Holistic Rubric Validity and Reliability in Essay Assessment Using Rasch Model in Blended Learning Program

Yenni Arif Rahman(✉), Nurhayati S., Fiza Asri Fauziah Habibah, and Fadilah Fadilah

Universitas Bina Sarana Informatika, Jakarta Pusat, Jakarta, Indonesia
yeni.yar@bsi.ac.id

Abstract. The covid-19 pandemic has forced teachers to adapt with learning method suitable with distance learning, and blended learning program is one of the best options around. Then the necessary assessment of output of this method should be available, valid, and reliable. This study aims to determine the validity and reliability of a holistic scoring rubric developed by Jacob et al. to assess students' writing in the Blended Learning program by using Rasch model during Covid-19 pandemic. The participants were 55 EFL learners who had taken 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 Minister software outputs used for data analysis: the "statistical summary output" to obtain figures and data in general, *output item one-dimensionality* to obtain reliability information, and *Fit-Order Items* to obtain *item validity*. The results of the reliability were reflected in *Cronbach's alpha* value (*a*) 0.91, *item reliability* 0.81, and *person reliability* 0.89 which show "excellent" reliability performance. The item fit order measured by *MNSQ*, *ZFTD*, and *PT. Measure correlation* with four assessment items: content, structure, diction, and mechanic fulfilled almost all ranges except for writing mechanics showing negative results on the *MNSQ*. It could be concluded that holistic rubric is the "Valid" measuring tool for assessing students' abilities in various test settings that require a rubric for students' essay assessment in blended learning program.

Keywords: Essay Assessment · Holistic Rubric · Rasch Model · Validity and Reliability

1 Introduction

When teachers plan to assess students' writing, they should consider the appropriate approach. Rubrics are considered the best way to evaluate and grade students' writing assignments. This efficient and increasingly popular approach to writing assessment aim to assess the overall level of proficiency reflected in a given sample of students' writing. Regarding that, this study aims to determine the validity and reliability of a holistic rubric to assess students' writing skills using the Rasch model during distance learning

M. Kristiawan et al. (Eds.): OCERI 2023, ASSEHR 775, pp. 685–698, 2024.
https://doi.org/10.2991/978-2-38476-108-1_68

through the Blended Learning method. This research result is expected to contribute how to implement validity and reliability test to any writing assessment by using Rasch model. In addition, the research result could also be a basis for teachers in assessing students' writing and could confirm the validity and reliability of writing assessment rubric developed by Jacob et al. [1] so that the teachers could utilize this rubric for their assessment in any setting.

Regarding writing assessment methods, there are three types conventionally agreed and valid, namely: holistic, analytic, and multi-trait assessment, and they have been widely used in writing assessment in different context [2]. However, holistic rubrics are accepted very quickly than the others in recent days [3]. This is due to several reasons. The most evident one is holistic assessment requires a thorough assessment of the quality of a student's writing, without analyzing its specific features. So it makes the teachers to assess essays faster and easier without making much work to do [4]. Surely this is not the most accurate assessment around. However, in particular condition which time is the most important, this less-time consuming assessment offers efficiency the others cannot do. This is supported by Lai, Wolfe, & Vickers [5] in Dunya & Erguvan [6] which state that the assessment method must be chosen carefully before assessing student writing directly. So once 'time' is considered the most crucial factor in assessment then holistic rubric is the best option should be delivered.

Related to holistic scoring, the word Rubric always comes hand-in-hand with the term. The word Rubric itself implies to an assessment tool that describes a level of performance on a particular task and is used to assess outcomes in various performance-based contexts [7]. In other word, education rubric is an assessment tool for qualitative assessment of student performance. It incorporates criteria for assessing important dimensions of performance, as well as standards for achieving those criteria [8]. Rubrics tell teachers and students what basic skills teachers look for when they assess students' performance [9]. Rubrics are also powerful tools for measuring test taker performance. They provide opportunities for reliable judgments, rather than subjective judgments based solely on the rater's personal idiosyncrasies [10]. Among several benefits of using rubrics, providing consistency of grading between students, assignments, as well as between different raters are the main ones. Rubrics offer a way to provide validity in assessing complex talent, without ignoring the need for reliability [11]. Rubrics also promote learning by making criteria and standards clear to students and providing them with quality feedback [9].

