International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 11, Issue 6, 2020

Performance Assessment Analysis Using the Rasch Many-Facet Measurement Model for Junior High School Geometry Students

Rivo Panji Yudha[a*], Soeprijanto[b], Dinny Devi Triana[c], aFaculty of Science Education, Universitas 17 Agustus 1945 Cirebon[a], [b,c]West Java, Indonesia, Universitas Negeri Jakarta, Jakarta, Indonesia, Email: [a*]rivoyudha@yahoo.co.id,

The purpose of this study was to analyse a set of questions in geometry lessons using performance assessment through the Rasch Many-Facet Measurement Model approach. The data was collected from 250 junior high school students using 3 raters. Performance instruments in the form of rubrics are used to assess the process of students working on problems, and each question has a different rubric. The Rasch Many facet Measurement Model (MFRM) is used to analyse data to see three aspects namely facet person, rater agreement and domain difficulty. The results show that an appropriate index measures the high feasibility and low feasibility of the facet person in terms of number one through number five having a good enough item validity value and high reliability. The high feasibility and construct value following the MFRM shows that this measurement model can measure the accuracy of the scores for each facet.

Keywords: *Performance assessment, rater, many facet rasch measurement model (MFRM).*

Introduction

Assessment and Curriculum in Indonesia, namely 2013 Curriculum, requires teachers to be able to develop assessment instruments in accordance with learning objectives and aspects of integrated thematic learning. The instrument developed should be tested for quality so as to produce data that can be accounted for. The instrument used for data collection of student learning outcomes in integrated learning should meet five requirements, namely validity, reliability, objectivity, practical, and economical. The fact is that teachers in the field have not used quality assessment instruments. Akbar (2013: 4) states that currently, the tendency

of teachers in the field still uses tests as the main instrument in recording student learning outcomes. The instruments used have not been tested in terms of validity, reliability, and practicality.

Student performance assessment is one alternative assessment that is focused on two main activities, namely: Observation of the process during the performance of skills and evaluating the product or creativity. The skills displayed by students are the variables assessed (Yudha, Anggara, & Zulaeha, 2019). Assessment of student skills is based on a comparison between student performance and targets that have been set. Therefore, written and oral assessments alone cannot represent all of the desired assessments, especially with discussion material that requires students to be able to solve problems and determine attitudes, collaborate with group peers and others.

Performance assessments seek and gather information about students' abilities in understanding and applying knowledge and process skills in real situations (McTighe & Ferrata, 2010). Characteristics of performance assessment or authentic assessment, require meaningful assignments designed to represent performance, emphasise higher-level thinking and learning more complex, articulate criteria in advance so students know how to be evaluated, expect students to present their work in public whenever possible, and involves examining learning processes and products (Lund & Kirk, 2010).

The main problem in performance appraisal is also the problem of scoring analysis specifically due to many factors that influence the results of the skills assessment or performance assessment scoring analysis. The problem of scoring analysis on skills assessment or performance appraisal is more complicated than scoring analysis in the form of the problem description. In addition, the components that must be assessed are also difficult to score, generally because they are difficult to observe (unobservable). This will certainly result in invalid scoring results, and with the results therefore being inaccurate (unreliable).

The performance assessment presents a number of validity issues that are not easily dealt with by traditional approaches and criteria for validity research. These assessments usually give students substantial freedom in interpreting, responding to, and possibly designing assignments; they produce fewer independent responses, each of which is complex, reflecting the integration of various skills and knowledge; and they need an expert judgment for evaluation. Consequently, meeting criteria related to validity issues such as reliability, generalisation, and comparability of assessments, at least as defined and operationalised, becomes problematic. (Moss, 1992).

When evaluating performance evaluations that are mediated by assessors, it is important to recognise that the interpretation of each of these problems is directly related to the quality of

the ratings mediated through a group of assessors. Although some recent discussions about the validity of performance assessments mention the involvement of assessors in the assessment process, including appraisal training procedures, concerns with appraisal bias, and appraiser's consistency (Brennan, Robert L. Brennan, & Brennan, 2006; Slomp Prof., Corrigan, & Sugimoto, 2014), Specific validity concerns and issues related to the assessor have not yet been fully explored. On the contrary, the role of the assessor and the quality of the ranking in this assessment is generally displayed in discussions on issues of reliability and generalisation.

Some researchers have reported that teacher ratings are more reliable if rubrics are used (Jonsson & Svingby, 2007; Silvestri & Oescher, 2006). No studies have been found that show the negative effects of using rubrics (reducing reliability). As a result, many teachers have used rubrics with the assumption that they are increasing objectivity assessments, specifically regarding sending student writing. As a result, there is another assumption that grading without rubrics tends to be more subjective because it is based only on the subjective assessment of each student, and his overall impression on the author's style. With this in mind, teachers often decide that using rubrics is better than not using rubrics (Spandel, 2006). Difficulties in assessing performance assessment results often affect the ability to evaluate students accurately. The reliability and validity of the assessment tools need to be established. While cognitive learning outcomes are often assessed using multiple-choice exams or short answers, this assessment of actions does not often assess higher levels of cognition or other learning domains. Performance appraisals can also enable certain tasks to produce multiple scores in various content domains, which have practical and pedagogical appeal. Tasks designed to score in more than one content domain may not only reflect a more integrated approach to teaching but also motivate a more integrated learning approach.

Procedural issues, procedures used in skills assessment or performance appraisal are not effective so that it also affects the scoring results. The problem that usually occurs is the scorer (rater) has to suspend too many components. For scorers, the fewer components that must be assessed the better, but the scoring guideline must still make scoring guidelines that can represent all the important components that affect the quality of the final result. Another problem with this procedure is that generally the scorer (rater) is only one person, so it is challenging to be able to compare the results of the scoring considerations with others.

A common problem that has so far been a matter of reliability for this performance assessment relates to variations among assessors as sources of measurement errors. Emphasise the need to go beyond these indicators of consistency or agreement and evaluate the quality of rankings based on the requirements for invariant measurements. In contrast to analyses based on decomposing error variance into the overall source of measurement error, we highlight the importance of checking the accuracy of measurements related to individual

elements in aspects such as individual assessors, students, or rubric domains. In particular, the Rasch Many-Facet (MFR) model as a useful framework where it is possible to explore indicators of reliability and precision associated with various aspects of the assessment procedure while maintaining a focus on rater-invariant measurements.

