Validity and Inter-rater Reliability of the Scoring Rubrics for the Science Teacher TPACK Test Instrument

(Received 25 April 2021; Revised 31 May 2022; Accepted 31 May 2022)

Raden Ahmad Hadian Adhy Permana[1*], Ari Widodo[2]

[1]Banten Province Educational Quality Assurance Agency, Lebak, Indonesia

[2]Department of Biology Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

Corresponding Author: *r.ahmadhadian@gmail.com

DOI: 10.30870/jppi.v8i1.11164

Abstract

This study examined the content validity of the scoring rubric instrument for measuring science teachers' TPACK and the inter-rater reliability in using the instrument. This research was conducted as part of research and development which has been designed for the development of instruments for measuring teacher knowledge. The analysis carried out was a qualitative analysis based on triangulation of the three validators' validation results and quantitative analysis for inter-rater reliability based on the Intraclass Correlation Coefficient (ICC) obtained for each question. The validation involved three science education experts from the university to assess the suitability of the scoring rubrics in the technological pedagogical content knowledge (TPACK) framework. Inter-rater reliability examining involved 100 participants who answered 15 questions on the instrument and three experienced raters to assess the participants' answers. The validation results showed that the instrument content was valid for measuring the knowledge tested and had very high inter-rater reliability coefficient for all items The validation results show that qualitatively the contents of the instrument are valid for measuring the knowledge being tested and had an average inter-rater reliability coefficient of 0.94 (very high).

Keywords: Content Validity, Inter-rater Reliability, Scoring Rubrics, TPACK

INTRODUCTION

The essay question instrument has become an alternative choice considered better, especially in the aspect of the answer's authenticity compared to multiple-choice questions (Kastner and Stangla, 2011). If essay questions are used as a test instrument, the test taker does not have the opportunity to guess the answer, such as choosing the answer to a multiple-choice question (Rios and Wang, 2018). These advantages become the background for the essay question development as an instrument in high-consequence testing. On a large scale, the application of essay questions for exams has been supported by an automatic scoring system that uses information and communication technology tools (Shermis *et al*., 2010). In general, a good instrument must also have high validity and reliability as its characteristics (Hair, *et al*., 2010).

On the other hand, the obstacle in using essay questions for teacher competency assessment on a large scale is scoring (Williamson *et al*., 2010). Therefore, not much information has been obtained regarding the development of essay instruments for this purpose. Generally, the teacher's competency measurement utilizes objective tests or performance appraisals supported by technology in its application (Ministry of Education and Culture, 2015; Okhremtchouk, Newell and Rosa, 2013). Tests using multiple-choice instruments are deemed the most practical and efficient choice in large-scale teacher evaluations (Sumaryanta *et al*., 2018). Meanwhile, performance appraisal is generally carried out for teacher certification, where the number of participants can be limited in the exam stages (Stacey *et al*., 2020).

The essay question instrument referred to in the above description is a type of question that asks for a limited short essay answer, not a short answer or a long essay (Nitko and Brookhart, 2014). This type of question belongs to the type of open-ended questions or the type of constructed response questions (Wu, Tam, and Jen, 2016). This type of question asks the test taker to answer without being given an answer choice, and the answers requested are generally in the form of a sentence or several sentences that do not compose a paragraph. Essay questions provide several advantages over questions that provide answer choices, namely (1) reducing measurement errors due to random guesses, (b) eliminating corrective feedback to test takers by mistake, and (c) increasing the validity of the test construct (Rios and Wang, 2018). However, the general view for open-ended questions is that they have low reliability, so they are less likely to be used. According to Rademakers, Cate, and Bär (2005), this view may be since open-ended question tests were often not statistically analyzed; whereas, their research results indicate that instruments with open-ended questions can also have adequate reliability, such as multiple-choice questions.

The application of the essay question instrument requires an instrument for a more reliable scoring process, namely a rubric or scoring guide (Brown, 2009; Wallerstedt, Erickson and Wallerstedt, 2012). Rubrics for instrument assessment of essay questions are also known as scoring keys, which contain answer keys, types of answers partially correct, and the scores given (Nitko and Brookhart, 2014). Besides, rubrics are generally part of authentic assessment, especially in performance appraisals (Anderson and Krathwohl, 2001; Jonsson and Svingby, 2007), and the use of rubrics generally aims to improve the consistency of the assessment given (Jonsson and Svingby, 2007). Thus, the subjectivity in essay grading can be reduced due to using the rubric as a predefined evaluation scheme (Moskal, 2000).