Discussing validity, the term refers to a measurement that ensures a variable measures what should be measured [12]. A test will have high validity if it is able to run a function as a measuring tool. This means the instrument is able to provide precise measurement results and accurate in accordance with the purpose for which the instrument was developed [13]. If a test instrument, for example, has low validity, then the data generated through the test will be accepted as irrelevant or not accurate. In addition to referring to the accuracy in making measurements, validity instrument can also refer to the accuracy of the instrument. Valid instrument will have a high degree of accuracy in measurement. Accuracy in this sense is the instrument's ability to detect differences in the measured attribute even though the difference is small.

Regarding the method to measure validity and reliability, one of the popular and accurate approaches to measure them is provided by Rasch model. In recent years, the Rasch model, also called the *Theory of Item-Response* (IRT) or *Properties Latent Model*, has provided an alternative framework for understanding measurement and alternative strategies for assessing the quality of an instrument or questionnaire [14]. The application of the Rasch model can produce more reliable and valid instrument. The Rasch model can also prove that an instrument has a high level of validity and reliability [15]. This is because Rasch model provides useful statistics and offers a tremendous opportunity to investigate the validity of the instrument. The development of Rasch's model in educational science is a response to various weaknesses of the CTT paradigm [16]. The advantage of Rasch model over *Classical Test Theory* (CTT) is the ability to predict missing data, which is based on a systematic response pattern. This clearly makes the results of statistical analysis more accurate in the research conducted. In this way, the existing data can be processed as complete data so as to be able to produce standard error measurement values for instruments which can increase the accuracy of calculations [16].

Further, fundamental differences of the Rasch model when compared to CTT lies in how to treat raw scores in the process analysis. In CTT, the raw score in the form of a rating scale is directly analyzed and treated as data as if it had integer characters. While in the Rasch model, raw data cannot be directly analyzed, but must first be converted into 'odds ratio' then processed with algorithmic transformation into logit units as a manifestation the respondents probability responding to an item [17]. Referring to this procedure, then Rasch model can be used as a method to return data according to its natural condition [16]. This natural condition refers to the basic characteristics of quantitative data, which is a continuum. Through the Rasch model, an ordinal response can be transformed into the form of a ratio that has a higher level of accuracy by referring to the principle of probability. Another difference that distinguishes the Rasch model from the CTT is that in data analysis using the Rasch Model, the data fit the model. Whereas in the CTT, the model is selected based on the data. Based on this procedure, the use of the Rasch model in the validation of instrument will result in more holistic information and better meet the need of measurement.

2 Methods

55 EFL students who have taken the TOEFL iBT essay writing course were the participants in this study. Participants were high school and university students, with an assumed English proficiency level of Intermediate or higher. The TOEFL score above 500 proves this. With the assumption that an effective essay structure has been developed and consists of an introduction, content, and conclusion, essays typically have a length of five paragraphs. This generic structure is needed so that the results of the assessment by the raters is not bias related to the number of paragraphs that can be a plus and minus tendency of the assessment. This structure is also a reference for whether or not a student's writing can be used as a essay sample because uniform minimum requirements are needed so that research results can be appropriately validated. For data collection,

the Academic Writing instructor, who generally and explicitly provided directions for writing essays, was involved in collecting the essay sample.

Research Stages

In doing this research, the researchers went through 4 stages: first, identification of measurement objectives (determination of theoretical constructs). The constructs revealed are the validity and reliability of the holistic measurement level in determining the participants' writing ability. The level of holistic measurement has been determined by several writing ability items categorized into four: 1. Content and coherence, 2. Grammar, syntax, sentence structure, and cohesion, 3. Diction, and 4. Writing mechanics. The adaptation of those four items are presented in Table 1 by adapting the holistic Rubric developed by Jacobs et al. [1].

The second stage is scaling. The scaling method utilized is the *Likert* scale with five response/assessment options from Table 1, which are: proficient, fluent, expanding, developing, beginning and emerging (merged into one group). The interpretation of the rubric rating scale in Table 2 is as follows:

The third stage is assessment and scoring. To maintain the objectivity of each writing sample, the assessment and scoring process involves a rater with experienced scoring essay scores. The rater used the holistic rubric in Table 1 as a reference to rate the students' essays. The result then converted into *Likert* score in Table 2 to undergo analytical process of Rasch model in Ministep software.

The last stage is analysis. This study was analyzed using the Rasch model because it could see the interaction between participants and items at once. In the Rasch model, a value is not seen based on the raw score but a logit value that reflects the probability of selecting an item in a group of participants. This is used to anticipate the raw score of the *Likert* rating in the form of an ordinal that does not have the same interval between scores. The use of the Rasch model for polytomy data was developed by Andrich [18] still based on two basic theorems, namely the level of individual ability/agreement and the level of difficulty of the item to be approved [19]. The Ministep output used for data analysis is the output summary statistics (Fig. 1), to obtain reliable information is the output *item one-dimensionality* (Fig. 2), and for validity is the *Item Fit Order* (Fig. 3).