The problem of biased scoring, scoring (rater) tends to be challenging to eliminate problems, personal bias. When scoring the test participant's work, there is a possibility that the scorer (rater) has a generosity error problem which means the scorer tends to give a high score, despite the fact that the test taker's work is not good. It is also possible that the scorer has a severity error problem, which means that the scorer tends to score low, despite the fact that the test taker's work is good. Another possibility is that the scorer also tends to be able to give a moderate score, even though in reality the results of the test taker's work are good and some are not good. Another problem is the possibility that the scorer is interested or sympathetic to the test taker so that it is difficult for him to give an objective score (hello effect) (Engelhard Jr., 2002; Knoch, 2009; Myford & Wolfe, 2003, 2004; E. Wolfe, 2004).

Performance assessments typically use built-in response items. Such a thing requires the examinee to form a response, rather than choosing the correct answer from the alternatives given. To achieve the score to capture the intended proficiency, the rater must pay attention, interpret, and evaluate the responses given by the examinees. Thus the performance assessment process can be described as a complex and indirect process. Examinees respond to test items or assignments designed to represent the underlying construct (for example, writing skills), and assessors assess the quality of responses built on their understanding of it. Building and utilising a more detailed assessment rubric (Bejar, Williamson, & Mislevy, 2006; Cumming, 2007; Freedman & Calfee, 1983; Roever & McNamara, 2006; E. W. Wolfe, 1997). This long and possibly fragile interpretation evaluation chain highlights the need to investigate the psychometric quality of assessments mediated by rater carefully. One of the main difficulties facing researchers, and also practitioners, is the occurrence of rater variability.

The term rater variability generally refers to the variability associated with the characteristics of the rater and not to the performance of the examinees. In other words, rater variability is a component of unwanted variability that contributes to constructing variance that is irrelevant in the test participant's score. This type of variability obscures the construct that is measured and, therefore, threatens the validity and reasonableness of performance assessment (Brennan et al., 2006; Messick, 1995; Roever & McNamara, 2006; Weir, 2005). Related terms like the rater effect (Myford & Wolfe, 2003, 2004; E. Wolfe, 2004), rater error (Saal, Downey, & Lahey, 1980), or bias rater (Hoyt, 2000; Johnson, Penny, & Gordon, 2009), every touch on this aspect of the fundamental rater variability problem.

Rater variability is not a unity phenomenon but can manifest itself in various forms, each of which calls for careful research. Research has shown that rater can differ not only in the severity or leniency expressed when assessing test takers' performance, but also in the degree to which they adhere to the assessment rubric, in the way they interpret and use criteria in operational assessment sessions, in understanding and using rating scale categories, or the extent to which their rankings are consistent across all examinees, assessment criteria, performance assignments, testing time, and other aspects involved (Bachman & Palmer, 1996; Brown, 2003; Hamp-Lyons, 2007; S C Weigle, 2002).

The usual, or standard, approach to dealing with the variability of rater, especially in high-risk testing, consists of three components: rater training, independent appraisals with the same performance by two or more rater (repeat assessment), and building reliability among rater. The first component, the training rater, usually aims to familiarise the rater with the test format, test assignments, and assessment criteria. More specifically, the rater is trained to achieve a shared understanding of (a) the construct being measured, (b) the level, or level, the performance to be tested, (c) the criteria and associated descriptors that represent the construct. at each level of performance, (d) the scale category or rating scale, and (e) the overall level of difficulty of the items or assignments the examinees must respond to.

Ideally, the differences between rater that might still exist after the assessment should be so small that practically it is not essential; that is, reliability among raters must be as high as possible. Rater usually remains far from functioning alternately even after an extensive assessment process (Barrett, 2001; Cumming, 2007; Eckes, 2004, 2005; Elbow & Yancey, 1994; Hoyt & Kerns, 1999; Knoch, 2009; Kondo-Brown, 2002; Lumley & Mcnamara, 1995; Taylor & Falvey, 2007; Sara Cushing Weigle, 1998), and giving individual feedback to the rater also seems to have no sweeping effect (Catherine Elder, Barkhuizen, Knoch, & von Randow, 2007; Cathie Elder, Knoch, Barkhuizen, & von Randow, 2005). In addition, the selection of trained and experienced trainers has proven to be systematically different in their interpretation of the assessment criteria used routinely. Rather than forming a homogeneous, group that has the same understanding of how to interpret and use criteria, rater falls into the type of rater, with each type marked with a different focus of judgment. For example, some raters show a strong focus on criteria that refer to vocabulary and syntax, while others give more significant weight to accuracy or fluency. (Brown & Hill, 2016; Eckes, 2008).

One of the subjects in mathematics that is often considered difficult by most students when solving problems is the subject of geometry. Geometry is one of the important fields of mathematical studies, but in practice, there are still many difficulties in learning geometry from elementary to tertiary level. There are at least four main topics in the study of mathematics in schools, namely numbers, algebra, geometry and measurement, as well as statistics and opportunities. Geometry is one of the important topics in developing students'

thinking processes. By learning geometry, students will learn about geometric shapes and structures and analyse characteristics and relationships (Uno, 2014). This kind of activity will provide a stimulus to develop the thought process. Meanwhile, according to (Kennedy, Tipps, & Johnson, 2008) Geometry learning activities can activate creativity, develop problem-solving and reasoning abilities and can support other topics in mathematics.

In the problem of geometry, students' geometry skills can influence the success of students in solving a problem, especially geometry problems. There are five basic skills in learning geometry, namely: (1) visual skills, including the ability to: recognise a variety of flat shapes and spaces; classifying buildings based on observed characteristics; identify the centre, axis and plane of symmetry of a shape; summarising information based on visual observations, (2) verbal skills, including the ability to identify various structures according to their names; visualise the building according to its verbal description describing the given shapes and their properties; formulating the definition of the form of words used, (3) drawing skills, including the ability to: draw a given figure and give a mark at certain points; draw shapes according to their verbal definition; draw or construct buildings based on the properties given; drawing or constructing geometry models with examples of the deniers, (4) logic skills, including the ability of students to recognise the differences and similarities between given structures; classifying buildings according to their properties, determining whether a building is included or not in a class, explaining the relationship between buildings, (5) applied skills, including the ability to: recognise physical models of geometric shapes, draw physical models of geometric objects, use geometric models in problem-solving (Patel, Ballam, Strachan, & Northfield, 1996). Geometry subjects in performance assessment were chosen because the scope of the material facilitates students to perform the skills process and can demonstrate their performance including exploring geometric shapes, discovering properties, constructing conjectures and then testing them with proof strategies. Geometry is considered important, because in geometry subjects objects are related to fields and space.