The two main categories of rubrics are holistic and analytical, which are used according to the assessment's needs and context (Nitko and Brookhart, 2014). Holistic rubrics, which generally use qualitative generic words, are used to make overall judgments about the quality of a performance or product, while analytic rubrics, which describe quality in detail, are employed to score each dimension in the assignment (Dawson, 2017; Jonsson and Svingby, 2007). Additionally, assessment using a rubric can show the ability to understand and analyze conceptual knowledge and analyze procedural knowledge (Anderson and Krathwohl, 2001). In this essay assessment, the rubric used can be general or specific, depending on the scoring program employed (Burstein and Attali, 2005). Each of the holistic rubrics and analytic rubrics tends to be applied with different strategies, including the interpretation strategy and decision-making by the scoring rater (Li and He, 2015). Weinberger and Guetl (2011) utilized a flexible analytical rubric in their research to test a semi-automatic scoring system, and this rubric can be an alternative as a supporting mechanism for an automatic scoring system.

Furthermore, expert reviews of questions and answers are generally the main process for validating the instrument so that the suitability of the instrument construction and the content tested as well as the adequacy or breadth of the content in the instrument are known (Finch and French, 2018; Wu *et al*., 2016). The validity of the scoring rubric content is also related to the knowledge that the subject should have in the assessment carried out (Moskal and Leydens, 2002). Besides, the content validation of an instrument is not carried out through statistical procedures but by proving the relationship between instrument construction and the content being tested (Finch and French, 2018). In this regard, several previous studies validated the instrument's content through a review process by several experts according to the content tested (Cantabrana, Rodríguez, and Cervera, 2019; Ghazali, 2016; Widiansah, Kartono and Rusilowati, 2019). Validation by these experts can improve the quality of the rubrics and, at the same time, increase their practicality (Haynes, Richard, and Kubany, 1995; Ngang, Nair, and Prachak, 2014).

Moreover, the use of rubrics for broad-scale assessments generally involves more than one rater so that testing of agreement or reliability between raters is needed (Thorndike and Thorndike-Christ, 2014). The reliability between the raters (inter-rater reliability) can indicate the number of error variants present in the scoring results and thus characterizes the precision of the scoring performed by several raters (Finch and French, 2018). Performing a examination for this inter-rater reliability can be done by estimating the Pearson correlation value, the alpha coefficient, the mean correlation between the raters, or the intraclass correlation value (Fleenor, Fleenor and Grossnickle, 1996; Gisev, Bell and Chen, 2013; Harris, Grandgenett and Hofer, 2010; Koo and Li, 2016). This inter-rater reliability examination is also generally carried out on human and computer scoring in research on automatic essay scoring systems as an alternative to human scoring (Attali and Burstein, 2006; Santos, Verspoor and Nerbonne, 2012; Smolentzov, 2012). Notably, the inter-raters reliability estimate results can be the basis for improving the instrument or the decision to directly use the instrument (Harris *et al*., 2010).

The instrument content of the rubrics examined in this study was teacher knowledge within the framework of Technological Pedagogical Content Knowledge (TPACK). The TPACK framework is a concept that is most often the research focus because a deep understanding of technology is needed to use technology in effective learning, communication, problem-solving, and decision making according to the current context (Koehler and Mishra, 2009; Schmid, Brianza and Petko, 2020). This measurement of TPACK for science teachers using the instrument of essay questions and scoring rubrics is an effort to improve the efficiency of the assessment process and still get authentic results. Generally, the TPACK assessment uses a performance appraisal instrument or a questionnaire as an instrument in research (Adi Putra, Widodo and Sopandi, 2017; Agustin and Liliasari, 2017; Bertram and Loughran, 2012; Koehler and Mishra, 2006; Pamuk *et al.*, 2015), but the performance appraisal instrument has constraints when applied to a large number of teachers, and the questionnaire is subjective (Jüttner *et al.*, 2013).

Related to the general characteristics of a test instrument, the questions that become the problem in this research are: 1) How is the content validity of the scoring rubric that has been compiled for the assessment of science teacher knowledge? and 2) How is the inter-rater reliability in testing the scoring rubrics for the assessment of science teacher knowledge? This study aimed to investigate the quality and internal consistency of the scoring rubrics that have been made for the assessment of science teacher knowledge. This study’s results will specifically become the basis for the use of scoring rubrics that have been compiled in implementing the TPACK measurement instrument for science teachers. This study’s results are also expected to become a scoring model for descriptive answers in the teacher's TPACK knowledge test as a novelty in educational assessment.

METHOD

This research was conducted as part of research and development (R&D) designed for the development of instruments to measure teacher knowledge. The R&D procedure includes both qualitative and quantitative test stages. The R&D procedures carried out are: 1) Research and information collecting; 2) Planning; 3) Develop preliminary form of product; 4) Stage 1 testing (qualitative validation); 5) Main product revision; 6) Stage 2 testing (quantitative); 7) Operational product revision; 8) Stage 3 testing (factor analysis); 9) Final product revision; 10) Dissemination and implementation. The research results described in this article are the results of stage 1 and stage 2 testing in the R&D design.