Measuring Validity with Item Fit Order

Items fit means that the items behave consistently with what the model expects [12]. Some of the fit provided in the Rasch analysis are *Person Infit ZSTD, Person Outfit ZSTD, Person Infit MNSQ, Person Outfit MNSQ, Item Infit ZSTD, Item Outfit ZSTD, Item Infit MNSQ*, and *Item Outfit MNSQ* (Boone, Staver & Yale, 2014). The *MNSQ* value is used to monitor the suitability of the data with the model. The *MNSQ* value is always positive and moves from zero (0) to infinity (∞). The expected mean square value is 1 (one). A mean-square value for infit or outfit greater than one indicate that the observed data has 30% more variation than predicted by Rasch model. An infit or outfit value of less than 1, say 0.78 (1–0.22 = 0.78), indicates that the observed data has 22% less variation than predicted by the Rasch model [21].

Meanwhile, the expected z value is close to 0 (zero). When the observed data fit the model, the z value has a mean close to 0, and the standard deviation is 1. *ZSTD* values that are too large ($z > +2$) or too low ($z < -2$) indicate that the item is not

Table 1. Holistic Score: Rating and Criteria

Rating	Criteria
Proficient	1. Writes single or multiple paragraphs with clear introduction, fully develop idea, and clear introduction 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 meaning
Fluent	1. Writes single or multiple paragraphs with main idea and supporting detail, present idea logically, though some parts may not fully develop 2. Uses appropriate verb tense and a variety of grammatical and syntactical structures; errors in 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 meaning
Expanding	1. Organizes ideas in logical or sequential order with some supporting detail; begins to write a paragraph 2. Experiment with a variety of verb tenses, but does not use them consistently; subject/verb agreement errors; uses some compound and complex sentences; limited use of transitions 3. Vocabulary is appropriate to purpose but sometimes awkward 4. Use punctuation, capitalization, and mostly conventional spelling; errors sometimes interfere with meaning
Developing	1. Writes sentences around an idea; some sequencing present, but may lack of 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
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 word, pictures, and patterned phases 3. Copies from model 4. Little awareness of spelling, capitalization, or punctuation

compatible with the expected model. The standardized z-value (ZSTD) on infit and outfit can be either positive or negative. A negative *ZSTD* value indicates less variation compared to the model. The answer response is close to the Guttman-style response string model; all subjects with high abilities can answer correctly, and all subjects with low abilities answer incorrectly on the item. Meanwhile, a positive value indicates more

Table 2. Rubric Rating Scale

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

answer variations than the model. Response responses are irregular and unpredictable [21]. According to Boone et al. [20], the criteria used to check the appropriate items are:

1. Outfit Mean Square (MNSQ) value received: $0.5 < MNSQ < 1.5$
2. Accepted Z-standard (ZSTD) outfit value: $-2.0 < ZSTD < +2.0$

If the items in the two criteria are not met, the items are not good and need to be revised or replaced. In contrast to the level of difficulty of the item, which is consistent, the level of suitability of this item is strongly influenced by the size of the sample. Answer key errors, the number of individuals who carelessly work on the questions, and questions with low discriminating power can reduce the value of the item's suitability.

3 Result and Discussion

The results of data processing in the form of output Figs. 1, 2, and 3 in this finding section were obtained from the Ministep Rasch 4.8.2.0 software data processing to see the reliability and validity of the holistic scoring rubric of EFL learners' essays. There are several data outputs which determine the reliability and validity of the rubric. They are *person reliability*, *item reliability*, *infit* and *outfit* both *MNSQ* and *ZSTD*, *raw variance*, and *PT. Measure.* To ease the tracking and spotting, the determining data outputs have been marked with red square in Figs. 1, 2, and 3.

1. Instrument Reliability

The instrument reliability is tested to determine whether the holistic essay scoring rubric instruments are reliable, and whether it can be used as a measuring tool to measure learners' essay whenever and wherever it is used. The Summary statistic (Fig. 1) is used in the Ministep software to determine the reliability of the instrument to be tested. The summary statistics in Fig. 1 display both the person reliability (participant's side) and item reliability (item reviews), as well as the interaction between participants and items.