Methods

The performance evaluation of geometry material carried out in junior high school consisted of five sets of questions and five sets of different rubrics for each question. In this paper, the analysed questions referred to as Set-1 were obtained from one junior high school in the Cirebon Region, West Java, Indonesia. Problem Set-1 consists of 5 questions worked on by 250 students, which are then assessed by three teachers, the total data is 5 x 250 x 3 = 3750, and no data is lost in the analysis.

The results of the assessment by three teachers were done by manually scoring, and then data entry was made into the Microsoft Excel program. At the same time, an analysis program was prepared for the Minifac software version 3.7.2. which conducts a double rater analysis using

the Rasch modelling approach (Rasch Model). The program and data are then put together for analysis.

Results and Discussion

Assessing performance instruments, teachers must prepare at least 2 documents, namely: (1) Questions or worksheets or worksheets or work orders; (2) Observation instrument/observation sheet in the form of the rating scale (rating scale).

The observation sheet here is an instrument used to observe the appearance of aspects of performance skills observed. The observation sheet here is a rating scale. The rating scale is a list of questions/statements to assess the quality of implementation of the observed aspects of skills ranging from 1-4.

Arrange Problem Steps: (1) Look at the grid; (2) instruments (indicators) that have been made; (3) Formulate the form of questions or worksheets or worksheets or work orders based on indicators; (4) make answer questions about the scoring guidelines (rubrics). The preparation of the class performance assessment instrument grid refers to the Core Competencies, Basic Competencies, further bias can be seen in table 1.

Table 1: Performance Assessment Instrument Grid

No	Basic competencies	Indicators of Applied Skills	Number Question
1	Determine the position of sides, ribs, angles, lines and areas of flat side spaces	Ability to: recognize physical models of geometric shapes	1, 2
2	Making cube nets, beams, prisms and	Draw physical models of geometric objects	3
3	pyramid	Using geometric models in problem-solving.	4, 5

Problem Analysis Number One

In table 2 shows the output severity assessors provided by the FACETS program. All three appraisers have an acceptable match according to size can be achieved well in the range of 1.50 to 0.50, showing good consistency in the assessment. In the context of the rater aspect, the corresponding index can be interpreted as a measure of intra-rater reliability, where the difference between the fit size and the optimal value of 1.00 indicates the percentage of noise that cannot be explained in the response pattern (Wright & Linacre, 1994). The value of MNSQ infit and outfit for each rater in assessing student performance is to describe the

consistency of the rater in making an assessment. MNSQ Infit and Outfit values for each rater are within acceptable range of 0.5 and 1.50. Readings are within the acceptable range and this shows they have consistently rated each student and subsequently shows that their assessment data has a score of validity.

Table 2: Rater Measurement Report Item Number One

Total Score	Total Count	Obsvd Average	Fair(M) Averag	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N RATER
3508	1500	2.34	2.33	.22	.03	1.02	.7	1.02	.7	.96	.45	.44	1 P1
1318	500	2.64	2.68	-.22	.05	.93	-1.3	.93	-1.2	1.11	.40	.42	2 P2
2413.0	1000.0	2.49	2.51	.00	.04	.97	-.3	.98	-.3		.42		Mean (Count: 2)
1095.0	500.0	.15	.17	.22	.01	.05	1.1	.05	1.0		.02		S.D. (Population)
1548.6	707.1	.21	.25	.31	.02	.07	1.5	.07	1.4		.03		S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .22 Separation 5.04 Strata 7.06 Reliability .96
Model, Fixed (all same) chi-square: 52.9 d.f.: 1 significance (probability): .00

In table 2 shows the output severity assessors provided by the FACETS program. All three appraisers have an acceptable match according to size can be achieved well in the range of 1.50 to 0.50, showing good consistency in the assessment. In the context of the rater aspect, the corresponding index can be interpreted as a measure of intra-rater reliability, where the difference between the fit size and the optimal value of 1.00 indicates the percentage of noise that cannot be explained in the response pattern (Wright & Linacre, 1994). The value of MNSQ infit and outfit for each rater in assessing student performance is to describe the consistency of the rater in making an assessment. MNSQ Infit and Outfit values for each rater are within acceptable range of 0.5 and 1.50. Readings are within the acceptable range and this shows they have consistently rated each student and subsequently shows that their assessment data has a score of validity.

The most important values in Table 2, however, are the z scores from the Infit and Outfit rater measurements; these values indicate that rater 1, has a statistically significant set of responses, while Outfit measurements from other assessors can be interpreted as occurring randomly in this analysis. In practical terms, to manage performance assessments such as translation testing, assessors such as rater 1 will benefit from further panel discussions with other reviewers regarding rubrics that can be interpreted as errors.

Next is looking at the category probability curves for individual evaluators to assist in detecting the effects of central tendencies. The horizontal axis is the scale of the rater's performance; the vertical axis is the probability of observing a certain rank (from 0 to 1).

Separate curves are produced for each ranking scale category. When we look at one of these numbers, the main concern is whether the ranking scale category is widely separated on the logit scale, or not. If categories are placed broadly (and thus have very different peaks), then it might indicate that the evaluator shows a central tendency.