Data collection and procedures

The content validation of the scoring rubrics was conducted by experts in science education, namely three lecturers who teach at the Faculty of Mathematics and Natural Sciences at the Indonesia University of Education, Bandung. Each of the three validators filled out a questionnaire as evidence of validity. The validated aspects include (1) conformity of scoring rubrics with instrument indicators and questions, (2) the accuracy of the concept or the existence of misconceptions, and (3) the criteria for each score. The validators provided a qualitative decision instead of numbers. The results were combined data for further analysis. The content validation reference is the instrument indicator of the question which is compiled based on the TPACK operational framework approach, which consists of 4 aspects: Pedagogical Content Knowledge (PCK), Technological Content Knowledge (TCK), Technological Pedagogical Knowledge (TPK), and Technological Pedagogical Content Knowledge (TPCK) supporting transformative perceptions for the TPACK assessment (Koh, Chai and Tsai, 2013; Angeli, Valanides and Christodoulou, 2016). Following this concept, the content of the instrument is questions that evaluate the teacher's knowledge integrative in planning, implementing, and evaluating learning that integrates technology. Learning content is limited to electrical and photosynthetic materials as part of the science material taught in junior high schools. The instrument contains 15 questions which are divided into PCK questions (4 points), TCK questions (3 points), TPK questions (4 points), and TPCK questions (4 points).

The question instrument has been assessed on 100 junior high school science teachers in the Banten Province, Indonesia. The participants consisted of 38 men and 62 women with varying age ranges, namely 12% less than 30 years, 51% between 30 and 45 years, and 37% more than 45 years. For years of service, the participants were divided into 3 categories, namely new and uncertified teachers with less than 5 years of service (25%), moderately experienced teachers with 5 to 10 years of service (6%), and experienced teachers with more than 10 years of service (69%). The composition of the participants was obtained through convenience sampling. Furthermore, all participant answers were assessed by three scorers, namely the researcher and 2 lecturers of science education. The scoring results become the data for reliability testing between scorers. The three scorers are experienced in assessing essay answers, so no special training was needed to conduct the assessment in this study. The resulting data are numbers in the range 0 to 2 as the scores that have been set in the scoring rubrics for this study. The score data were then processed using SPSS 20 software, following a quantitative procedure to obtain the Intraclass Correlation Coefficient (ICC). Quantitative analysis was also conducted by calculating the percentage agreement between the 3 scorers.

Data analysis

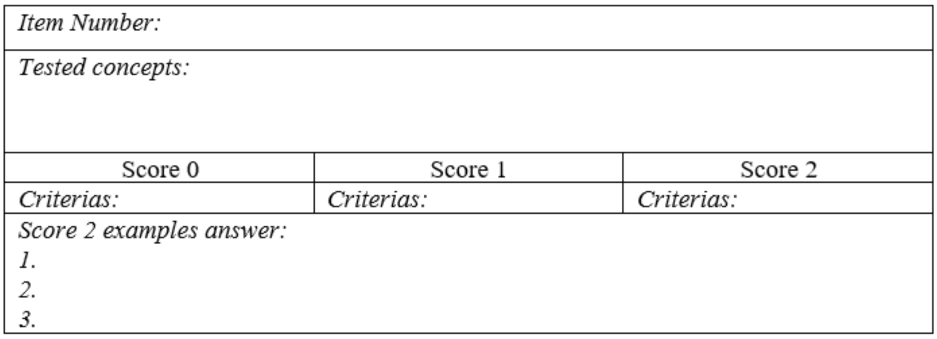
The data analysis was carried out in two stages: the qualitative data analysis stage for content validation results and the quantitative data analysis stage on the inter-rater reliability examination results. The qualitative stage was conducted by analyzing the validation results in the form of decisions and notes from the validators on the issue of scoring rubrics. The analysis results were a triangulation of the assessments given by the validators. Quantitative analysis was performed based on the ICC obtained and the character of each answer in the scoring rubrics.

RESULTS AND DISCUSSION

Description of the TPACK scoring rubrics

The format of the validated and examined answer scoring rubrics in this study is as shown in Figure 1.

The format shown in Figure 1 is the scoring guide or guide for each item. The box "tested concepts" contains an explanation of the concepts being tested on each item of the question. This concept is related to the TPACK framework and indicators that formed the basis for developing questions and answers to relevant problems. The type of essay question used was a limited essay answer so that the answer could be defined but had flexibility in the use of key terms or words. Thus, as a substitute for the answer key, the format was given the concept, criteria, and sample answers with a score of 2 as the highest score. The box "score criteria 0/1/2" covers a detailed explanation of the criteria according to the expected answers. An example of the criteria given for a score of 2 is that "the answer is in the form of two correct, complete, and relevant reasons for the importance of 21st-century skills as student learning outcomes." On the same question, a score of 1 was given on the criterion "only one correct and relevant answer." Meanwhile, for a score of 0 on this question, the criterion was "not giving an answer, or the answer does not match the concept of learning outcomes 21st-century skills". The last box in the rubric format consists of some (maximum 3) examples of answers given a score of 2. These examples of answers were given with considerations, among other things, to increase the similarity of the raters' perceptions because in the research conducted, there was no moderation between the raters as a support for the use of the rubrics (Brown, 2009).