The outputs of the Summary Statistics in Fig. 1 provide information on the quality of the instrument (*item*) and sample (*person*) in essays. It also indirectly explains the interaction between *person* and *item.* So, what needs to be discussed from Fig. 1 is *person measure* to show the average score of participants in the instrument viewed from the *mean measure* score which is 0.73. *Cronbach's alpha* obtained from the "*Test" reliability* score, which score 0.91 is used to measure *reliability* and the interaction between person and item. The score of *person reliability* which scores 0.89 is used to

see the consistency of the answers from participants, and *item reliability* (0.81) which is the reliability score of the questions is to determine the quality of the items in the instrument. In addition, Fig. 1 displays *INFIT* and *OUTFIT MNSQ* which score 0.95 and 0.74 for *person reliability*. While *INFIT* and *OUTFIT ZSTD* of *item reliability* score 0.95 and 0.74. Figure 1 data outputs also present ZSTD scores –0,25 and –0.57 for item reliability, and separation is 2.09.

2. Instrument Validity

The validity of the instrument is used to test whether the holistic scoring rubric instrument is accurate to measure essays from blended learning program. This section is a step to explain the interaction between the subject (*person*) and the item (*test items*). The data outputs used in the Ministep software are *Item one-dimensionality* in Fig. 2 and *Item Fit Order* in Fig. 3. The *item one-dimensionality* means all items just measure one factor which is the validity. *Item fit order* explains whether the item (in Table 1) has functioned normally to measure what should be measured. Analysis of the validity of this holistic essay score rubric instrument in the Ministep software is called *fit* and *misfit* test (valid and invalid items). The criteria used to check the item including fit or misfit can be done by analyzing the output of this item fit order. The following is the figures of the two table outputs which show information on the validity criteria of the instrument:

TABLE 3.1 DATA MINISTEP.xlsx ZOU594WS.TXT Aug 11 2022 6:2(
INPUT: 55 PERSON 4 ITEM REPORTED: 55 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.(

SUMMARY OF 55 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	14.4	4.0	.73	1.67	.67	-.29	.74	-.30
SEM	.4	.0	.77	.08	.11	.15	.16	.14
P.SD	2.7	.0	5.63	.59	.83	1.07	1.19	1.01
S.SD	2.8	.0	5.69	.60	.83	1.08	1.20	1.02
MAX.	19.0	4.0	9.53	2.54	4.03	3.48	6.46	2.84
MIN.	10.0	4.0	-8.01	1.09	.03	-1.02	.02	-1.06

REAL RMSE 1.85 TRUE SD 5.32 SEPARATION 2.88 PERSON RELIABILITY .89
MODEL RMSE 1.77 TRUE SD 5.35 SEPARATION 3.02 PERSON RELIABILITY .90
S.E. OF PERSON MEAN = .77

PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .91 SEM = .82
STANDARDIZED (50 ITEM) RELIABILITY = .99

SUMMARY OF 4 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	198.0	55.0	.00	.38	.95	-.25	.74	-.57
SEM	3.8	.0	.54	.00	.13	.64	.18	.47
P.SD	6.6	.0	.93	.00	.23	1.11	.31	.82
S.SD	7.6	.0	1.07	.01	.27	1.28	.36	.94
MAX.	203.0	55.0	1.55	.39	1.24	1.06	1.24	.71
MIN.	187.0	55.0	-.71	.38	.66	-1.67	.44	-1.47

REAL RMSE .40 TRUE SD .84 SEPARATION 2.09 ITEM RELIABILITY .81
MODEL RMSE .38 TRUE SD .84 SEPARATION 2.20 ITEM RELIABILITY .83
S.E. OF ITEM MEAN = .54

Fig. 1. Output Summary Statistics in Ministep

```
    Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                        Eigenvalue   Observed   Expected
Total raw variance in observations   =      20.7945 100.0%        100.0%
  Raw variance explained by measures =      16.7945  80.8%         79.8%
    Raw variance explained by persons =      14.3319  68.9%         68.1%
    Raw Variance explained by items  =       2.4626  11.8%         11.7%
  Raw unexplained variance (total)   =       4.0000  19.2% 100.0%  20.2%
    Unexplned variance in 1st contrast =      1.8916   9.1%  47.3%
    Unexplned variance in 2nd contrast =      1.3439   6.5%  33.6%
    Unexplned variance in 3rd contrast =       .7111   3.4%  17.8%
    Unexplned variance in 4th contrast =       .0557    .3%   1.4%
    Unexplned variance in 5th contrast =       .0013    .0%    .0%
```

Fig. 2. Item Unidimensionality Output in Ministep

Figure 3 shows the standardized residual variance in eigenvalue unit. This means eigenvalue reveals the item information unit. What important and determines the result is the *raw variance explained by measure* column where the score is 80.8% and the *unexpended variance in the 1st contrast* column where the *eigen value* is 1.8916 and the *observed* score is 9.1%. The expected percentage of raw variance explained by measures shows 79.8% while *unexpended variance in 1st contrast* display 47.3%. Both of the percentage is important to determine the reliability of the item.