Figure 1. Category Probability Curves for "No Effect" Large Scale Rater item One

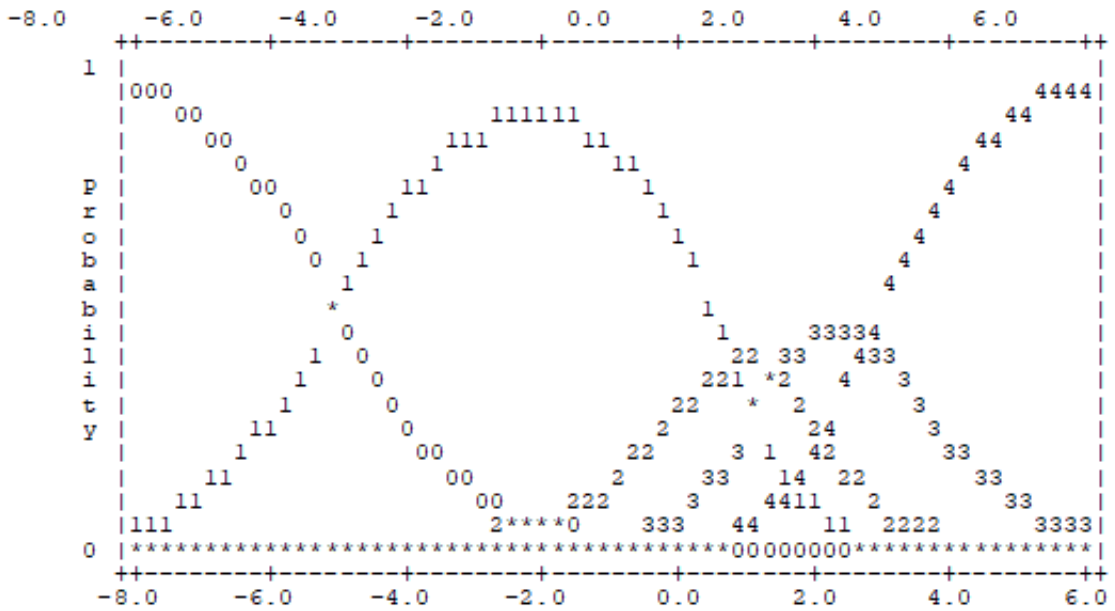


Figure 1 shows the category probability curve from the ranking analysis for Rater which does not show a central tendency. Note that, although there are separate peaks for each category on the scale of nature, the probability curves of individual categories are quite narrow.

Problem Analysis Number Two

The result of item number two analysis of the first is the fixed chi-square test of the hypothesis that all rates show the same calibrated level of performance (ie, that all raters share the same performance measure, after accounting for measurement errors). Non-significant chi-square values indicate the effect of group randomness.

The results of the chi-square rater test are shown in the bottom line of Table 3, Rater Measurement Report. The chi-square value of 45.8 with 2 degrees of freedom is statistically significant ($p < .000$), indicating that there might not be any group-level randomness effect present in this simulation data set.

Table 3: Rater Measurement Report Item Number Two

Total Score	Total Count	Obsvd Average	Fair(M) Averag	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N RATER
1447	600	2.41	2.41	.26	.05	1.12	2.2	1.11	2.1	.84	.44	.40	1 P1
1554	600	2.59	2.60	-.03	.05	.87	-2.5	.88	-2.4	1.15	.36	.40	2 P2
1631	600	2.72	2.73	-.24	.05	1.00	.0	1.00	.0	1.00	.40	.40	3 P3
1544.0	600.0	2.57	2.58	.00	.05	1.00	-.1	1.00	-.1		.40		Mean (Count: 3)
75.4	.0	.13	.13	.20	.00	.10	2.0	.10	1.9		.03		S.D. (Population)
92.4	.0	.15	.16	.25	.00	.12	2.4	.12	2.3		.04		S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .22 Separation 3.78 Strata 7.06 Reliability .96
Model, Fixed (all same) chi-square: 52.9 d.f.: 1 significance (probability): .00

Second, the rater separation ratio. This ratio is a measure of the spread of the rater performance measure relative to the accuracy of the steps. Low-level separation ratios show the effect of group randomness level. The rater separation ratio of 3.78 (shown in the second row from the bottom of Table 3) shows that the spread of rater performance measures is 3 times greater than the accuracy of the measurements. This indicator does not suggest group-level randomness effects. The rater separation index, this indicator links the number of statistically different rater performance levels. The low level separation index shows the effect of group randomness level.

Furthermore, a separation index of 3.78 and reliability of 0.93 indicates that these evaluators show good inter-appraisal reliability. The FACETS program calculates reliability as a measure of variance in a sample; therefore, low values among the appraisal samples mean that appraisers are relatively homogeneous in valuation, desirable features in terms of valuers and convergent validity indicators (Wright, 1996). Although the separation index 5.37 can be considered low, being greater than 1.00 indicates that all three assessors approach heterogeneous rater. Likewise, the fixed-effect test 2 shows that these evaluators are statistically different in rank, that is, the nature of independence in the valuer's parameters can be determined ($X^2 = 45.8$, df 2, p <0,000) which gives a difference in validity. The results of both the separation index and the chi-square value agree and point in the same direction.

Next is looking at the category probability curves for individual evaluators to assist in detecting the effects of central tendencies. The horizontal axis is the scale of the rater's performance; the vertical axis is the probability of observing a certain rank (from 0 to 1).

Separate curves are produced for each ranking scale category. When we look at one of these numbers, the main concern is whether the ranking scale category is widely separated on the logit scale, or not. If categories are placed broadly (and thus have very different peaks), then it might indicate that the evaluator shows a central tendency.

Figure 2. Category Probability Curves for "No Effect" Rater Item Two

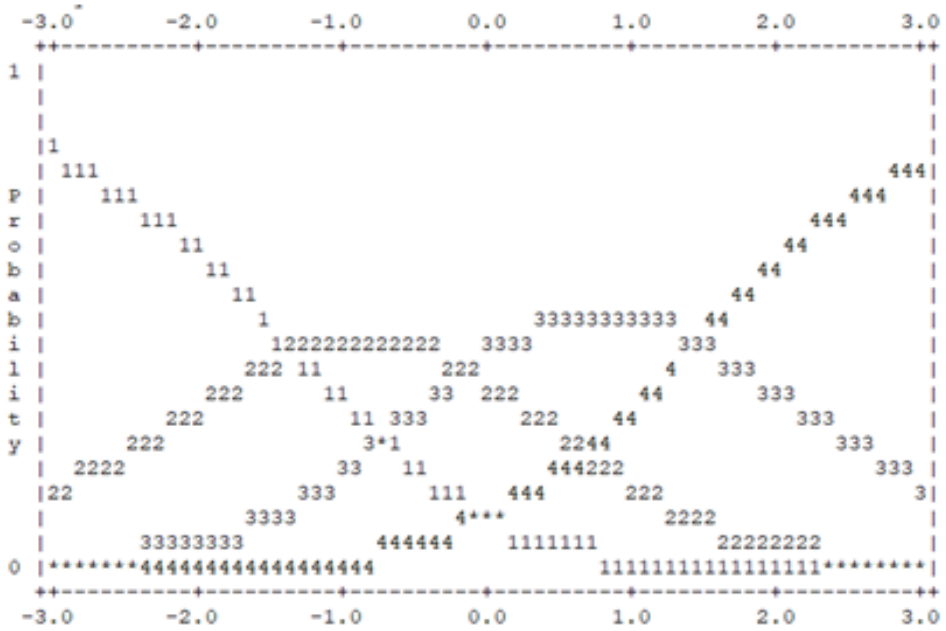


Figure 2 shows the category probability curve from the ranking analysis for Rater which does not show a central tendency. Note that, although there are separate peaks for each category on the scale of nature, the probability curves of individual categories are quite narrow.

