KONSTRUKTIVISME, Vol.15 No.1 Januari 2023
ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index

Person and Item Validity and Reliability in Essay Writing Using Rasch Model

Received:
02 January 2023
Accepted:
16 January 2023
Published:
19 January 2023

[1]*Yenni Arif Rahman
[1]English Department, Faculty of Communication and Language
Bina Sarana Informatika University
[1] Jl. Kramat Raya No.98, RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450, Indonesia
E-mail: [1] yeni.yar@bsi.ac.id*

*Corresponding Author

Abstrak— Penelitian ini bertujuan untuk mengkaji reliabilitas dan validitas iteman dan siswa dengan menggunakan model Rasch. Tingkat kemampuan menulis siswa dinilai melalui empat konstruk yang diturunkan menjadi dua puluh empat item pernyataan dalam bentuk rubrik penilaian. Penelitian ini melibatkan 40 siswa Bahasa Inggris sebagai bahasa asing yang telah mengikuti kelas menulis esai di kelas TOEFL iBT dimana sampel penulisan diambil bekerja sama dengan pusat bahasa tempat diadakannya kelas TOEFL iBT. Metode penelitian ini menggunakan model Rasch sebagai pendekatan analisis kuantitatif dengan menggunakan tiga output statitistic data: "*statistical summary output*" untuk mendapatkan angka dan data secara umum, statistik item untuk mendapatkan validitas iteman, dan statistik partisipan untuk mendapatkan validitas partisipan. Hasil reliabilitas tercermin dalam nilai *alpha Cronbach* (*a*) 0,92, reliabilitas iteman 0,60, dan reliabilitas partisipan 0,68 yang menunjukkan kinerja reliabilitas yang "cukup baik". Keempat konstruk penilaian: konten, struktur, diksi, dan mekanik memenuhi kesesuaian item yang diukur dengan OUTPUT MNSQ dan OUTPUT ZSTD walaupun konstruk mekanik penulisan valid dengan catatan. Person fit yang diidentifikasi dengan teknik INFIT MNSQ menunjukkan tujuh siswa misfit yang perlu studi lebih lanjut untuk menemukan sumber misfit.

Kata Kunci: validitas; reliabilitas; menulis essay; model rasch.

Abstract— *This study aims to examine the reliability and validity of items and persons using the Rasch model. The students' writing skills were assessed through four constructs elaborated into twenty-four items statements in the form of a rubric. The participants were 40 EFL learners who had taken an essay writing course in TOEFL iBT class whereas the writing samples were taken in collaboration with a language center wherein the TOEFL iBT class was held. The research method employed the Rasch model as a quantitative analysis approach by using three Ministep software outputs used for data analysis: the "statistical summary output" to obtain figures and data in general, item statistics to obtain item validity, and person statistics to acquire person validity. The results of the reliability were reflected in Cronbach's alpha score (a) 0.92, item reliability 0.60, and person reliability 0.68 which show "acceptable" reliability performance. The four assessment constructs: content, structure, diction, and mechanic fulfill item fit measured by OUTPUT MNSQ dan OUTPUT ZSTD though mechanic constructs pass the item fit with a note. The person fit order identified by INFIT MNSQ shows seven students are misfits which need further assessment to find the source of a misfit.*

Keywords: *validity; reliability; essay writing; rasch model.*

I. INTRODUCTION

Proficient writing skill is one of the markers of students' success in the academic field. For this reason, some English proficiency tests mandatorily incorporate writing skills into their tests. One of the best examples is the iBT TOEFL test which is a benchmark to test students' writing skills in English. This test is also a prerequisite for foreign students who wish to continue their studies at various foreign universities. Therefore, students who intend to continue studying abroad or for other academic purposes pay great attention to developing writing skills by participating in various writing skills improvement programs. At this point, each student's essay output will be assessed by the instructor to appraise whether there has been an increase in students' writing skills regularly. This increase is identified by holding a pretest, while test, or post-test. At this point, the method of identification, the reliability of both student and item measurement, and the validity or accuracy of the measurement hold the strategic point along the continuum of assessment [1], [2], [3].

Related to the writing test measurement, the Rasch model with the help of rubric as item measurement can be utilized to assess students' reliability and validity toward the item and at the same time can measure students' validity of their essay [4]. The accuracy and effectiveness of the Rasch model for measuring responses is a fairly practical choice to measure students' writing skills [5]. This is evidenced by the growing popularity of using the Rasch model in item response measurement (IRT). It is reasonable because the Rasch model itself offers the advantages of quantitative tests which are not found in the classical model [2]. One of them is its ability to predict missing data based on systematic response patterns (scalogram format) [1]. This makes the results of statistical analysis more accurate. In classical models, it is customary to treat missing data with a score of zero, even if the percentage rate of missing data is high, the analysis cannot provide satisfactory conclusions. However, with its predictive ability, Rasch modeling produces the best possible score from the missing data [6].

Several studies have been conducted regarding the validity and reliability of writing assessments using the Rasch model. Tan [7] studied the performances of the writing rating scale toward multiple raters using the Rasch model. This study discusses multiple raters who utilized a revised rating scale (analytical rubric) to discriminate performances for essay scoring. Meanwhile, Erguvan & Dunya [8] focused on their study on the rater severity of instructors using a multi-trait rubric in a freshman composition course. The researcher only found those two studies regarding writing assessment using the Rasch model with a different focus. The lack of research on the validity and reliability of writing assessments using the Rasch model prompted this

ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index

research. Particularly research on the validity and reliability of "holistic rubrics" using the Rash model which has never been done before. The results of research on the performance of the holistic scoring rubric in writing assessment using the Rasch model are needed as a comparison of the effectiveness of the holistic rubric scale as one of the options for writing assessment rubrics.