Item Number:		
Tested concepts:		
Score 0	Score 1	Score 2
Criterias:	*Criterias:*	*Criterias:*
Score 2 examples answer: *1.* *2.* *3.*		

Figure 1. TPACK Scoring Rubrics Format

The scoring scale on the rubrics was divided into three levels: 0, 1, and 2. A score of 0 can also be interpreted as having no knowledge or not understanding (absent), a score of 1 means the lack of understanding (deficient) category, and a score of 2 means understanding (sufficient-mastery). The equalization of scores with these categories is in accordance with the instrument preparation’s objectives, namely measuring the teacher’s knowledge level as part of the evaluation, and it is in line to use the rubrics itself, namely determining the level of answers as an expression of the knowledge possessed (Brophy, 2013; Moskal and Leydens, 2002).

Content validation stages

The content of the scoring rubrics, consisting of concepts and criteria, was examined qualitatively to obtain the content validity and determine the instrument’s quality as a guide for giving a score, not based on the scoring system or results (Keith, 2003). Qualitative examining as the initial was considered appropriate because it could measure the accuracy of the concept, depth, and breadth of the assessment tool. The number of experts involved also became a calibration for the quality of the rubrics. The first phase results of qualitative examination are in the table 1.

In Table 1, it is shown that there were eight items considered not valid yet in content, either related to indicators, concepts, or criteria for scoring. If the validator did not give an OK code, it did not mean that the scoring rubrics for an item were immediately considered invalid and had to be discarded or replaced. The items marked with (I), (M), or (C) were then reviewed according to the explanation given by each validator.

Table 1. Summary of the Results of the TPACK Scoring Rubrics Validation

Item	V1	V2	V3
1	OK	(I)	OK
2	OK	OK	OK
3	OK	OK	OK
4	(C)	(I)(M)(C)	(M)
5	(C)	(I)(M)(C)	OK
6	OK	OK	OK
7	OK	OK	OK
8	(C)	(I)(M)(C)	OK
9	OK	(I)(M)	OK
10	OK	OK	OK
11	OK	(I)(M)	OK
12	OK	(I)(M)	(M)
13	OK	OK	OK
14	OK	(I)(M)	OK
15	OK	OK	OK

Note. OK = valid, (C) = innacurate criteria, (I) = problems relate to indicator, (M) = problems in concept

For items marked with (I), the validator (V) considered that the items were not in accordance with the indicators for the questions and answers on the rubric. However, all codes (I) were only given by validator 2, while the other two validators did not give the same code. Thus, the items were then deemed only to need to be corrected according to the notes given by the validator 2. Likewise, for other items, a code (M) or (C) was obtained from a validator only. Meanwhile, for items that received the same mark from two validators, there had to be a fundamental change as an improvement to the rubrics, namely items 4, 5, and 8 on the scoring criteria, and item 12 required conceptual changes. No items got the same mark from the three validators at once, so no scoring rubric had to be replaced entirely or invalid on certain items.

The first aspect measured using these rubrics was the pedagogical content knowledge (PCK) aspect. This first aspect was compiled by items 3, 6, 11, and 14. If the instrument and scoring rubrics for this aspect were valid, the instrument could accurately measure the integration between knowledge of the subject matter and how to teach it comprehensively according to student characteristics (Bilici *et al*., 2013; Shulman, 2015). The validation results showed that the scoring rubrics for this aspect still needed to be improved; among others, in item 11, the concept of the contextual application needed to be clarified, and item 14 was related to the answer to the problem question given as a concept.

The second aspect of the rubrics validated in this study was technological content knowledge (TCK). Measuring the TCK concept is about the use of technology to represent specific topics in learning materials (Bilici, Guzey, and Yamak, 2016; Cox and Graham, 2009; Koehler and Mishra, 2006). This aspect of the TCK was compiled by the answers to questions 2, 4, and 10. The scoring rubrics for item 4 still needed improvement in the concept explanation and the criteria construction. The construction criteria were adjusted to the improvements in the concepts considered inaccurate by validators 1 and 2.

The measurement for the third aspect was technological pedagogy knowledge (TPK). The concept of TPK, which was the operational basis for this research, is the complexity of pedagogical knowledge for learning that utilizes the latest technology. Validation of this aspect is related to the conceptual reference that the teacher has the knowledge to use the right technology tools to facilitate student learning as part of TPACK (Bilici *et al*., 2013; Cox & Graham, 2009; Koehler & Mishra, 2006, 2009). This aspect was validated in the answer rubrics for questions 1, 9, 12, and 13. Improvements in the formulation of criteria were needed to clarify the specifications for the differences in scoring 2 and 1 in the answers to questions 9 and 12.