Problem Analysis Item Number Three

The results of the analysis of question number three using FACETS software which in problem number four uses the rubric of 3 facet ratings. The results of analysis number three provide a Wright map. Figure 3 presents a map of the steps for the two judging sessions produced by the computer program FACETS. Three raters were included in the original analysis by the essay question.

Table 4: Rater Measurement Report Item Number Three

Total Score	Total Count	Obsvd Average	Fair(M) Averag	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N RATER
1138	500	2.28	2.29	.08	.05	.99	-.2	.99	-.2	1.06	.37	.36	3 P3
1738	750	2.32	2.31	.06	.04	1.00	.0	.99	-.1	1.01	.37	.36	1 P1
1844	750	2.46	2.46	-.14	.04	1.01	.3	1.01	.2	.95	.35	.36	2 P2
1573.3	666.7	2.35	2.36	.00	.05	1.00	.0	1.00	.0		.36		Mean (Count: 3)
310.9	117.9	.08	.08	.10	.00	.01	.2	.01	.2		.01		S.D. (Population)
380.7	144.3	.10	.09	.12	.01	.01	.3	.01	.2		.01		S.D. (Sample)

Model, Populn: RMSE .05 Adj (True) S.D. .08 Separation 1.85 Strata 2.79 Reliability .77
Model, Fixed (all same) chi-square: 14.1 d.f.: 2 significance (probability): .000

The results of the analysis of item number two are the first from the chi-square rater test shown in the bottom line of Table 4, Rater Measurement Report. The chi-square value of 14.1 with 2 degrees of freedom is statistically significant (p <.000), indicating that there may be no group-level randomness effect present in this simulation data set.

Second, the rater separation ratio. This ratio is a measure of the spread of the rater performance measure relative to the accuracy of the steps. Low-level separation ratios show the effect of group randomness level. The rater separation ratio of 1.85 (shown in the second row from the bottom of our Table 4) shows that the spread of rater performance measures is almost 2 times greater than the accuracy of the measurements. This indicator does not suggest group-level randomness effects. The rater separation index, this indicator links the number of statistically different rater performance levels. The low level separation index shows the effect of group randomness level.

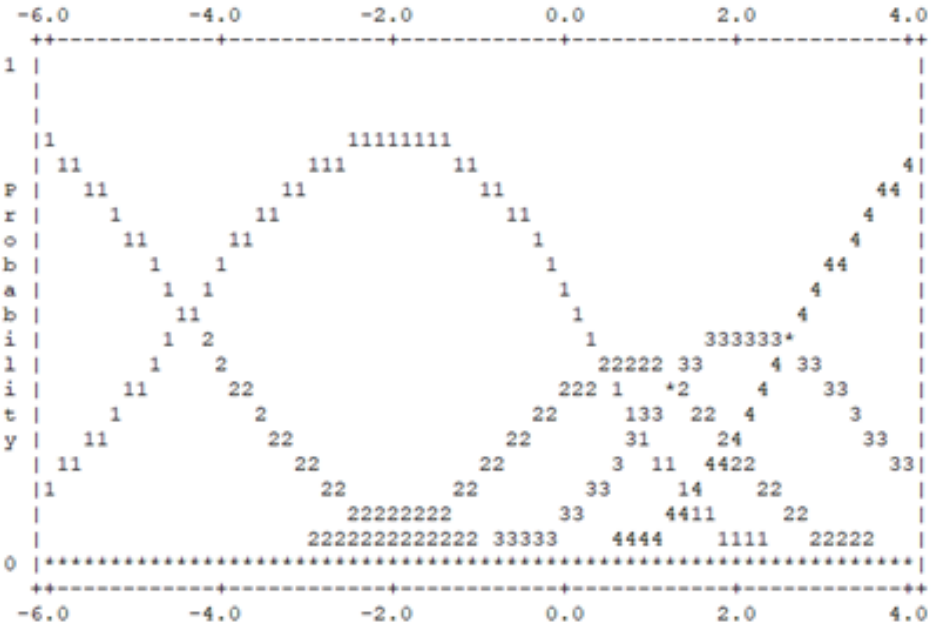
Furthermore, a separation index of 1.85 and reliability of 0.77 indicates that this assessor shows sufficient inter-assessor reliability. The FACETS program calculates reliability as a measure of variance in a sample; therefore, low values among the appraisal samples mean that appraisers are relatively homogeneous in valuation, desirable features in terms of valuers and convergent validity indicators (Wright, 1996). Although the separation index of 1.85 can be considered low, being greater than 1.00 indicates that all three assessors approach heterogeneous rater. Likewise, the fixed-effect test 2 shows that these evaluators are statistically different in rank, that is, the nature of independence in the parameter of the assessor can be established (X^2 = 14.1, df 2, p <0,000) which gives a difference in validity. The results of both the separation index and the chi square value agree and point in the same direction.

Next is looking at the category probability curves for individual evaluators to assist in detecting the effects of central tendencies. The horizontal axis is the scale of the rater's performance; the vertical axis is the probability of observing a certain rank (from 0 to 1).

Separate curves are produced for each ranking scale category. When we look at one of these numbers, the main concern is whether the ranking scale category is widely separated on the logit scale, or not. If categories are placed broadly (and thus have very different peaks), then it might indicate that the evaluator shows a central tendency.

Figure 3 shows the category probability curve from the ranking analysis for Rater which does not show a central tendency. Note that, although there are separate peaks for each category on the scale of nature, the probability curves of individual categories are quite narrow.

Figure 3. Category Probability Curves for "No Effect" Rater Item Three



Problem Analysis Number Four

Next there are four group-level statistical indicators related to rater performance included in the output of the analysis using a ranking scale model.

First, the chi-square test remains from the hypothesis that all rates show the same calibrated performance level (ie, that all raters share the same performance measure, after accounting for measurement errors). Non-significant chi-square values indicate the effect of group randomness.

