Pakistan Journal of Life and Social Sciences
www.pjlss.edu.pk

https://doi.org/10.57239/PJLSS-2024-22.2.001046

RESEARCH ARTICLE

Characteristics of The Fifth-Grade Natural and Social Sciences Test in Sukoharjo Regency, Central Java, Indonesia

Suwarto Suwarto[1]*, Nur Rokhimah Hanik[1], Tri Wiharti[1], Arini Hidayah[2], Zakiyah Zakiyah[3] *

[1]Faculty of Education and Teacher Training, Veteran Bangun Nusantara University, Sukoharjo, Indonesia
[2]Surakarta University, Surakarta, Indonesia
[3]Doctoral Student in English Language Education, State University of Malang, Malang, Indonesia

ARTICLE INFO

Received: Oct 14, 2024
Accepted: Nov 21, 2024

Keywords

Distractors
Item Fit
Validity
Reliability
Item Difficulty
Item Discrimination

***Corresponding Author:**

suwartowarto@yahoo.com

ABSTRACT

The purpose of this study: (1) to describe the content validity of the fifth-grade Natural and Social Sciences test, (2) to determine the reliability of the test, (3) to describe the percentage of difficulty levels of the questions: easy: moderate: difficult, (4) to describe the percentage of discrimination of the Natural and Social Sciences test questions, (5) to describe the effectiveness of the distractor, (6) to describe the test with the ability of the test takers, and (7) to describe all items fit with the Rasch Model. The population of the study was all answer sheets for the Natural and Social Science tests of fifth-grade students throughout Sukoharjo Regency, Central Java, Indonesia, totaling 3000 answer sheets. A sample of 188 answer sheets was taken randomly. The test that was studied was a multiple-choice test with four choices. The number of questions was ten. The data in the form of test answer sheets was obtained through documentation. After all the data was collected, the Quest program was used to analyze the data. Research results: (1) The validity of the fifth-grade Natural and Social Sciences test questions has met the validity in terms of content. (2). The reliability of the Natural and Social Sciences test questions is 0.910. (3). The percentage of the level of difficulty of the questions: easy: moderate: difficult = 0%: 90%: 10%. (4). The percentage of the discriminatory item of the questions in the very good category: good = 50%: 50%. (5). There are 30 distractors consisting of 29 effective distractors and 1 less effective distractor. (6). The Natural and Social Sciences test questions have a level of difficulty that is equivalent to the ability of the test participants. (7). All Natural and Social Sciences test questions are fit with the Rasch Model.

INTRODUCTION

Assessment of learning outcomes at the elementary school level plays an important role in ensuring the effective achievement of educational goals. Natural Sciences (IPA) and Social Sciences (IPS) subjects in fifth-grade are crucial subjects, which not only aim to provide theoretical knowledge, but also shape students' scientific and social attitudes. Natural Sciences (IPA) and Social Sciences (IPS) are combined under the name IPAS (Natural and Social Sciences). Evaluation through tests is one of the main methods used to measure the extent to which students have understood the material being taught. However, the quality and characteristics of these tests have often not been analyzed in depth, even though this greatly influences the accuracy of the assessment and the success of learning.

According to Sudijono (2011), learning outcomes assessment is a systematic process for determining the level of achievement of predetermined learning objectives. In this context, the test used as an evaluation tool must have characteristics that meet certain standards, such as validity, reliability, distinguishability, and an appropriate level of difficulty. In Sukoharjo Regency, Central Java, Indonesia, it is still rare to find research that specifically discusses the characteristics of tests in science and social studies subjects in fifth-grade. In fact, as stated by Arikunto (2013), the validity and reliability of assessment instruments is the key to obtaining accurate and fair evaluation results.

This study aims to examine the characteristics of the science and social studies tests used in fifth-grade in Sukoharjo Regency. The focus of this research includes analysis of validity, reliability, distinguishability, and level of difficulty of questions. By carrying out this analysis, it is hoped that a clear picture of the quality of the tests used can be obtained, as well as recommendations for improving the evaluation instrument in the future. It is hoped that the results of this research can make a significant contribution to improving the quality of education in Sukoharjo Regency, as well as becoming a reference for teachers and policy makers in developing more effective evaluation tools.

LITERATURE REVIEW

The literature review in this research will discuss various concepts and previous research that are relevant to test characteristics, especially in the context of IPAS subjects in the fifth-grade of elementary school. This review includes a discussion of validity, reliability, item difficulty, item discrimination, distractors which are the main focus in this research.

Test Validity

Validity is the extent to which a test is able to measure what it is supposed to measure. Validity is an important component in assessment, because it determines the extent to which the test results reflect the ability or knowledge to be measured. According to Messick (1995), validity is not only related to measurement accuracy, but also to the interpretation of test results and the implications of their use. Messick states that validity is the integrity of all the evidence that supports the interpretation and use of a test. Therefore, validity analysis is an important first step in evaluating the quality of a test.