Regarding the Rasch model, its presence as a new measurement system aims to overcome the limitations of the classical measurement system or Classical Test Theory (CTT) [9], [10], [11]. In the classical measurement, both the person and the item parameters which are the results of the analysis of the item difficulty level and the item discrimination index are group dependent [12]. In terms of difficulty level, the classification of item difficulty level will change when given to different sample groups [13], whereas, in the case of the discrimination index, higher scores tend to be obtained from heterogeneous samples and lower scores are obtained from homogeneous samples [14]. This dependence results in CTT being unable to describe the ability of the sample and limiting test development because it complicates analysis [15], as well as the emergence of theoretical difficulties in applying CTT to several measurement situations, for example when equating or computerized adaptive testing [16]. In modern test theory, item parameters do not change even though they are estimated from different sample groups [17]. This means that modern test theory provides a uniform measurement scale [13], so that sample groups can be tested with a different set of items, according to their level of ability and the scores can be directly compared [18].

Another feature provided by the Rasch model is a probabilistic unidimensional test which states that (1) the easier the question, the more likely it is that students will respond to the question correctly, and (2) the greater the ability students have, the more likely they will answer questions correctly compared to less able students [13]. Therefore, in this model, only one item parameter is known which is the item difficulty level, while the discrimination index parameter is assumed to be equal to one [19].

In general, the person and item fit of the model to the data is a major concern when implementing analysis using modern test approaches [20]. If the data deviates greatly from the Rasch model, the causes need to be considered and the person or item that does not fit may need to be deleted [6], [21]. Therefore, specifically for the Rasch model, there are two types of fit, namely item fit and person fit, which illustrate the measurement validity of the Rasch model [22] and can be used to detect differences between empirical data and Rasch model data [23], [24].

Item fit describes the extent to which the pattern of a person's response to an item is consistent with the responses of other people responding to other items, while person fit indicates

the extent to which a person's pattern of performance on the test is consistent through items that are also responded to by other people [22] [25]. The urgency of item fit and person fit research cannot be negotiated considering that both of them have a strong symmetrical relationship [20], where both play an important role in test construction, especially about evaluation issues and item selection and in making decisions on test scores based on individual response results [26]. Therefore, through item fit, errors that occur during the calibration phase of instrument development can be detected. For example, if there is an item that has a different power parameter that is not good, then the item fit statistic will identify this problem [20]. Meanwhile, person fit can show whether there are deviations in response patterns (leading to scores that are too high or too low) due to cheating, careless responding, lucky guessing, and random responding [27] [28]. This means that respondents who are included in-person fit are only able to answer items correctly when the item has a level of difficulty below their ability.

II. METHOD

Participants in this study were 40 EFL students of Eloquensi English Language Centre who had taken the TOEFL iBT essay writing course. They were students of high school and college who has intermediate or above English proficiency levels. The participants were required to write essays of five paragraphs, assuming that an efficient essay format has been constructed and comprises an introduction, content, and conclusion. This general structure is required to prevent bias in the rater's evaluations caused by the number of paragraphs, which can have a positive or negative impact on the rating.

This study uses the Rasch model as a basis for analysis because it can see the interaction between respondents and items at the same time. In the Rasch model, a score is not seen based on a raw score, but a logit score that reflects the probability of selecting an item in a group of respondents [2], [7]. This is used as an anticipation of the raw score of the Likert rating which is in the form of ordinal which does not have the same interval between the scores. The use of the Rasch model for polytomous data was developed by Andrich while still based on two basic theorems, namely the level of individual ability/agreement and the level of difficulty of items to agree on [29]. The output used for data analysis is output summary statistics (Figure 1.) to obtain reliable information as well as the output of unidimensionality items (Figure 2.) and Fit Order items (Figure 3.) for validity.

In this study, the measurement used was a holistic rubric by Jacob et al [30]. This measurement rubric uses six levels of measurement consisting of proficient, fluent, expanding,

ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index

developing, beginning, and emerging. In the criteria column, it can be observed that in general there are four types of writing ability elements that are assessed, namely content, structure, diction, and mechanics. The first element is content which consists of an introduction, ideas or body paragraphs, and the ability to write ideas logically. The second ability is a structure that not only assesses the ability to apply grammar correctly in sentences but also how one paragraph is composed of various types of sentences (simple, compound, complex, and compound-complex sentences). The third ability assessed is diction which not only assesses the respondent's ability to use vocabulary correctly but also variations in the use of words in one paragraph so that there is no repetition of words in the same paragraph. The last ability to be assessed is writing mechanics which includes the use of punctuation, correct spelling, and capital letters.

TABLE 1. JACOB ET AL HOLISTIC RUBRIC

Rating	Criteria
Proficient	1. Writes single or multiple paragraphs with a clear introduction, fully develop the idea, and presents the idea logically 2. Uses appropriate verb tense and a variety of grammatical and syntactical structures; uses complex sentences effectively; uses smooth transitions 3. Uses varied, precise vocabulary 4. Has occasional errors in mechanics (spelling, punctuation, and capitalization) which do not detract from the meaning
Fluent	1. Writes single or multiple paragraphs with main idea and supporting detail, presents idea logically, though some parts may not fully developed. 2. Uses appropriate verb tense and a variety of grammatical and syntactical structures; errors in the sentence do not detract from meaning; uses transitions 3. Uses varied vocabulary appropriate for the purpose 4. Has few errors in mechanics which do not detract from the meaning
Expanding	1. Organizes ideas in logical or sequential order with some supporting detail; begin to write a paragraph 2. Experiment with a variety of verb tenses, but do not use them consistently; subject/verb agreement errors; use some compound and complex sentences; limited use of transitions 3. Vocabulary is appropriate to purpose but sometimes awkward 4. Use punctuation, capitalization, and most conventional spelling; errors sometimes interfere with meaning
Developing	1. Writes sentences around an idea; some sequencing is present, but may lack cohesion 2. Write in present tense and simple sentences; has difficulty with subject/verb agreement, run-on sentences are common; begin to use compound sentences 3. Uses high-frequency words; may have difficulty with word order; omit endings or words 4. Uses some capitalization, punctuation, and transitional spelling; errors often interfere with meaning