Table 5: Rater Measurement Report Item Number Four

Total Score	Total Count	Obsvd Average	Fair(M) Averag	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N RATER
1138	501	2.27	2.21	-1.21	.05	1.08	1.5	1.08	1.4	.85	.35	.42	3 P3
1712	749	2.29	2.27	-1.28	.04	1.01	.3	1.02	.3	1.04	.41	.42	2 P2
1790	750	2.39	2.38	-1.41	.04	.93	-1.6	.93	-1.6	1.06	.48	.42	1 P1
1546.7	666.7	2.31	2.29	-1.30	.04	1.01	.1	1.01	.1		.41		Mean (Count: 3)
290.7	117.1	.05	.07	.08	.00	.06	1.3	.06	1.3		.05		S.D. (Population)
356.1	143.5	.06	.09	.10	.01	.07	1.6	.07	1.6		.07		S.D. (Sample)

Model, Populn: RMSE .04 Adj (True) S.D. .07 Separation 1.59 Strata 2.46 Reliability .82
Model, Fixed (all same) chi-square: 10.5 d.f.: 2 significance (probability): .00
Inter-Rater agreement opportunities: 1751 Exact agreements: 498 = 28.4% Expected: 553.4 = 31.6%

The results from the chi-square rater test are shown in the bottom line of our Table 4.18, Rater Measurement Report. The chi-square value of 10.5 with 2 degrees of freedom is statistically significant (p <.000), indicating that there may not be any group-level randomness effect present in this simulation data set.

Second, the rater separation ratio. This ratio is a measure of the spread of the rater performance measure relative to the accuracy of the steps. Low-level separation ratios show the effect of group randomness level. The rater separation ratio of 1.59 (shown in the second row from the bottom of our Table 4.18) shows that the spread of rater performance measures is almost more than 2 times greater than the accuracy of the measurements. This indicator does not suggest group-level randomness effects. The rater separation index, this indicator links the number of statistically different rater performance levels. The low level separation index shows the effect of group randomness level.

Third, the rater separation index of 2.46 shows that almost three strata of rater performance are statistically different in this tariff sample. (Note that the rater separation index does not appear as part of the Facet output. We need to calculate this index manually using the formula (4G + 1) / 3, where G is the rater separation ratio, which is entered as part of Faceted Output. In this example , [4 (1.59) + 1] / 3 = 2.46). Again, there is no evidence here of the effect of group-level randomness.

Fourth, the reliability of the rater separation index. This indicator is a measure of the spread of rater performance measures relative to its accuracy, showing the extent to which appraisers have been able to distinguish reliably between tariffs in terms of their performance. The low-level separation reliability index shows the effect of group-level randomness.

The reliability of the rater separation is shown in the second row from the bottom of our table. A sufficient level of reliability of rater separation (0.77) implies that appraisers can reliably differentiate between rates in terms of their performance. Therefore, this indicator does not suggest group-level randomness effects in this data set.

Next is looking at the category probability curves for individual evaluators to assist in detecting the effects of central tendencies. The horizontal axis is the scale of the rater's performance; the vertical axis is the probability of observing a certain rank (from 0 to 1).

Separate curves are produced for each ranking scale category. When we look at one of these numbers, the main concern is whether the ranking scale category is widely separated on the logit scale, or not. If categories are placed broadly (and thus have very different peaks), then it might indicate that the evaluator shows a central tendency.

Figure 4. Category Probability Curves for "No Effect" Rater Item Four

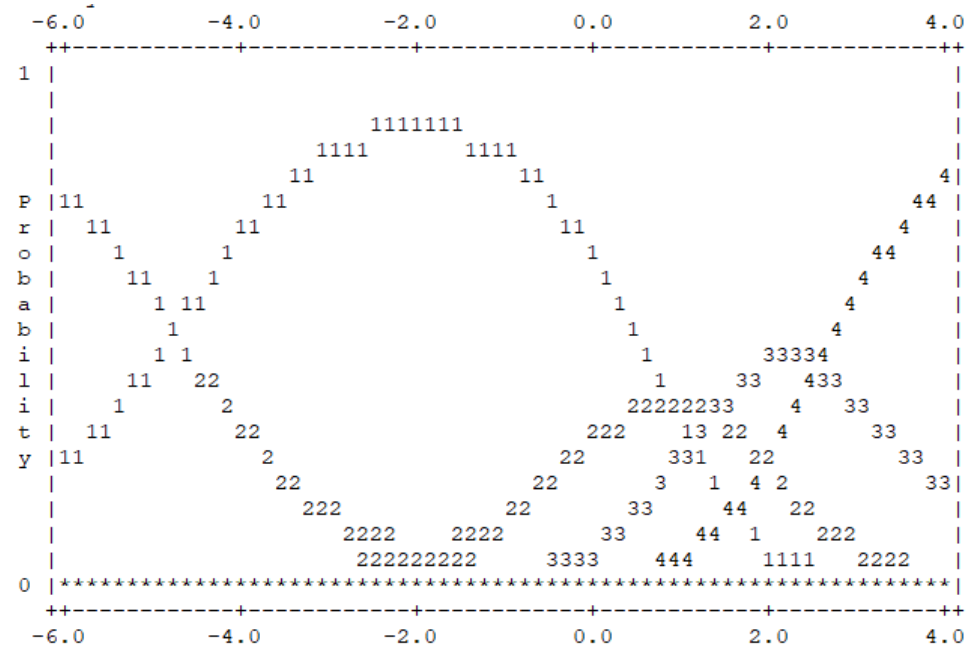


Figure 4 shows a category probability curve from the analysis for Valuers that shows a central tendency to the right. Note that there are separate and distinct peaks for each rating scale category, and that these categories are widespread across the rater performance measurement scale. In general, when the assessor shows a central tendency, it increases the likelihood of observing in the deepest scale categories, and that results in a wide separation from the category threshold, especially in the middle of the rater performance measurement distribution

Problem Analysis Number Five

The results of the first item number five analysis are the chi-square rater test shown in the bottom line of our Table 6, Rater Measurement Report. The chi-square value of 8.9 with 2 degrees of freedom was statistically significant ($p < .001$), indicating that there might not be any group-level randomness effect present in this simulation data set.

Second, the rater separation ratio. This ratio is a measure of the spread of the rater performance measure relative to the accuracy of the steps. Low-level separation ratios show the effect of group randomness level. The rater separation ratio of 2.01 (shown in the second row from the bottom of our Table 6) shows that the spread of rater performance measures is 2 times greater than the accuracy of the measurements. This indicator does not suggest group-level randomness effects. The rater separation index, this indicator links the number of statistically different rater performance levels. The low level separation index shows the effect of group randomness level.