Test Reliability

Reliability refers to the consistency of test results when repeated under the same conditions. A reliable test will provide consistent results if administered under the same conditions to the same group. Brown (2014) in his research shows that test reliability is very important to ensure that test results can be trusted and used as a basis for decision making. Reliability is the primary prerequisite for validity because without reliability, the interpretation of test results will be flawed.

Item Difficulty

The item difficulty of the questions is an important aspect in designing tests, because questions that are too easy or too difficult can affect the test's ability to differentiate between students who understand the material well and those who don't. According to research conducted by Hambleton and Jones (1993), the ideal item difficulty is when the questions are able to clearly differentiate between students who have adequate knowledge and skills and those who do not. An appropriate item difficulty will help create a fair and effective test.

Table 1: The Category of the Item Difficulty

P = The item difficulty	Category
$P > 0.700$	Easy
$0.300 \leq p \leq 0.700$	Moderate
$P < 0.300$	Difficult

(Suwarto at al., 2023).

Item Discrimination

The item discrimination is the ability of a question to differentiate between students with high abilities and students with low abilities. Questions with high differential power will be more effective in assessing variations in ability between students. As stated by Ebel and Frisbie (1991), adequate discrimination is critical to ensuring that the items on a test truly measure differences in ability among students. This research will evaluate the differentiating power of questions to determine the effectiveness of the test in accurately measuring student abilities.

Table 2: The Category of Item Discrimination

Item Discrimination	**Category**
0.40-1.00	Very good
0.30-0.39	Good
0.20-0.29	Fair
0.00-0.19	Poor
Negative r_{pbis}	Low-performing students got the correct answers more than high-performing students

(Dichoso & Joy, 2020).

Relevant Research Reviewer

The research entitled "Characteristics of Fifth-Grade Natural and Social Science Tests in Sukoharjo Regency, Central Java, Indonesia" is an important contribution to the field of educational evaluation. This study provides a comprehensive picture of the quality of tests used to assess student learning outcomes in fifth-grade in IPAS subjects. This research focuses on analyzing test characteristics, including validity, reliability, level of difficulty, and differentiability of questions. Through a systematic approach, the results of this research provide useful insights for educators and policy makers in improving the quality of evaluation instruments.

One of the important findings from this research is the identification of a gap between the curriculum objectives and the characteristics of the tests used. Some of the tests analyzed showed low validity, meaning that they may not fully measure the competencies expected according to the curriculum. Afriani, R. (2019) emphasized the importance of valid and reliable assessment instruments to ensure that the assessments carried out truly reflect students' abilities. This finding is in line with previous studies which show that evaluation instruments are often not valid enough to describe actual learning outcomes.

This research also found that some tests had an unequal level of difficulty, with questions that were either too easy or too difficult, which had the potential to cause distortions in the assessment of students' abilities. This supports Dewi, Patonah, & Sukamto (2023) statement that tests that have the right level of difficulty will be able to differentiate between students with high and low abilities better.

In addition, this research highlights the importance of item discrimination in measuring test effectiveness. Questions with low discrimination power are unable to differentiate between students who really understand the material and students who just answer randomly. This shows the need for improvements in test design to ensure that each question has sufficient differentiation power, as suggested by Wulandari, Ramli, & Muzzazinah (2022).

Overall, this research makes a significant contribution to understanding the characteristics of tests used in elementary schools, especially in Sukoharjo Regency. These findings can form the basis for improvements in planning and developing more effective and fair tests, which in turn could improve the quality of education in the region.

This finding is in line with research conducted by Taherdoost (2016), which states that validity is a key component in the development of effective assessment instruments, as it ensures that the instrument actually measures what it is supposed to measure.

Apart from that, this research also found that the level of difficulty of questions in science and social studies tests is often unequal, which can cause bias in assessing students' abilities. This supports the statement from a study conducted by Brookhart (2018), where he stated that questions with an appropriate level of difficulty can differentiate between students who have different understandings, and it is important to maintain a balanced distribution of difficulty levels in the test.

However, this study also has several limitations, especially in terms of generalization of the results. This research is limited to one district, namely Sukoharjo Regency, so the results may not be generalizable to other regions with different characteristics. Additionally, although this research has provided important insights into test characteristics, further research is needed to explore how these findings can be applied in the context of curriculum development and teaching practices.

Overall, this research provides a strong basis for improving assessment instruments at the elementary school level, especially in science and social studies subjects. It is hoped that these findings can become a reference for educators, policy makers and other researchers in efforts to improve the quality of educational evaluation.