Continuation of table 1

Rating	Criteria
Beginning	1. Begin to convey meaning through writing 2. Write predominantly phrases and patterned or simple sentences 3. Uses limited or repetitious vocabulary 4. Uses temporary (phonetic) spelling
Emerging	1. No evidence of idea development or organization 2. Uses single words, pictures, and patterned phases 3. Copies from model 4. Little awareness of spelling, capitalization, or punctuation

Then the six measurement levels are interpreted into five Likert ratings which can be seen in table 2. Interpretation of scores into a Likert scale is required so that the raw scores obtained from the scoring results can be further processed through mini step software.

TABLE 2. RUBRIC RATING SCALE

Scale	*Likert* Score
Proficient	5
Fluent	4
Expanding	3
Developing	2
Emerging & Beginning	1

III. FINDING AND DISCUSSION

Three output data results are used to reveal the validity and reliability of persons and items from student essays. The first output data used are summary statistics. Then the second output data is item statistics to determine misfit items, and the third output data is person statistics used to determine misfit persons. The data outputs in this study were obtained from the use of Rasch model analysis using mini step software, which is specifically statistical software for Rasch modeling.

KONSTRUKTIVISME, Vol.15 No.1 Januari 2023
ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index

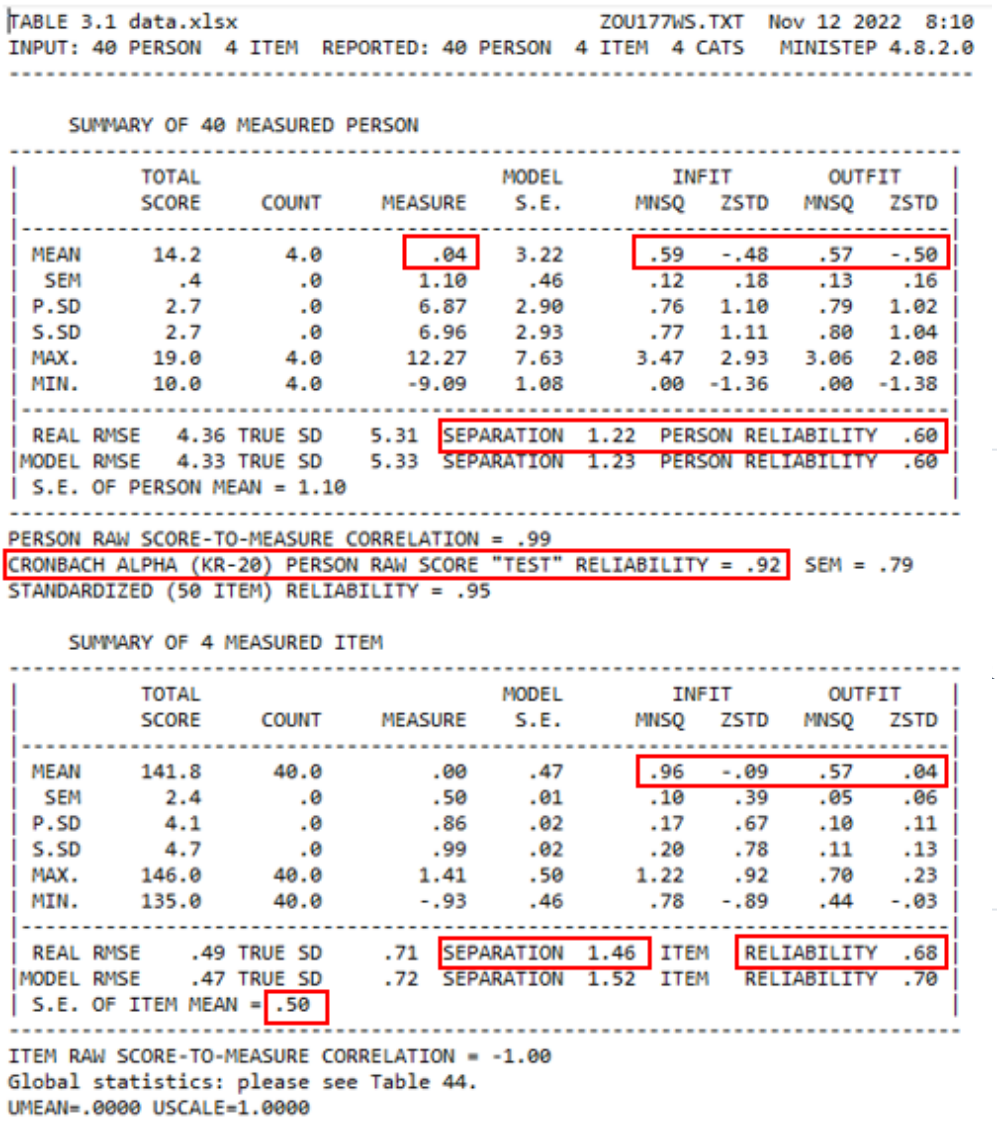


TABLE 3.1 data.xlsx ZOU177WS.TXT Nov 12 2022 8:10
INPUT: 40 PERSON 4 ITEM REPORTED: 40 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.0

SUMMARY OF 40 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	14.2	4.0	.04	3.22	.59	-.48	.57	-.50
SEM	.4	.0	1.10	.46	.12	.18	.13	.16
P.SD	2.7	.0	6.87	2.90	.76	1.10	.79	1.02
S.SD	2.7	.0	6.96	2.93	.77	1.11	.80	1.04
MAX.	19.0	4.0	12.27	7.63	3.47	2.93	3.06	2.08
MIN.	10.0	4.0	-9.09	1.08	.00	-1.36	.00	-1.38