The final aspect measured in this study was integrating pedagogical content and technology knowledge (TPCK). This aspect could be measured in isolation from a transformative perspective. The general concept reference for this aspect is the knowledge that the teacher has to teach the material specifically using the latest technology and according to the planned strategy (Angeli *et al*., 2016; Cox & Graham, 2009; Jang & Chen, 2010; Koehler & Mishra, 2009). The items for this aspect were numbers 5, 7, 8, and 15. Validation of the answer scoring rubrics for these questions showed that for the answer to question number 5, it was necessary to improve the concept explanation so that it did not cause bias, and number 8 needed to be adjusted with the improvements in the question instrument. The two questions were considered unclear in the criteria for scores 1 and 2, so they needed to be more specific.

The qualitative validation of the rubrics aimed to achieve the validity of the content and construction of the rubrics in accordance with the framework and material concept. Characteristics of essay questions related to this are the tendency of subjectivity and complexity of answers to a given problem (Ornstein, 1992; Valenti, Neri and Cucchiarelli, 2003; Nehm and Haertig, 2012). The answer's subjectivity and complexity will make the correct answer to the question written in various sentences. Therefore, the rubrics made are also expected to capture various answers, not limited to one or two standard sentences as the correct answer.

Quantitative stages

The scoring rubrics assessed in the second stage are the result of improvements based on the results of qualitative testing. The second stage to assess the scoring rubrics in this study is the quantitative test stage. The summary of the results of the reliability calculation in the form of the Intra-class Correlation Coefficient (ICC) and the agreement of the three scorers are shown in Table 2. The ICC obtained from the calculation of the scoring results by the three scorers on the test results of 100 participants was used as the basis for assessing the quality of the scoring rubrics. The percentage agreement was shown to strengthen confidence in the character of the scoring rubric, but it was not used as a determinant of the quality of the rubrics assessed. For this discussion, it is limited to assessing the scoring rubrics so that the results of the validity and reliability tests on the test items that have been prepared are not shown.

The result of the scoring rubrics examination shown in Table 2 was the ICC coefficient for all questions more than 0.9 with a mean of 0.96. This calculation’s results can be interpreted that the scoring rubric instrument in this study had very strong internal consistency and inter-rater reliability (Hair *et al*., 2010; Harris *et al*., 2010; Moskal & Leydens, 2002; Pallant, 2011). Meanwhile, the average percentage of agreement was 84%. This agreement percentage was the equal score given by the three raters. For a score agreed upon by a minimum of two raters, the result was more than 99%, and it means that only <1% of the raters gave different scores from all the answers examined. This indication signifies that the scoring rubrics could support scoring to eliminate the raters’ subjectivity and consistency constraints.

Table 2. The Quantitative Results of the Scoring Rubrics

Item	ICC	Agreement
1	0.97	91%
2	0.94	76%
3	0.97	84%
4	0.95	80%
5	0.98	88%
6	0.98	91%
7	0.95	73%
8	0.94	87%
9	0.96	75%
10	0.96	87%
11	0.98	89%
12	0.99	90%
13	0.92	72%
14	0.99	95%
15	0.96	87%
average	0.96	84%

The examination results based on the aspects in the TPACK framework also revealed that the mean was in a very strong category for the ICC. These results were also supported by the percentage of agreement among the three raters, which reached more than 80% for all aspects. It means that in every aspect, this scoring rubric instrument was adequate quantitatively. The average test value for each of these aspects can be seen in Table 3.

The inter-rater reliability value for the scoring rubric can be influenced by the complexity of the criteria set, the variation in answers given by the subject, and the range of scores used. The high complexity of the criteria will tend to make the rater give more different decisions and a higher range of scores. The advantages of a three-tier scoring rubric, such as those made in this study, include the potential for agreement and inter-rater reliability to be higher than using more scoring levels. Variations in subject answers are not influenced by the rubrics used in scoring but are influenced by the questions' difficulty level and the habit of using terms. The difference in answers can cause different raters to give different scores even though the answers are of the same essence and can reduce scoring agreement and reliability.

Table 3. Average Reliability Value of Each TPACK Aspect

Aspects	Average	
	ICC	Agreement
PCK	0.98	90%
TCK	0.95	81%
TPK	0.96	82%
TPCK	0.96	84%

Another potential is the inconsistency of scoring when the answers given are in the form of explanations in long sentences. Even though the character of this instrument is a short essay, there were still subjects that might answer with quite a long explanation. Some answers exceeded the question request. Like a question requested to make a two-point or two-point answer, the subject answered more than what was asked for. It also has the potential to reduce the consistency of the raters when viewed from the scores given in detail. These possibilities are essential to consider in preparing rubrics for scoring an effective answer (Brown, 2009; Brophy, 2013).