Figure 4 shows *Expected score ICC* (Item Characteristic Curve) which can be used to see the item misfit. The ideal curve model (*ogive curve*) is represented by red curve. The response of the participants to those four items in graphic 1. Content, 2. Structure, 3. Diction, 4. Mechanic are represented by black dots connected by blue line (characteristic curve). These black dots also represent observed scores in each of intervals of the trait. The two black curves which lie on top and bottom the *ogive curve* of those 4 items represent the interval of confidence curve. If those black dots are outside the *confidence interval* curve, it can be interpreted that the misfit response has occurred. However, from all those four items, the *characteristic curve* is inside the *confidence interval* which means misfit response didn't occur to those four items.

There are three instruments will be discussed in this section. The result will determine the reliability and validity of items of Holistic rubric. Instrument reliability result is

```
TABLE 14.1 DATA MINISTEP 2.xlsx                ZOU738WS.TXT  Aug 11 2022  7: 9
INPUT: 55 PERSON  4 ITEM  REPORTED: 55 PERSON  4 ITEM  4 CATS   MINISTEP 4.8.2.0
-------------------------------------------------------------------------------
PERSON: REAL SEP.: 2.88  REL.: .89 ... ITEM: REAL SEP.: 2.09  REL.: .81

         ITEM STATISTICS:  ENTRY ORDER

-------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL          MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|            |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM       |
|------------------------------------+----------+----------+-----------+-----------+------------|
|     1    203     55    -.71     .39|1.24  1.06| .74  -.49|  .90   .88| 76.4  83.5| Content    |
|     2    187     55    1.55     .38|1.11   .58|1.24   .71|  .88   .88| 83.6  81.6| Structure  |
|     3    199     55    -.13     .38| .80  -.97| .53 -1.02|  .89   .88| 83.6  81.5| Diction    |
|     4    203     55    -.71     .39| .66 -1.67| .44 -1.47|  .89   .88| 90.9  83.5| Mechanic   |
|------------------------------------+----------+----------+-----------+-----------+------------|
| MEAN   198.0   55.0     .00     .38| .95   -.2| .74   -.6|           | 83.6  82.5|            |
| P.SD     6.6     .0     .93     .00| .23   1.1| .31    .8|           |  5.1   1.0|            |
-------------------------------------------------------------------------------------------------
```