REAL RMSE 4.36 TRUE SD 5.31 SEPARATION 1.22 PERSON RELIABILITY .60
MODEL RMSE 4.33 TRUE SD 5.33 SEPARATION 1.23 PERSON RELIABILITY .60
S.E. OF PERSON MEAN = 1.10

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .92 SEM = .79
STANDARDIZED (50 ITEM) RELIABILITY = .95

SUMMARY OF 4 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	141.8	40.0	.00	.47	.96	-.09	.57	.04
SEM	2.4	.0	.50	.01	.10	.39	.05	.06
P.SD	4.1	.0	.86	.02	.17	.67	.10	.11
S.SD	4.7	.0	.99	.02	.20	.78	.11	.13
MAX.	146.0	40.0	1.41	.50	1.22	.92	.70	.23
MIN.	135.0	40.0	-.93	.46	.78	-.89	.44	-.03

REAL RMSE .49 TRUE SD .71 SEPARATION 1.46 ITEM RELIABILITY .68
MODEL RMSE .47 TRUE SD .72 SEPARATION 1.52 ITEM RELIABILITY .70
S.E. OF ITEM MEAN = .50

ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

FIGURE 1. SUMMARY STATISTICS

Summary statistics in figure 1 provides general information about the respondents and the instruments used as well as the interactions between person and item [2], [7], [8]. Table 1 tells that person measure = 0.04 which indicates the average score of respondents in the essay writing instrument. The average score that is more than logit 0.0 shows the tendency of respondents who can meet the standards of the existing rubric. Cronbach's alpha score is used to measure reliability, namely the interaction between person and item as a whole. Figure 1 tells Cronbach alpha score = 0.92. Person and item reliability both show how far measurements produce the same information. In other words, if the measurement is carried out by another party, it will not produce too much different result. The differences that appear are interference that can still be tolerated.

KONSTRUKTIVISME, Vol.15 No.1 Januari 2023
ISSN: 1979-9438 (Print) / 2442-2355 (Online)
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index
Email: konstruktivisme@unisbablitar.ac.id

However, if there are striking differences in the results in the same sample with different researchers, several things can be examined, namely: similarity over time (stability), parallel instruments (equivalence), elements in the instrument (internal consistency), and rater agreement.

Figure 1. also shows other overt findings regarding the person and item reliability. The finding of person reliability in figure 1 displays a score of 0.60 and item reliability in figure 1 produces a score of 0.68. Other data that can be used to measure person and item reliability are the INFIT MNSQ and OUTFIT MNSQ. In figure 1 it can be seen in the person table the scores for both are 0.96 and 0.57 respectively. Meanwhile, the ideal score is 1.00 where the closer to the ideal score, the better. And for the INFIT ZSTD and OUTFIT ZSTD scores, based on figure 1, they are -0.48 and -0.50. While the ideal score is 0.0 where the closer to the ideal score, the better the quality is The grouping of persons and items can be identified from the separation score. The greater the separation score, the better the quality of the instrument in terms of all respondents and items because it can identify groups of respondents and groups of items.

TABLE 10.1 data.xlsx ZOU744WS.TXT Nov 14 2022 11: 7
INPUT: 40 PERSON 4 ITEM REPORTED: 40 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.0

PERSON: REAL SEP.: 1.22 REL.: .60 ... ITEM: REAL SEP.: 1.46 REL.: .68

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
1	143	40	-.24	.47	1.22	.92	.70	.23	A .91	.89	80.0	83.4	Content
2	135	40	1.41	.46	1.00	.06	.61	-.03	B .88	.88	85.0	81.4	Structure
4	146	40	-.93	.50	.84	-.45	.53	-.01	b .88	.89	87.5	86.5	Mechanic
3	143	40	-.24	.47	.78	-.89	.44	-.03	a .91	.89	85.0	83.4	Diction
MEAN	141.8	40.0	.00	.47	.96	-.1	.57	.0			84.4	83.7	
P.SD	4.1	.0	.86	.02	.17	.7	.10	.1			2.7	1.8	

TABLE 10.3 data.xlsx ZOU744WS.TXT Nov 14 2022 11: 7
INPUT: 40 PERSON 4 ITEM REPORTED: 40 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.0

FIGURE 2. ITEM STATISTICS

The second output data generated is item statistics in figure 2 which displays the four items used as parameters for assessing student essays, namely content, structure, mechanics, and diction. Figure 2 also provides information on misfit items sorted from the most inappropriate (top). To check fit and misfit items, one can use the INFIT MNSQ score of each item; the average score and standard deviation are added up, then compared, and a logit score that is greater than this score indicates a misfit item. Figure 3 displays the number of logit items from MEAN and P.SD: 0.96 + 0.17 = 1.13.

KONSTRUKTIVISME, Vol.15 No.1 Januari 2023
ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisablitar.ac.id/index.php/konstruktivisme/index

TABLE 6.1 data.xlsx ZOU177WS.TXT Nov 12 2022 8:10
INPUT: 40 PERSON 4 ITEM REPORTED: 40 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.0