The information provided in Table 1 exhibited that the relationship between the ICC value and the percentage agreement was not always directly proportional. For example, if a comparison was made between question number 2 and question number 7, question number 2 had a higher percentage of the agreement but a lower ICC value than question number 7. Meanwhile, for other questions, most of what can be seen are that a question with a higher ICC reliability value would also have a higher agreement value than other questions or vice versa.

The raters' subjectivity and the consistency weakness are some of the main obstacles in getting objective essay answers. A good scoring rubric is a rubric that can reduce these constraints so that the resulting score has a better confidence value, as expressed by Jonsson & Svingby (2007), Moskal & Leydens (2002), and Nkhoma *et al*. (2020). The rubrics compiled and used in this study had gone through the examining and revision stages. The rubric for scoring the answers was considered to have met the criteria and specifications as a good scoring rubric (Dawson, 2017). The criteria for good scoring rubrics include validity, reliability, and support for the scoring of the question instruments.

CONCLUSION

Qualitative examining by three validators revealed that the scoring rubrics had valid content as an instrument to measure the junior high school science teachers' TPACK on the electricity and photosynthesis material. This feasibility was shown by the triangulation results of the three validators' opinions on the instrument being examined, in which all validators did not provide an evaluation to change the instrument items. The notes given were generally in the form of suggestions for improvement to improve the relationship between the statement and the indicators or the criteria details to determine the answer score clearer.

Inter-rater reliability in the form of the Intraclass Correlation Coefficient (ICC) for all items uncovered that the coefficient was in a very high category (more than 0.9). These results indicate that the scoring carried out by the researcher, as one of the raters and the instrument compilers, had consistency with the scores carried out by the two other raters who were not involved in the instrument preparation. These facts and ICC scores are evidence that the scoring rubrics studied can be used with high confidence and showed sufficient consistency in scoring answers according to the questions to measure science teachers' TPACK.

SUGGESTIONS

This study has not yet produced a ready-to-use scoring rubric instrument for all material content taught by science teachers in junior high schools. The validity and reliability measured in this study might have been sufficient, but for the development of similar instruments on different materials, expert evaluators, or participants, it is possible to have different results. Besides, instrument examining in this study did not take into account the overall science teacher population and a statistically adequate sample. Therefore, this study's results cannot be fully generalized to different contexts. Future research could be undertaken to expand the measured content and measure teacher knowledge on other subjects within a similar framework.

REFERENCES

Adi Putra, MJ, Widodo, A and Sopandi, W 2017, 'Science Teachers' Pedagogical Content Knowledge and Integrated Approach', *Journal of Physics: Conference Series*, vol.895, no.1.

Agustin, RR and Liliasari, L 2017, 'Investigating Pre-Service Science Teachers (PSTs)' Technological Pedagogical Content Knowledge Through Extended Content Representation (CoRe) Investigating Pre- Service Science Teachers (PSTs)' Technological Pedagogical Content Knowledge Through Ext', *Journal of Physics: Conference Series PAPER*, 812

Anderson, LW and Krathwohl, DR 2001, *A Taxonomy for Learning, Teaching, and Assessing A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman. Inc.

Angeli, C, Valanides, N and Christodoulou, A 2016, 'Theoretical Considerations of Technological Pedagogical Content Knowledge', in Herring, MC, Koehler, M J, and Mishra, P (eds) *Handbook of Technological Pedagogical Content Knowledge (TPACK) for Educators*. 2nd edn. New York and Oxon: Routledge, pp. 11–32.

Attali, Y and Burstein, J 2006, 'Automated Essay Scoring With e-rater V.2', *The Journal of Technology, Learning, and Assessment*, vol.4, no.3.

Bertram, A and Loughran, J 2012, 'Science Teachers' Views on CoRes and PaP-eRs as a Framework for Articulating and Developing Pedagogical Content Knowledge', *Research in Science Education*, vol.42, no.6, pp. 1027–47.

Brophy, TS 2013, 'Writing Effective Rubrics'. University of Florida, pp. 1–9.

Brown, GTL 2009, 'The reliability of essay scores: The necessity of rubrics and moderation', in Meyer, LH *et al.* (eds) *Tertiary assessment and higher education student outcomes: Policy, practice and research*. Wellington, N.Z.: Ako Aotearoa, pp. 43–50.

Burstein, J and Attali, Y 2005, 'Automated Essay Scoring With E-rater V. 2.0', *Journal of Technology, Learning, and Assessment*, vol.4, no.3.

Bilici, SC *et al*. 2013, 'Technological Pedagogical Content Knowledge Self-Efficacy Scale (TPACK-SeS) for Pre-Service Science Teachers: Construction, Validation, and Reliability Suggested Citation', *Eurasian Journal of Educational Research*, vol.52, pp. 37–60.