Fig. 3. Standardized Residual Variance in *Eigenvalue* Unit in Ministep

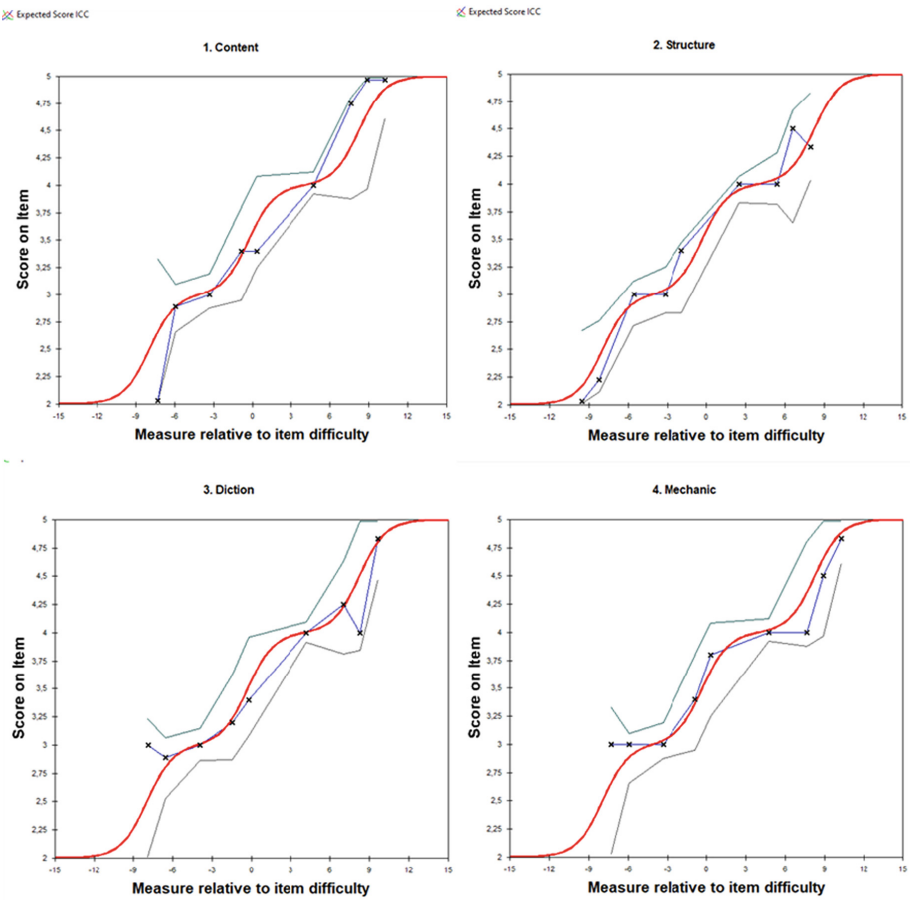

Fig. 4. Expected Score ICC Curves

represented by Fig. 1 which shows test reliability and item reliability. The following points are observed and inferred from them:

1. *Person measure* in Fig. 1 shows 0.73. The score is greater than logit 0.0. Which indicates the tendency of students' abilities to exceed standard criteria for writing skills within the items (Fig. 1). From the Table 2 the exact reliability is interpreted as "good" which implies that the students have understood and are able to apply content, structure, diction, and mechanic in their essay and the lesson has successfully delivered the material to the students.
2. *Cronbach's alpha* value is used to measure the interaction between students and items which is in the output of Fig. 1. The finding in Fig. 1 that *Cronbach's alpha* value is 0.91, which is higher than 0.8 in Table 3 the interpretation of the score is "Excellent". This means that the items are reliable to be used as the instrument to measure the students' essays.

Table 3. Reliability Test Index Based on *Cronbach's 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

(Sumintono & Widhiarso, 2015)

3. The score of *Person Reliability* and *Item Reliability* in the output of Fig. 1 both show 0.89 and 0.81. The *Person Reliability* shows that the consistency of the participants is "excellent". The interpretation is inferred from *Cronbach's alpha* index in Table 2 as reference. The *Item Reliability* score shows an "excellent" level so that the quality of the items used in the instrument is reliable as well.
4. The score of *INFIT MNSQ* and *OUTFIT MNSQ*, as well as *INFIT ZSTD* and *OUTFIT ZSTD*, both can be seen in Fig. 1 shows the following results: *INFIT MNSQ* has a person score of 0.67 and an item score of 0.95, *OUTFIT MNSQ* has a person score of 0.74 and an item score of 0.74. It can be seen that the scores shown in the table person and items from *INFIT MNSQ* and *OUTFIT MNSQ* are all closer to the value of 1.00, because the closer the data is to the value of 1.00, the better the quality is. Then for *INFIT ZSTD* the person score is –0.29, the item score is –0.25, *INFIT ZSTD* the person score is –0.30, and the item score is –0.57. Through these scores, the ideal score is 0.0, which means that the closer to the ideal value, the better the quality is. For items, all scores tend to be closer to 0.0, but for person it is still far, meaning that there are indeed participants who are not good at completing the given task.
5. The grouping of persons and items can be seen from the separation value. The greater the value of separation, the better the quality of the instrument in terms of overall participants and items because it can identify groups of participants and groups of items. For the separation person, the score in Fig. 1 is 2.88 and the score for the separation item is 2.09.

Based on the previous review from 1–5, the holistic rubric instrument has an "excellent reliability" with the following details:

Based on Table 3 the interpretation of *Cronbach's alpha* (0.91) is 'excellent'. This shows that there is a match between the item and the person (participant). Then the consistency of the answers from the participants (person reliability) is 'excellent' with

Table 4. The Result of Reliability Test

Cronbach's alpha	**Interpretation**	**Item Reliability**	**Interpretation**	**Person Reliability**	**Interpretation**	**Summary**
0,91	Excellent	0,81	Excellent	0,89	Excellent	Reliable

the quality of the instrument items (item reliability) is 'excellent' as well. So it was concluded that the holistic score instrument studied is "reliable".