PERSON: REAL SEP.: 1.22 REL.: .60 ... ITEM: REAL SEP.: 1.46 REL.: .68

PERSON STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
27	13	4	-3.21	1.18	3.47	2.93	3.06	1.77	A .33	.29	25.0	74.2	S27
35	19	4	12.27	1.25	1.82	1.23	2.64	1.59	B -.63	.39	50.0	76.7	S35
40	18	4	10.97	1.08	1.94	2.14	2.18	2.08	C -.68	.38	25.0	65.4	S40
30	14	4	-1.99	1.08	1.94	2.05	2.16	2.02	D -.68	.38	25.0	65.5	S30
20	14	4	-1.99	1.08	1.57	1.39	1.75	1.46	E -.27	.38	25.0	65.5	S20
18	11	4	-7.81	1.23	1.53	.89	1.45	.74	F -.16	.39	50.0	76.4	S18
24	11	4	-7.81	1.23	1.53	.89	1.45	.74	G -.16	.39	50.0	76.4	S24
10	17	4	9.73	1.21	1.15	.45	.99	.32	H .16	.28	75.0	75.1	S10
37	13	4	-3.21	1.18	1.11	.38	.96	.27	I .16	.29	75.0	74.2	S37
1	14	4	-1.99	1.08	.73	-.66	.67	-.65	J .68	.38	75.0	65.5	S1
21	10	4	-9.09	1.08	.73	-.67	.67	-.65	K .68	.38	75.0	65.4	S21
22	10	4	-9.09	1.08	.73	-.67	.67	-.65	L .68	.38	75.0	65.4	S22
26	14	4	-1.99	1.08	.73	-.66	.67	-.65	M .68	.38	75.0	65.5	S26
6	13	4	-3.21	1.18	.67	-.48	.53	-.31	N .63	.29	75.0	74.2	S6
5	19	4	12.27	1.25	.46	-.82	.36	-.70	O .94	.39	100.0	76.7	S5
7	19	4	12.27	1.25	.46	-.82	.36	-.70	P .94	.39	100.0	76.7	S7
15	19	4	12.27	1.25	.46	-.82	.36	-.70	Q .94	.39	100.0	76.7	S15
16	11	4	-7.81	1.23	.44	-.83	.34	-.73	R .94	.39	100.0	76.4	S16
19	11	4	-7.81	1.23	.44	-.83	.34	-.73	S .94	.39	100.0	76.4	S19
23	11	4	-7.81	1.23	.44	-.83	.34	-.73	T .94	.39	100.0	76.4	S23
25	11	4	-7.81	1.23	.44	-.83	.34	-.73	t .94	.39	100.0	76.4	S25
28	11	4	-7.81	1.23	.44	-.83	.34	-.73	s .94	.39	100.0	76.4	S28
2	12	4	-5.45	1.87	.04	-.90	.04	-.95	r .00	.21	100.0	92.5	S2
3	12	4	-5.45	1.87	.04	-.90	.04	-.95	q .00	.21	100.0	92.5	S3
4	12	4	-5.45	1.87	.04	-.90	.04	-.95	p .00	.21	100.0	92.5	S4
13	12	4	-5.45	1.87	.04	-.90	.04	-.95	o .00	.21	100.0	92.5	S13
17	12	4	-5.45	1.87	.04	-.90	.04	-.95	n .00	.21	100.0	92.5	S17
29	12	4	-5.45	1.87	.04	-.90	.04	-.95	m .00	.21	100.0	92.5	S29
8	16	4	4.59	7.63	.00	-1.36	.00	-1.38	l .00	.05	100.0	99.6	S8
9	16	4	4.59	7.63	.00	-1.36	.00	-1.38	k .00	.05	100.0	99.6	S9
11	16	4	4.59	7.63	.00	-1.36	.00	-1.38	j .00	.05	100.0	99.6	S11
12	16	4	4.59	7.63	.00	-1.36	.00	-1.38	i .00	.05	100.0	99.6	S12
14	16	4	4.59	7.63	.00	-1.36	.00	-1.38	h .00	.05	100.0	99.6	S14
31	16	4	4.59	7.63	.00	-1.36	.00	-1.38	g .00	.05	100.0	99.6	S31
32	16	4	4.59	7.63	.00	-1.36	.00	-1.38	f .00	.05	100.0	99.6	S32
33	16	4	4.59	7.63	.00	-1.36	.00	-1.38	e .00	.05	100.0	99.6	S33
34	16	4	4.59	7.63	.00	-1.36	.00	-1.38	d .00	.05	100.0	99.6	S34
36	16	4	4.59	7.63	.00	-1.36	.00	-1.38	c .00	.05	100.0	99.6	S36
38	16	4	4.59	7.63	.00	-1.36	.00	-1.38	b .00	.05	100.0	99.6	S38
39	16	4	4.59	7.63	.00	-1.36	.00	-1.38	a .00	.05	100.0	99.6	S39
MEAN	14.2	4.0	.04	3.22	.59	-.5	.57	-.5			84.4	83.7	
P.SD	2.7	.0	6.87	2.90	.76	1.1	.79	1.0			24.8	13.0	

FIGURE 3. PERSON STATISTICS

The same technique in item checking is used to find out the misfit of persons by looking at the INFIT MNSQ score of each person; the average score and standard deviation are added up, then compared, a logit score that is greater than this score indicates a person who is a misfit. Total logit items from MEAN and P.SD: 0.59 + 0.76= 1.35.

The summary statistics in figure 1 provide general information about the respondents and the instruments as well as the interactions between person and item. The results of the following

analysis also describe how person and item validity and reliability in groups interact with each other. Summary statistics in figure 1 show that person measure = 0.04 which indicates the average score of respondents in the essay writing instrument. The average person measure score is 0.04 > logit 0.0 so it shows the tendency of respondents who can fulfill the ability indicators listed in the existing rubric.

Figure 1. shows that the Cronbach alpha score = 0.92. and according to table 3. It can be inferred that the interaction between a person (identified from Cronbach's alpha score) shows "excellent" results. In other words, the reliability between a person and an item is “excellent” in general.