Bilici, SC, Guzey, SS and Yamak, H 2016, 'Assessing pre-service science teachers' technological pedagogical content knowledge (TPACK) through observations and lesson plans', *Research in*

Science and Technological Education, vol.34. no.2, pp. 237–51.

Cantabrana, JLL, Rodríguez, MU and Cervera, MG 2019, 'Assessing teacher digital competence: The construction of an instrument for measuring the knowledge of pre-service teachers', *Journal of New Approaches in Educational Research,* vol.8, no.1, pp. 73–8.

Ghazali, NHM 2016, 'A Reliability and Validity of an Instrument to Evaluate the School-Based Assessment System: A Pilot Study', *International Journal of Evaluation and Research in Education (IJERE),* vol. 5, no.2), pp. 148–57.

Cox, S and Graham, C 2009, 'Using an elaborated model of the TPACK framework to amalyze and depict teacher knowledge', *TechTrends,* vol.53. no.5, pp. 60–9.

Dawson, P 2017, 'Assessment rubrics: towards clearer and more replicable design, research and practice', *Assessment and Evaluation in Higher Education,* vol.42, no.3, pp. 347–60.

Finch, WH and French, BF 2018, *Educational and Psychological Measurement*, *Educational and Psychological Measurement*.

Fleenor, JW, Fleenor, JB and Grossnickle, WF 1996, 'Interrater reliability and agreement of performance ratings: A methodological comparison', *Journal of Business and Psychology,* vol.10, no.3, pp. 367–80.

Gisev, N, Bell, JS. and Chen, TF 2013, 'Interrater agreement and interrater reliability: Key concepts, approaches, and applications', *Research in Social and Administrative Pharmacy*. Elsevier Inc, vol.9, no.3, pp. 330–38.

Hair, JH *et al.* 2010, *Multivariate Data Analysis*. 7th edn. New York: Pearson.

Harris, J, Grandgenett, N and Hofer, M 2010, 'Testing a TPACK-based technology integration assessment rubric', *Teacher Education and Professional Development Commons,* pp. 3833–3840.

Haynes, SN, Richard, DCS and Kubany, E. S 1995, 'Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods', *Psychological Assessment,* vol.7, no.3, pp. 238–47.

Jang, SJ and Chen, KC 2010, 'From PCK to TPACK: Developing a Transformative Model for Pre-Service Science Teachers', *Journal of Science Education and Technology,* vol.19, no.6, pp. 553–64.

Jonsson, A and Svingby, G 2007, 'The use of scoring rubrics: Reliability, validity and educational consequences', *Educational Research Review,* vol. 2, no.2, pp. 130–44.

Jüttner, M *et al.* 2013, 'Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK)', *Educational Assessment, Evaluation and Accountability,* vol.25, no.1, pp. 45–67.

Kastner, M and Stangla, B 2011, 'Multiple choice and constructed response tests: Do test format and scoring matter?', *Procedia - Social and Behavioral Sciences,* vol.12, pp. 263–73.

Keith, TZ 2003, 'Validity and Automated Essay Scoring Systems', in Shermis, M D and Burstein, J (eds) *Automated Essay Scoring: A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, pp. 147–168.

Koehler, MJ & Mishra, P 2006, 'Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge', *Teachers College Record*, vol.108, bo.6, pp. 1017–54.

Koehler, MJ and Mishra, P 2009, 'What Is Technological Pedagogical Content Knowledge?', *Contemporary Issues in Technology and Teacher Education (CITE)*, vol.9, no.1, pp. 60–70.

Koh, JHL, Chai, CS and Tsai, CC 2013, 'Examining practicing teachers' perceptions of technological pedagogical content knowledge pathways: A structural equation modeling approach', *Instructional Science*, vol.41, no.4, pp. 793–809.

Koo, TK and Li, MY 2016, 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research', *Journal of Chiropractic Medicine*. vol.15, no.2, pp. 155–63.

Li, H and He, L 2015, 'A comparison of EFL raters' essay-rating processes across two types of rating scales', *Language Assessment Quarterly*, vol.12, no.2, pp. 178–212.

Ministry of Education and Culture 2015, *Pedoman Pelaksanaan Uji Kompetensi Guru*.

Moskal, BM 2000, 'Scoring rubrics: What, when and how?', *Practical Assessment, Research and Evaluation*, vol.7, no.3, pp. 2000-1.

Moskal, BM and Leydens, JA 2002, 'Scoring Rubric Development: Validity and Reliability', in Boston, C. (ed.) *Understanding Scoring Rubrics A Guide for Teachers*. Maryland: Printing Images. Inc, pp. 25–33.

Nehm, RH and Haertig, H 2012, 'Human vs. Computer Diagnosis of Students' Natural Selection Knowledge: Testing the Efficacy of Text Analytic Software', *Journal of Science Education and Technology*, vol.21, no.1, pp. 56–73.