Validity concerns with how far the test items are able to measure what should be measured according to a specific concept or conceptual definition has been set [22]. In Rasch analysis, the validity test is identified with the *item unidimensionality* [16]. The items is considered unidimensional if there is only one variance or latent variable responsible for systematic variations found within the item variance. And this idea is used to test the unidimensionality of items if there are no correlated residuals between the items [23]. Regarding that, Rasch analysis uses component analysis principle of the standardized residual variance (in Eigenvalue units) [16]. The validity test based on *item unidimensionality* can be seen in the raw variance value explained by measures. Interpretation of *items unidimensionality* based on *raw variance values explained by measures* was indicated by scores $> 20\%$ is considered acceptable, $> 40\%$ good and $> 60\%$ excellent. In addition, the *Eigenvalues* and *observed* in *unexplained variance 1st contrast* are used to find out whether there are problematic and misfit items. The *eigenvalues* itself must be less than 3 to indicate there are no problematic items and the *observed values* must be less than 15% to indicate the item is fit (item fit). The results of validity processing of Rasch model with Ministep 4.8.2.0 are presented in Table 4.

Based on Table 4 the results of the *raw variance values explained by measures* (80.8%) indicate that the overall resilience test item is in the *excellent* category. Furthermore, based on the values *observed* in the *unexplained variance 1st contrast*, there is no tendency for discrepancies in the items so that they can be used, and the *eigenvalues (1.8916)* less than 3 indicate that there are no problematic items so that further analysis doesn't need to be carried out and the items themselves don't contain any problems.

Item fit can explain whether the items function normally to measure or not. The *outfit means-square, outfit z-standard,* and *point measure correlation* are the criteria used to see the level of *item fit* (Boone et al., 2014). The criteria used to check the suitability of the items are presented in Table 5.

Based on Table 5, if all three criteria are fit on the item, it can be interpreted that the item is *acceptable* and it is ascertained that the quality of the item is *good* and can be used. Whereas if there are only two criteria or one criterion that fit, the item can still be used, maintained, and does not need to be changed so that it can be categorized as *acceptable* and can be used. However, if the three criteria don't fit then it can be interpreted that the items are not valid and it can be ascertained that the items are *unacceptable* so that they need to be corrected or replaced. The results of *item fit order* are presented in Table 6.

Table 6 shows the item validation mechanism in the instrument. The data in the table shows that item number 1, content, is *valid*. The conclusion is obtained from the *MNSQ*

Table 5. The Result of Instrument Validity

Raw variance explained by measure	**Interpretation**	***unexplained variance 1st contrast***		**Interpretation**
		Eigenvalue	***observed***	
80.8%	Excellent	1.8916	9.1%	No-Misfit Item

Table 6. 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
Point Measure Correlation	0,4 < *PT Measure Corr* < 0,85

(Sumintono & Widhiarso, 2013).

Table 7. The Result of *Item Fit Order*

Number	Item	Outfit		*PT Measure Corr.*	Item Fit Criteria			Interpretation
		MNSQ	*ZFTD*		*MNSQ*	*ZFTD*	*PT Measure Corr.*	
1	Content	0,74	–0,49	0,90	Fit	Fit	Misfit	Valid
2	Structure	1,24	0,71	0,88	Fit	Fit	Misfit	Valid
3	Diction	0,53	–1,02	0,89	Fit	Fit	Misfit	Valid
4	Mechanic	0,44	–1,47	0,89	Misfit	Fit	Misfit	Valid with note

and *ZFTD* values that meet the threshold of Table 5. While the *PT Measure Corr value.* Passed the upper threshold of 5 digits. Even so, the two criteria that meet the criteria can be interpreted that the *content* has met the validity of the item value so that it can be maintained. The same thing happened to items no. 2 and 3, *structure* and *diction*, showing the same pattern of results and conclusions as item no. 1. While item no. 4, *mechanics*, showed a valid value conclusion with a note that there was a possibility that the items in the rubric of holistic mechanic scores must be corrected. The other possibility is the students don't pay attention to the aspects of a good writing convention or the items have low discriminating power so that they can reduce the item's validity value which is signified by discrepancy between the *MNSQ* and *ZFTD* scores.

4 Conclusion

Tables 3 and 7 inform that the use of the Rasch model in instrument reliability and validity produce more holistic information about the instrument under study and better meet the definition of measurement. The Rasch model result analysis in Table 3 also shows that the holistic rubric for scoring students' essay was *reliable* with the 'excellent' criteria. Table 6 also tells that four items, content, structure, diction, and mechanic, were *valid* so that the rubric of holistic score is valid and can be utilized to assess the EFL students' essay in various test settings that require rubrics. However, even though the overall result of the four items shows 'valid', but the actual result of 'mechanic' in Table 6 tells it doesn't satisfy all the Rasch model standard. This factor then needs to be

recognized and need further research to investigate whether the discrepancy of the score occurs in person/participant or the measurement item level.