TABLE 3. CRONBACH ALPHA

Cronbach’s alpha	Interpretation of Internal Consistency
$a > 0,8$	Excellent
$0,7 < a \leq 0,8$	Good
$0,6 < a \leq 0,7$	Acceptable
$0,5 < a \leq 0,6$	Questionable
$a < 0,5$	Poor

Summary statistics also display person and item reliability. The interpretation of both person and item reliability is then consulted to the Cronbach alpha in table 3. Since the score of person and item reliability is 0.60 and 0.68 respectively then both scores are within the range of $0.6 < a \leq 0.7$ which is interpreted as “acceptable”.

The ideal score of INFIT MNSQ and OUTPUT MNSQ is 1.00. Meanwhile, INFIT MNSQ and OUTFIT MNSQ scores in figure 1 are 0.96 and 0.57 respectively. So it is inferred that the reliability is “good” because it is closer to the ideal score of 1. The score of INFIT ZSTD and OUTFIT ZSTD, based on figure 1 are 0.48 and 0.50. While the ideal score is 0.0 where the closer to the ideal score, the better the quality. Meanwhile, the two scores move closer to 0.0 which can be concluded that the reliability of the person and item is "good".

The clustering of persons and items can be identified from the separation score. The greater the separation score, the better the quality of the instrument is in terms of all respondents and items. In other words, it can identify a wider group of subjects (able - unable) and item groups (difficult - easy). In figure 1 the separation score is 1.46. Then the formula to calculate stratum separation is H = [(4 X separation) + 1] / 3. So H = [(4 X 1.46) + 1] / 3= 2.28 (rounded 2) which means there are 2 groups of questions. However, the individual separation index recorded at 2.28 is still weak according to Fisher [31] because it can only produce two levels/strata of respondents' ability involved in the study. The condition of not being able to separate individuals into more

than two strata may be due to the low quality of the items or individual separation. However, the "acceptable" item reliability indicates that this instrument is sufficient and may be used to conduct real research.

TABLE 4. PERSON AND ITEM RELIABILITY

Cronbach's alpha	Interpretation	Item Reliability	Interpretation	Person Reliability	Interpretation	Summary
0,92	Excellent	0,68	Acceptable	0,60	Acceptable	Reliable

Table 4 accumulates the conclusions of the item and person reliability whose final results are contained in the summary. Table 4 shows "excellent" Cronbach's alpha score and the interpretation of item reliability and person reliability are "acceptable". The accumulation of these three criteria concludes that the items and persons are "reliable".

Item Validity/ Item Fit

In the Rasch model, the validity test is known as the unidimensionality item [2]. Undimensionality items are used to evaluate whether the instrument can measure what should be measured. One method to seek the item unidimensionality/item fit is by measuring three criteria: the outfit means-square, outfit z-standard, and point measure correlation [21], [23]. According to Linacre [32], two output statistics can be used to assess item fit in the Rasch model, namely infit (inlier-sensitive or information-weighted fit) and outfit (outlier-sensitive or information-weighted fit). Both outputs are generally reported in the form of the mean squared (MNSQ) and z-standardized (ZSTD). MNSQ is the average of the residuals square for an item, and ZSTD (standard form) is a transformation from the mean squared value with sample size correction [23]. Therefore, in this study, to identify whether the items are proven to be fit or misfit, then Outfit Mean Square statistics (MNSQ) output needs to be interpreted. Table 5 [29] provides the score range of those two criteria.

TABLE 5. THE ITEM FIT SCORE RANGE

Criteria	Score Range
Outfit mean square (MNSQ)	0,5 < *MNSQ* < 1,5
Outfit Z-*standard* (ZSTD)	-2,0 < *ZSTD* < +2,0

Further, To interpret item fit via MNSQ, Linacre [33] suggested rules of the thumb to assess the implication item fit toward measurement, which is MNSQ > 2,0 which means undermine the measurement; 1,5 < MNSQ ≤ 2,0 which means does not have the significance to the measurement; 0,5 ≤ MNSQ ≤ 1,5 which is valuable to measurement; dan MNSQ < 0,5 which is interpreted as useless for the measurement. Based on figure 2 and compare with the item fit score range in table 5 the result of item fit is presented in table 6.

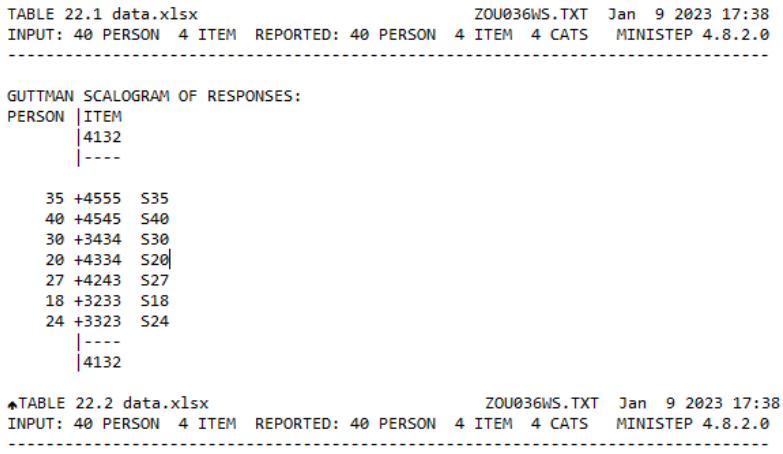
TABLE 6. ITEM FIT RESULT

Number	Item	Outfit		Interpretation
		MNSQ	*ZFTD*	
1	Content	0,70	0,23	Valid
2	Structure	0,61	-0,03	Valid
3	Diction	0,53	-0,01	Valid
4	Mechanic	0,44	-0,03	Valid with note