Ngang, T K, Nair, S and Prachak, B 2014, 'Developing Instruments to Measure Thinking Skills and Problem Solving Skills among Malaysian Primary School Pupils', *Procedia - Social and Behavioral Sciences*. vol. 116, pp. 3760–64.

Nitko, AJ, & Brookhart, SM 2014, *Educational Assessment of Students Sixth Edition*. 6th edn, *Pearson New International Edition*. 6th edn. Essex: Pearson.

Nkhoma, C *et al.* 2020, 'The Role of Rubrics in Learning and Implementation of Authentic Assessment: A Literature Review', in *Proceedings of the 2020 InSITE Conference*, pp. 237–276.

Nunnally, JC and Bernstein, IH 1994, *Psychometric Theory*. 3rd edn. United States: McGraw-Hill.

Okhremtchouk, IS, Newell, PA and Rosa, R 2013, 'Assessing pre-service teachers prior to certification: Perspectives on the performance assessment for california teachers (PACT)', *Education Policy Analysis Archives*, 21(July).

Ornstein, AC 1992, 'Essay Tests: Use, Development, and Grading', *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, vol.65, no.3, pp. 175–7.

Pallant, J 2011, *SPSS Survival Manual: A step by step guide to data analysis using SPSS 4th edition*. 4th edn. Crows Nest NSW: Allen & Unwin.

Pamuk, S *et al.* 2015, 'Exploring relationships among TPACK components and development of the TPACK instrument', *Education and Information Technologies*, vol.20, no.2, pp. 241–63.

Rademakers, J, Cate, TJT and Bär, PR 2005, 'Progress testing with short answer questions', *Medical Teacher,* vol.27, no.7, pp. 578–82.

Rios, JA and Wang, T 2018, 'Essay Items', in Frey, B. B. (ed.) *The SAGE Encyclopedia of Educational, Measurement, and Evaluation*. Thousand Oak CA: SAGE Publications, Inc, pp. 602–605.

Santos, V, Verspoor, M and Nerbonne, J 2012, 'Identifying important factors in essay grading using machine learning', *Language Testing and Evaluation Series (International Experiences in Language Testing and Assessment)*, 28(January), pp. 295–309.

Schmid, M, Brianza, E and Petko, D 2020, 'Developing a short assessment instrument for Technological Pedagogical Content Knowledge (TPACK.xs) and comparing the factor structure of an integrative and a transformative model', *Computers and Education*, 157.

Shermis, MD *et al.* 2010, 'Automated essay scoring: Writing assessment and instruction', *International Encyclopedia of Education*, pp. 20–26.

Shulman, LS 2015, 'PCK It Genesis and Exodus', in Berry, A, Friedrichsen, P, and Loughran, J (eds) *Re-examining Pedagogical Content Knowledge in Science Education*. 1st edn. New York and London: Routledge, pp. 3–13.

Smolentzov, A 2012, *Automated Essay Scoring: Scoring Essays in Swedish*. Stockholm.

Stacey, M *et al.* 2020, 'The development of an Australian teacher performance assessment: lessons from the international literature', *Asia-Pacific Journal of Teacher Education*. vol. 48, no.5, pp. 508–19.

Sumaryanta *et al.* 2018, 'Assessing Teacher Competence and Its Follow-up to Support Professional Development Sustainability', *Journal of Teacher Education for Sustainability*, vol.20, no.1, pp. 106–23.

Thorndike, RM and Thorndike-Christ, T 2014, *Measurement and Evaluation in Psychology and Education.* 8th edn, *Journal of the American Statistical Association*. 8th edn. London: Pearson Education.

Valenti, S, Neri, F and Cucchiarelli, A 2003, 'An Overview of Current Research on Automated Essay Grading', *Journal of Information*

Technology Education: Research, vol. 2, pp. 319–30.

Wallerstedt, S, Erickson, G and Wallerstedt, SM 2012, ‘Short Answer Questions or Modified Essay questions – More Than a Technical Issue’, *International Journal of Clinical Medicine*, vol.3, no.1, pp. 28–30.

Weinberger, A and Guetl, C 2011, ‘Analytical Assessment Rubrics to facilitate Semi-Automated Essay Grading and Feedback Provision Mohammad AL-Smadi’, in *Proceedings of the ATN Assessment Conference*. Curtin University Perth, Western Australia, pp. 170–177.

Widiansah, KN, Kartono and Rusilowati, A 2019, ‘Development of Assessment Instruments Mathematic Creative Thinking Ability on Junior High School Students’, *Journal of Research and Educational Research Evaluation*, vol. 8, no. 1, pp. 84–90.

Williamson, D *et al*. 2010, ‘Automated Scoring for the Assessment of Common Core Standards’, *Educational Testing Service.*, (July).

Wu, M, Tam, HP and Jen, T H 2016, *Educational Measurement for Applied Researchers*, *Educational Measurement for Applied Researchers*.