Table 6: Rater Measurement Report Item Number Five

Total Score	Total Count	Obsvd Average	Fair(M) Averag	- Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N RATER
1738	750	2.32	2.31	.31	.04	1.01	.1	1.01	.3	1.05	.44	.40	1 P1
1750	750	2.33	2.33	.29	.04	.95	-1.2	.94	-1.3	1.12	.47	.40	2 P2
1245	500	2.49	2.46	.12	.05	1.07	1.3	1.10	1.8	.76	.22	.40	3 P3
1577.7	666.7	2.38	2.36	.24	.05	1.01	.1	1.02	.3		.38		Mean (Count: 3)
235.3	117.9	.08	.07	.08	.00	.05	1.1	.06	1.3		.11		S.D. (Population)
288.2	144.3	.10	.08	.10	.01	.06	1.3	.08	1.6		.14		S.D. (Sample)

Model, Sample: RMSE .05 Adj (True) S.D. .09 Separation 2.01 Strata 3.02 Reliability .80
Model, Fixed (all same) chi-square: 8.9 d.f.: 2 significance (probability): .001

Third, the rater separation index of 3.02 shows that more than three strata of rater performance are statistically different in this tariff sample. (Note that the rater separation index does not appear as part of the Facet output. We need to calculate this index manually using the formula (4G + 1) / 3, where G is the rater separation ratio, which is entered as part of Faceted Output. In this example , [4 (2.01) + 1] / 3 = 3.02). Again, there is no evidence here of the effect of group-level randomness.

Fourth, the reliability of the rater separation index. This indicator is a measure of the spread of rater performance measures relative to its accuracy, showing the extent to which appraisers have been able to distinguish reliably between tariffs in terms of their performance. The low-level separation reliability index shows the effect of group-level randomness.

Furthermore, a separation index of 2.01 and reliability of 0.80 indicates that these assessors show sufficient inter-assessor reliability. The FACETS program calculates reliability as a measure of variance in a sample; therefore, low values among the appraisal samples mean that appraisers are relatively homogeneous in valuation, desirable features in terms of valuers and convergent validity indicators (Wright, 1996). Although the separation index of 2.01 can be considered low, being greater than 1.00 indicates that all three assessors approach heterogeneous rater.

Next is looking at the category probability curves for individual evaluators to help detect the effects of the central tendency to the right. The horizontal axis is the scale of the rater's performance; the vertical axis is the probability of observing a certain rank (from 0 to 1).
Separate curves are produced for each ranking scale category. When we look at one of these numbers, the main concern is whether the ranking scale category is widely separated on the logit scale, or not. If categories are placed broadly (and thus have very different peaks), then it might indicate that the evaluator shows a central tendency.

Figure 5. Category Probability Curves for "No Effect" Rater Item Five

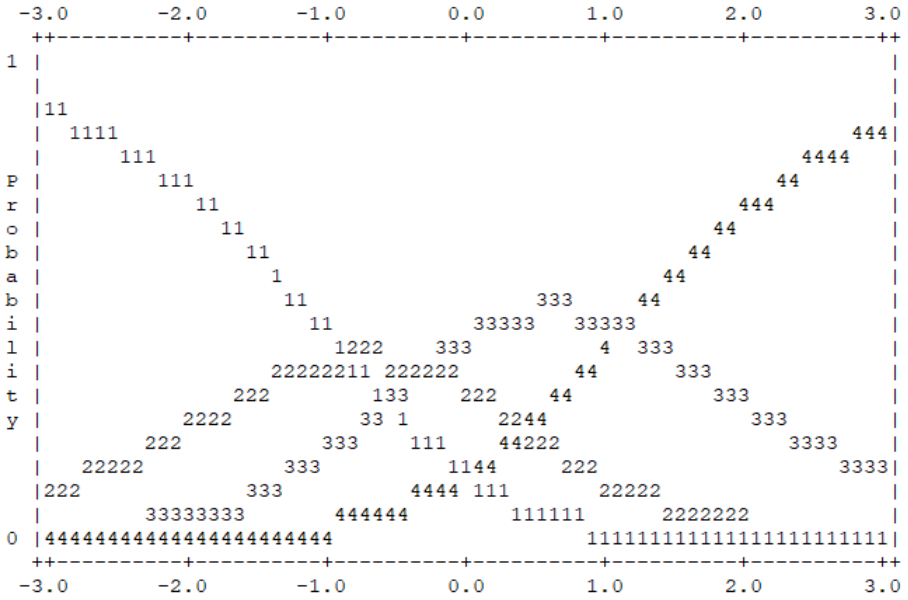


Figure 5 shows a category probability curve from the analysis for Valuers that shows a central tendency. Note that there are separate and distinct peaks for each rating scale category, and that these categories are widespread across the rater performance measurement scale. In general, when the assessor shows a central tendency, it increases the probability of observing in the deepest scale category, and that results in a wide separation from the category threshold, especially in the middle of the rater performance measurement distribution.

Discussion

One possible reason for such results is that students from class eight in junior high school know very well that their performance on the exam will not be counted in their school records, which may limit their performance on the exam. The results might also imply that developing students' abilities to a higher level in completing challenging performance tasks can take longer than we expect. In this case, further evidence is needed before we can make definitive conclusions. However, it is clear that there are no negative effects from using the newly discovered assessment strategy on student performance.

The results of this study indicate that teachers and students are able to handle performance tasks. Although the effect of using the new strategy is in some cases unclear, this study observed a positive effect on students' academic achievement and their level of anxiety about mathematics learning. In addition, it is clear that no negative impacts were observed. In short, overall results appear to support teachers using contextual problems in real life situations and open inquiry in students' mathematics learning.

From this study, then, we conclude that mathematics performance assessment is closely related to direct learning activities that can be developed to provide reliable measures of mathematical achievement, steps that distinguish student learning geometry and that differ from performance test scores. Performance assessments are expensive and time-consuming to develop and manage, and a large number of assignments are needed to provide reliable student achievement estimates.

Of particular concern is the consequential validity of the performance assessment. On a small scale, on average, scores are lower than a large scale on a performance test, and this difference is greater for students at hand than the traditional curriculum. However, based on qualitative analysis on small and large scale in the performance process, the two groups did not differ in their approach to assessment not in the strategy they used to solve the problem and not in their specific types of mistakes. Unfortunately, data are not available to make the same student analysis hand on the curriculum.