Person Validity/Person Fit

The second output data generated is item statistics in figure 3 which displays the personal statistics of 40 students. Figure 3 also provides information about person misfits which are sorted from the least suitable (at the top). The same technique in item checking is used to find out the person misfit by analyzing the INFIT MNSQ score of each person; the average score and standard deviation are added up, then compared, a logit score that is greater than this score indicates a person who is a misfit. Total logit items from MEAN and P.SD: 0.59 + 0.76= 1.35. There are 7 students who are greater than the number of logit items, namely the score of person S27 (3.47), S35 (1.82), S40 (1.94), S30 (1.94), S20 (1.57), S18 (1.53), S24 (1.53) INFIT MNSQ. In other words, 7 students have not met one or more items (content, mechanic, structure, and diction) due to some reasons mentioned by Karabatsos (2003) and Meijer (1996). This also means the ability of 7 students who have different response patterns than the rest of the students cannot be predicted by the model/rubric [34]. Whereas through the pattern of responses, the accuracy of the responses of each student for each item can be depicted [2]. The other finding in figure 3 also states that the rest 33 students (figure 3) are below total logit items (1.35) which indicates the person fit. The person fit means the 33 students have relatively logical response patterns. One method to identify the causes of a person misfit is by utilizing the Gutman matrix or scalograms. The Guttman matrix can provide valuable information because the items have been sorted from the hardest to the easiest item (1: content, 2: structure, 4: mechanic, 3: diction). This matrix can also show unidimensionality data [16]. Below is the identification of 7 students (S20, S35, S40, S30, S27, S18, S24) who are classified as person misfits based on the Guttman matrix:

KONSTRUKTIVISME, Vol.15 No.1 Januari 2023
ISSN: 1979-9438 (Print) / 2442-2355 (Online)
DOI: 10.35457/konstruk.v15i1.2618
Website: https://ejournal.unisbablitar.ac.id/index.php/konstruktivisme/index



```
TABLE 22.1 data.xlsx                          ZOU036WS.TXT  Jan  9 2023 17:38
INPUT: 40 PERSON  4 ITEM  REPORTED: 40 PERSON  4 ITEM  4 CATS   MINISTEP 4.8.2.0
-------------------------------------------------------------------------------

GUTTMAN SCALOGRAM OF RESPONSES:
PERSON |ITEM
       |4132
       |----

    35 +4555  S35
    40 +4545  S40
    30 +3434  S30
    20 +4334  S20
    27 +4243  S27
    18 +3233  S18
    24 +3323  S24
       |----
       |4132

▲TABLE 22.2 data.xlsx                         ZOU036WS.TXT  Jan  9 2023 17:38
INPUT: 40 PERSON  4 ITEM  REPORTED: 40 PERSON  4 ITEM  4 CATS   MINISTEP 4.8.2.0
-------------------------------------------------------------------------------
```

FIGURE 4. GUTTMAN MATRIX

Based on the Guttman matrix/Guttman scalogram of responses, the students' numbers 20, 35, 40, 30, 27, 18, and 24 are classified as person misfits in the Rasch model. This conclusion is derived from students' unusual response patterns, namely being able to respond on difficult items (the hardest to easiest item, namely 1: content, 2: structure, 4: mechanic, 3: diction).) but cannot respond correctly on easier items. The assumption is based on a definition of the Rasch model which states students with lower abilities have a lower chance to solve more difficult questions/items.

IV. CONCLUSION

The use of the Rasch model in the validation instrument produces more holistic information about the instrument being studied and better meets the definition of measurement. Based on table 4, the Cronbach alpha score is "excellent" which shows the reliability of items and persons is well correlated. In the same table, the person reliability is declared "acceptable" and the results of item reliability are also interpreted as "acceptable" as well (see table 3. Cronbach alpha for criteria).

The item validity which includes the structure, diction, and mechanic is lower than the combined score of MEAN and P.SD which is concluded that the three items are "valid". Whereas the "content" item score is greater than the combined score of MEAN plus P.SD indicates a "misfit item" that needs to be reviewed again. Meanwhile, in terms of person validity, there were 7 misfit students, namely S27 (3.47), S35 (1.82), S40 (1.94), S30 (1.94), S20 (1.57), S18 (1.53), S24 (1.53) so the instructor needs to take the necessary classroom action.

This research, however, involved a limited number of participants and this limitation needs to be acknowledged. Further research should pursue the same issue—person and item validity and reliability using the Rasch model with holistic rubric scoring—with a larger number

of samples. Furthermore, it is suggested to involve multiple-rater to rate students' essays with the Multi-Facet Rasch model method to acquire more accurate person and item reliability and validity with a holistic rubric scoring scale.

