, 2007.
- [13] K.T. Han and R.K. Hambleton, "User's Manual: WinGen (Center for Educational Assessment Report No. 642)," Amherst, MA Univ. Massachusetts, Sch. Educ., 2007.
- [14] L.M. Rudner, *PARAM Calibration Software Logistic IRT Models* (freeware). 2012.
- [15] F. Niklas, C. Nguyen, D.S. Cloney, C. Tayler, and R. Adams, "Self-report measures of the home learning environment in large scale research: Measurement properties and associations with key developmental outcomes," *Learn. Environ. Res.*, vol. 19, no. 2, pp. 181–202, 2016.
- [16] R. Guyer and N.A. Thompson, "User's Manual for Xcalibre item response theory calibration software." Assessment Systems Corporation St. Paul, MN, 2011.
- [17] D. Rizopoulos, "ltm: An R package for latent variable modeling and item response theory analyses," *J. Stat. Softw.*, vol. 17, no. 5, pp. 1–25, 2006.
- [18] I. Partchev, G. Maris, and T. Hattori, "Irtos: A collection of functions related to item response theory (IRT). R package version 0.2. 1." 2017.
- [19] R.C. Team, "R: a language and environment for statistical computing. Version 3.2. 2. Vienna: R Foundation for Statistical Computing; 2014." 2017.
- [20] J.P. Meyer, *jMetrik*, version 4.1.0. 2017.
- [21] L. Cai, S.H.C. Du Toit, and D. Thissen, "IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]," Chicago, Sci. Softw. Int., 2011.
- [22] C.R. Houts and L. Cai, "flexMIRT: Flexible multilevel item factor analysis and test scoring user's manual," Seattle, WA Vector Psychom. Gr., 2016.
- [23] B.O. Muthén, L.K. Muthén, and T. Asparouhov, *Regression and mediation analysis using Mplus*. Muthén & Muthén Los Angeles, CA, 2017.
- [24] M. Pomplun, M.D.H. Omar, and M. Custer, "A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model," *Educ. Psychol. Meas.*, vol. 64, no. 4, pp. 600–616, 2004.
- [25] R.K. Hambleton and R. ROVINELL, "Fortran-Iv Program For Generating Examine Response Data From Logistic Test Models," *Behavioral Science*, Vol. 18, No. 1. Int Soc Systems Science 428 East Preston St, Baltimore, MD 21202 USA, p. 74, 1973.
- [26] F.B. Baker, "GENIRV: A program to generate item response vectors," Unpubl. manuscript, Univer-sity Wisconsin, Lab. Exp. Des. Madison, 1982.
- [27] E. Muraki, "RESGEN item response generator," *ETS Res. Rep. Ser.*, vol. 1992, no. 1, pp. i–56, 1992.
- [28] H. Fang and G. Johanson, "WINIRT: A Windows-based item response theory data generator with an equating and differential item functioning simulation guide," 2005.
- [29] M. Yoes, "PARDSIM," *Comput. software*. Saint Paul, MN Assess. Syst. Corp., 1997.
- [30] R.D. Bock and M. Aitkin, "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, vol. 46, no. 4, pp. 443–459, 1981.
- [31] N.A. Volodin and R.J. Adams, "Identifying and estimating a D-dimensional item response model," 1995.
- [32] M.L. Wu, R.J. Adams, and M.R. Wilson, "ConQuest: Multi-aspect test software [computer program]," Camberwell, Vic. Aust. Coun. Educ. Res., 1997.
- [33] R. Tate, "A comparison of selected empirical methods for assessing the structure of responses to test items," *Appl. Psychol. Meas.*, vol. 27, no. 3, pp. 159–203, 2003.

- [34] T. Sims, "Comparison of IRTPRO 3 and Mplus 7 for multidimensional item response item parameter and examinee ability estimation," 2017.
- [35] B. Sumintono and W. Widhiarso, Aplikasi pemodelan rasch pada assessment pendidikan. Trim komunikata, 2015.
- [36] W.J. Boone, J.R. Staver, and M.S. Yale, Rasch analysis in the human sciences. Springer, 2013.
- [37] S. Alagumalai, D.D. Curtis, and N. Hungi, Applied Rasch measurement: A book of exemplars. Springer, 2005.
- [38] Safari, Evaluasi Pendidikan: Penyusunan Kisi-kisi, Penulisan, dan Analisis .Butir Soal. Berdasarkan Kurikulum 2013, Menuju Penilaian Abad 21. Jakarta: Indonesia: Esensi, divisi Penerbit Erlangga, 2019.