Acknowledgments. The researchers indebted to two institutions which have facilitated the research from the start until its completion. Without their invaluable contribution, this research would never be accomplished. We would like to acknowledge:

1. Universitas Bina Sarana Informatika
2. Elokuensi, International Language Center

References

1. 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.
2. Y. Y. and L. W. P. Richard P. Bagozzi, "Assessing Construct Validity in Organizational Research," Sage Publ. Inc., vol. 36, no. 3, pp. 421–458, 1991, doi: https://doi.org/10.2307/2393203.
3. B. Huot, "Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know," Coll. Compos. Commun., vol. 41, no. 2, pp. 201–213, 1990, doi: https://doi.org/https://doi.org/10.2307/358160.
4. S. Dikli, "Assessment at a distance: Traditional vs. Alternative Assessments," Turkish Online J. Educ. Technol., vol. 2, no. 3, pp. 13–20, 2003.
5. D. V. E. Lai, E. Wolfe, "Halo Effects and Analytic Scoring : A Summary of Two Empirical Studies Research," Lang. Test. Asia. Springer Open, vol. 10, no. 1, 2012, doi: https://doi.org/10.1186/s40468-020-0098-3.
6. 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.
7. P. Hafner, J., & Hafner, "Quantitative analysis of the rubric as an assessment tool: An empirical study of student peergroup rating," Int. J. Sci. Educ., vol. 25, no. 12, 2003, doi: 1509–1528. https://doi.org/10.1080/0950069022000038268.
8. G. Jonsson, Anders; Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," Educ. Res. Rev. Sci. Direct., vol. 2, no. 2, pp. 130–144, 2007, doi: https://doi.org/10.1016/j.edurev.2007.05.002.
9. J. Arter, J., & McTighe, Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. California: Corwin Press, 2001.
10. N. T. Carr, "A comparison of the effects of analytic and holistic rating scale types in the context of composition tests," Issues Appl. Linguist., vol. 11, no. 2, pp. 207–241, 2000.
11. S. M. Kemp, J. E., Morrison, G. R., & Ross, Designing effective instruction (2nd ed.). Upper Saddle River: Prentice Hall, 1998.
12. P. S. (2006) Cooper, D.R., & Schindler, Business Research Methods. USA: Mcgraw Hill, 2006.
13. H. Mohajan, "Two Criteria for Good Measurements in Research: Validity and Reliability," Ann. Spiru Haret Univ., vol. 17, no. 3, pp. 58–82, 2017.
14. T. H. Nguyen, H. R. Han, M. T. Kim, and K. S. Chan, "An Introduction to Item Response Theory for Patient-Reported Outcome Measurement," Patient, vol. 7, no. 1, p. 23, 2014, doi: https://doi.org/10.1007/S40271-013-0041-0.

15. D. Rachman, T. & Napitupulu, "Rasch Model for Validation a User Acceptance Instrument for Evaluating E-learning System," CommIT (Communication &Information Technol. J., vol. 11, no. 1, pp. 9–16, 2017.
16. W. Sumintono, B. & Widhiarso, Applications of the Rasch Model to Social Science Research. Trim Komunikata Publishing House, 2013.
17. S. Wibisono, "Rasch Model Application for Validation of Religious Fundamentalism Measurement Instruments for Muslim Respondents," J. Pengukuran Psikol. Dan Pendidik. Indones., vol. 5, no. 1, 2018, doi: https://doi.org/10.15408/jp3i.v5i1.9239.
18. D. Andrich, "A rating formulation for ordered response categories," Psychometrika, vol. 43, no. 1, pp. 561–573, 1978, doi: https://doi.org/https://doi.org/10.1007/BF02293814.
19. B. Misbach, I. H., & Sumintono, "Development and Validation of the Instrument 'Student Perceptions of Teacher Moral Character' in Indonesia with the Rasch Model," . PROCEEDING Semin. Nas. Psikometri, pp. 148–162, 2014.
20. 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.
21. C. Bond, T., & Fox, Applying the Rasch Model. Routledge, 2015.
22. P. Djaali, & Muljono, Measurement in the field of education. Jakarta: PT.Grasindo, 2008.
23. P. . Lazarfeld, "Latent Structure Analysis," S. Koch (Ed), Psychol. A Study a Sci., vol. 3, no. 1, pp. 476–543, 1959.