Finally, we want to show that this research is the first step for us to explore the possible effects of using performance tasks on teacher teaching and student mathematics learning. More research is needed on the impact of using these new strategies in teaching and learning in various aspects. In addition, given the complexity of teaching and learning practices, there is a long way to go for us to fully understand how the new assessment strategies can be used effectively to improve the quality of teaching and learning, especially with different students (for example, with various abilities).

Conclusion

Performance assessment in all aspects is objective because it does not use pencil and paper tests, but the score depends on the teacher or rater. Therefore, this study shows that appraisal tools or appropriate instruments are needed so that the data can be analysed objectively and fit with a specific measurement model that can assess student performance quantitatively. Analysis of children's response data based on performance assessment using the Rasch Many-facet measurement model is to assess the student's performance process. Each instrument item used during the assessment can be proven by testing the statistical fit. In addition, training can improve assessment standards using consideration and agreement between rater. In conclusion, the Rasch Many-facet measurement model is suitable for use in taking approximately a large number of domains and the components measured are based on performance assessments namely items of performance, the influence of rater and person.

Acknowledgments

Acknowledgments to Mathematics teachers and junior high school students in Cirebon City / Regency who have assisted in the data collection process, also thank Dr. Soeprijanto, M.Pd., and Dr. Dinny Devi Triana, M.Pd who always supports the process of making this article.

REFERENCES

Bachman, L. F., & Palmer, a. S. A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. *Oxford Applied Linguistics.* https://doi.org/10.2307/328718

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human Scoring. In *Automated scoring of complex tasks in computer-based testing*.

Bond, T. G., & Fox, C. M. (2001). The Question of Model Fit. In *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. https://doi.org/10.1111/j.1745-3984.2003.tb01103.x

Brennan, R. L., Robert L. Brennan, & Brennan, R. L. (2006). Educational Measurement. Fourth Edition. ACE/Praeger Series on Higher Education. In *Praeger*.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*. https://doi.org/10.1191/0265532203lt242oa

Brown, A., & Hill, K. M. (2016). On Common Ground? Dow Raters Perceive Scoring Criteria in Oral Proficiency Testing (Thomas Eckes). In *Tasks and Criteria in Performance Assessment*. https://doi.org/10.3726/978-3-653-05394-4/7

Cumming, A. (2007). Book reviews: Lumley, T. 2005: Assessing second language writing: the rater's perspective. Frankfurt: Peter Lang (Volume 3, Language Testing and Evaluation Series, edited by Rüdiger Grotjahn and Günther Sigott). 368 pp. ISBN 3-631-53327-6 US-ISBN 0-8204-7. *Language Testing*. https://doi.org/10.1177/0265532207076366

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge. *Diagnostica*. https://doi.org/10.1026/0012-1924.50.2.65

Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly*. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*. https://doi.org/10.1177/0265532207086780

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing*. https://doi.org/10.1016/1075-2935(94)90006-X

Elder, Catherine, Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*. https://doi.org/10.1177/0265532207071511

Elder, Cathie, Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*. https://doi.org/10.1207/s15434311laq0203_1

Engelhard Jr., G. (2002). Monitoring raters in performance assessments. In *Large-scale assessment programs for ALL students: Development, implementation, and analysis*.

Freedman, S. W., & Calfee, R. C. (1983). Experimental design and cognitive theory. In *Research on writing: Principles and methods*.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*. https://doi.org/10.1016/j.asw.2007.05.002

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*. https://doi.org/10.1037/1082-989X.5.1.64

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*. https://doi.org/10.1037/1082-989X.4.4.403

Johnson, R., Penny, J., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. *Journal of Educational Measurement*.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*. https://doi.org/10.1016/j.edurev.2007.05.002

Kennedy, L. M., Tipps, S., & Johnson, A. (2008). Guiding Children ' s Learning of Mathematics. In *Bulletin of the American Mathematical Society*.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*. https://doi.org/10.1177/0265532208101008

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*. https://doi.org/10.1191/0265532202lt218oa

Linacre, J. M. (1994). Constructing measurement with a Many-Facet Rasch model. In *Objective measurement: Theory into practice: Volume 2*.

Linacre, J., & Wright, B. D. (1994). Dichotomous Mean Square Chi-square fit statistics. *Rasch Measurement Transactions1*.

Lumley, T., & Mcnamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*. https://doi.org/10.1177/026553229501200104

Lund, J. L., & Kirk, M. F. (2010). What is Continuous Performance-Based Assessment? In *Performance-based assessment for middle and high school physical education*.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. https://doi.org/10.1037/0003-066X.50.9.741

Moss, P. A. (1992). Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research*. https://doi.org/10.3102/00346543062003229

Myford, C. M., & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*.

Patel, P., Ballam, L., Strachan, D., & Northfield, T. C. (1996). *All use subject to JSTOR Terms and Conditions PAPERS. 74*(1), 8–9.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.88.2.413

Silvestri, L., & Oescher, J. (2006). Using Rubrics to Increase the Reliability of Assessment in Health Classes. *International Electronic Journal of Health*.

Slomp Prof., D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A canadian study. *Research in the Teaching of English*.

Spandel, V. (2006). Speaking My Mind: In Defense of Rubrics. *English Journal*. https://doi.org/10.2307/30046656

Taylor, L. B., & Falvey, P. (2007). IELTS collected papers : research in speaking and writing assessment. In *Studies in language testing 19*.

Uno, B. H. (2014). Teori Motivasi & Pengukurannya. *Personnel Review*.

Weigle, S C. (2002). Scoring Procedures for Writing Assessments. In *Assessing Writing*.

Weigle, Sara Cushing. (1998). Using FACETS to model rater training effects. *Language Testing*. https://doi.org/10.1177/026553229801500205

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*. https://doi.org/10.1519/15184.1

Wolfe, E. (2004). Identifying Rater Effects Using Latent Trait Models. *Psychology Science*.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolins, L., Wright, B. D., & Masters, G. N. (1983). Rating Scale Analysis: Rasch Measurement. *Journal of the American Statistical Association*. https://doi.org/10.2307/2288670

Wright, B. D., & Stone, M. H. (1979). Best Test Design. Rasch Measurement. In *The measurement model*.

Yudha, R. P., Anggara, D. S., & Zulaeha, O. (2019). Authentic assessment instruments for performance in mathematics learning in elementary schools. *Journal of Physics: Conference Series*, *1321*(3). https://doi.org/10.1088/1742-6596/1321/3/032012