REFERENCES

[1] C. Bond, T., & Fox, *Applying the Rasch Model*. Routledge, 2015.
[2] W. Sumintono, B. & Widhiarso, *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata Publishing House, 2013.
[3] & I.-C. A. Paul C. Price, Rajiv Jhangiani, *Research Methods of Psychology (4th ed)*. Victoria, BC: BCCAMPUS, 2020.
[4] S. A. Osman, S. I. Naam, O. Jaafar, W. H. W. Badaruzzaman, and R. A. A. O. K. Rahmat, "Application of Rasch Model in Measuring Students' Performance in Civil Engineering Design II Course," *Procedia - Soc. Behav. Sci.*, vol. 56, pp. 59–66, Oct. 2012, doi: 10.1016/J.SBSPRO.2012.09.632.
[5] G. Engelhard, "The measurement of writing ability with a many-faceted Rasch model.," *Appl. Meas. Educ.*, vol. 5, no. 3, pp. 171–191, 1992.
[6] M. S. Boone, W. J., Staver, J. R., Yale, M. S., Boone, W. J., Staver, J. R., & Yale, "Item Measures. Rasch Analysis in the Human Sciences," pp. 93–110, 2014, doi: https://doi.org/10.1007/978-94-007-6857-4_5.
[7] S. Tan, "Validation of an Analytic Rating Scale for Writing: A Rasch Modeling Approach.," *Tabaran Inst. High. Educ. Iran. J. Lang. Test.*, vol. 3, no. 1, 2013.
[8] E. I. D. & A. D. B.A, "Analyzing rater severity in a freshman composition course using many facet Rasch measurement.," *Lang. Test. Asia. Springer Open*, vol. 10, no. 1, 2020, doi: https://doi.org/10.1186/s40468-020-0098-3.
[9] K. Ashraf, Z.A., & Jaseem, "classical and modern methods in item analysis of test tools," *Int. J. Res. Rev.*, vol. 7, no. 5, pp. 397–403, 2020.
[10] S. Meyer, J.P., & Zhu, "air and equitable measurement of student learning in MOOCs: an introduction to item response theory, scale linking, and score equating," *J. Res. Pract. Assess.*, vol. 8, no. 1, pp. 26–39, 2013.
[11] H. B. Yilmaz, "A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test," *nternational Electron. J. Elem. Educ.*, vol. 11, no. 5, pp. 539–545, 2019, doi: https://dx.doi.org/10.26822/iejee.2019553350.
[12] X. Fan, "em response theory and classical test theory:An empirical comparison of their item/person statistics," *Educ. Psychol. Meas.*, vol. 58, pp. 357–381, 998.
[13] C. Magno, "Demonstrating the difference between classical test theory and item response theory using derived test data," *he Int. J. Educ. Psychol. Assess.*, vol. 1, pp. 1–11, 2009.
[14] A. A. Bichi, "Classical test theory: An introduction to linear modelling approach to test and item analysis," *Int. J. Soc. Stud.*, vol. 2, pp. 27–33, 2016.
[15] R. W. Hambleton, R.K., & Jones, "omparison of classical test theory and item response theory and their applications to test development," *ducational Meas. Issues Pract.*, vol. 12, pp. 38–47, 1993, doi: https://doi.org/10.1111/j.1745-3992.1993.tb00543.
[16] & R. H. J. Hambleton, R.K., Swaminathan, H., *Fundamental of item response theory*. London: Sage Publishing, 1991.
[17] N. Rezaee, R., Shafiayan, M., Jafari, P., & Zarifsanaiey, "Invariance of item difficulty parameter estimates based on classical test theory and item response theory," *J. Adv. Pharm. Educ. Res.*, vol. 8, pp. 156–161, 2018.
[18] S. Anastasi, A. & Urbina, *Psychological testing*. New York: Prentice Hall, 2002.

[19] K. S. Maier, "A Rasch hierarchical measurement model," *J. Educ. Behav. Stat.*, vol. 26, pp. 307–331, 2001, doi: https://doi.org/10.3102%2F10769986026003307.
[20] S. . Reise, "A comparison of item- and person-fit methods of assessing model-data fit in IRT," *Appl. Psychol. Meas.*, vol. 14, no. 2, pp. 127–137, 1990, doi: https://doi.org/10.1177%2F014662169001400202.
[21] A. Boone, W.J., & Noltemeyer, "Rasch analysis: A primer for school psychology researchers and practitioners," *Cogent Educ.*, vol. 4, no. 1, pp. 1–13, 2017, doi: https://doi.org/10.1080/2331186X.2017.1416898.
[22] M. Wright, B., & Stone, *Measurement essentials (2nd ed.)*. Wilmington: Wide Range, Inc., 1999.
[23] C. M. Bond, T.G., & Fox, *pplying the Rasch model: Fundamental measurement in the human sciences (2nd ed.).* Mahwah, NJ: awrence Erlbaum Associate, 2007.
[24] N. L. A. Zubairi, A.M., & Kassim, "Classical and rasch analyses of dichotomously scored reading comprehension test items," *Malaysian J. ELT Res.*, vol. 2, no. 1, pp. 1–20, 2006.
[25] L. M. Razak, N. bin Abd, Khairani, A.Z. bin, & Thien, "Examining quality of mathemtics test items using rasch model: Preminarily analysis," *Procedia - Soc. Behav. Sci.*, vol. 69, pp. 2205–2214, 2012, doi: https://doi.org/10.1016/j.sbspro.2012.12.187.
[26] M. Rost, J., & von Davier, "A conditional item-fit index for rasch model," *Appl. Psychol. Meas.*, vol. 18, no. 2, pp. 171–182, 1994, doi: https://doi.org/10.1177/014662169401800206.
[27] G. Karabatsos, "Comparing the abberant response detection performance of thirty-six person-fit statistics," *Appl. Meas. Educ.*, vol. 16, no. 4, pp. 277–298, 2003.
[28] R. R. Meijer, "Person-fit research: an introduction," *Appl. Meas. Educ.*, vol. 9, no. 1, pp. 3–8, 1996.
[29] B. Misbach, I. H., & Sumintono, "Pengembangan dan Validasi Instrumen 'Persepsi Siswa Tehadap Karakter Moral Guru' di Indonesia dengan Model Rasch," . *PROCEEDING Semin. Nas. Psikometri*, pp. 148–162, 2014.
[30] H. Jacobs., Holly. L., Stephen, A., Zingkgraf., Deanne. R., Wormuth, V., Faye, H., Jane, B., *Testing ESL Composition: A Practical Approach*. Rowley: Newbury House Publishers, Inc, 1981.
[31] W. Fisher, "Rating scale instrument quality criteria," *Rasch Meas. Trans.*, vol. 1, 2007.
[32] J. M. Linacre, "KR-20/Cronbach alpha or Rasch person reliability: Which tells us the truth?," *Rasch Meas. Trans.*, vol. 11, pp. 580–581, 2002.
[33] J. M. (Linacre, "What do infit and outfit mean-square and standardized mean?," *Rasch Meas. Trans.*, vol. 16, p. 878, 2002.
[34] E. V. J. Smith, "Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective," *J. Appl. Meas.*, vol. 2, no. 3, pp. 281–311, 2001.