The problem Statement

In the learning process, assessing learning outcomes through tests is one method used to evaluate student achievement. However, not all tests used in the evaluation process have good quality, especially in terms of validity, reliability, items difficulty, items discrimination, and distractors. For this reason, this research seeks to answer several important questions related to the characteristics of IPAS tests in fifth-grade:

1. Do the IPAS tests in fifth-grade have sufficient validity?

2. What is the reliability of the IPAS test in fifth-grade?

3. How do the item difficulty of the questions contain in the IPAS tests in fifth-grade?

4. How do item discrimination in the IPAS tests in fifth-grade?

5. How does each distractor function?

6. Does the test match the student's ability being measured?

7. Do all item fit?

METHOD

The research design is quantitative research. Quantitative research because the researcher calculates test characteristics (level of item difficulty, item discrimination, distractor function, item fit, and test reliability). The research took place from February to July 2024. The research population was all answer sheets on the Natural Sciences and Social Sciences tests of grade 5 students throughout Sukoharjo Regency, Central Java, Indonesia, a total of 3000 answer sheets. A random sample of 188 answer sheets was taken. The test that has been studied is a multiple-choice test with four choices. The total number of questions is ten items. Data in the form of test answer sheets was obtained through documentaries. After all the data is collected, the Quest program is used to analyze the data. Content validation was carried out through studies by measurement and education experts (Sudarwati at al., 2023).

RESULTS AND DISCUSSION

The research results and discussion include aspects of validity, reliability, item difficulty, item discrimination, and distractors. Following are the main results of this study.

Test Validity

This research found that the validity of the tests used in IPAS subjects in fifth-grade had met validity in terms of content. Content validity has been carried out by measurement experts and each indicator in the question grid has been represented in the question items. Content validity is very important to measure student competence. This is in line with the findings presented by Anastasi and Urbina (1997), who stated that validity is the most fundamental thing in test development, because a valid test must be able to measure what it should measure. According to the theory of validity proposed by Messick (1989), validity must be seen as a comprehensive concept that includes a variety of evidence that supports the interpretation and use of tests. Messick asserts that validity includes not only evidence about what the test measures, but also the implications of the test's use.

Test Reliability

The reliability of the IPAS achievement test is 0.910. The reliability of the IPAS test is greater than 0.700, so the reliability of the IPAS test is reliable (Suwarto, 2016). This is in accordance with statements by Nunnally and Bernstein (1994), who emphasized that a test that has high reliability is one that provides consistent and reliable results. This is in line with what was expressed by (Sudarwati at al., 2023), that the instruments used must be reliable. Nunnally (1978) stated that reliability is the basis of any valid form of measurement, because without reliability, the results of the test cannot be trusted.

Item Difficulty

The item difficulty of the questions in the fifth-grade science test varies, with some questions being too easy and others being too difficult. The lowest item difficulty index was 0.239 for question 10 and the highest item difficulty index was 0.647 for question 5. Based on this index, it can be concluded that the fifth-grade science and social studies test question was question 10, while the easiest test item was question 5. The results of item difficulty levels based on categories in the fifth-grade science and social studies achievement tests are presented in table form as follows.

Table 3: The Item Difficulty Result of the IPAS

P = The item difficulty	Category	Items	Total	Percentage
P > 0.700	Easy	-	0	0
$0.300 \leq p \leq 0.700$	Moderate	1,2,3,4,5,6,7,8,9	9	90
P < 0.300	Difficult	10	1	10

Ideally, the distribution of items difficulty should be balanced to ensure that the test can measure different levels of student ability. This research supports the view of Lord (1980), who stated that questions of moderate difficulty are the most effective in differentiating between students of different abilities. This unequal difficulty shows that revisions in question design are needed to achieve a more optimal distribution of items difficulty. The items difficulty of the questions is an important factor that influences the test's ability to measure various levels of student ability. The results of this study indicate that the distribution of items difficulty of questions in the IPAS tests is not balanced. This can result in tests being ineffective in evaluating students' overall abilities. Crocker and Algina (1986) explained that the ideal level of item difficulty should separate students of different abilities, while still providing a fair opportunity to all students. Thus, adjustments need to be made in the design of the questions to achieve a better balance in the item difficulty.

Based on Table 3, it can be concluded that there are 0 questions which are included in the easy category in terms of the item difficulty of the test questions. The percentage of difficulty of questions that fall into the easy category is 0/10 x 100%= 0%. There are 9 items included in the medium category. The percentage of difficulty of questions that fall into the medium category is 9/10 x 100%= 90%. There is 1 question that is included in the difficult category. The percentage of difficulty of questions that fall into the difficult category is 1/10 x 100%= 10%. Based on the percentage of item difficulty for each category, it can be concluded that the most dominant test item difficulty category is the easy category (55%) and the most dominant test item difficulty category is the medium category (90%) and the smallest test item difficulty category is the difficult category (10 %).

Item Discrimination

The item discrimination of questions was also analyzed in this research. There are five items in the test that show very good discrimination, which means that these questions are effective in differentiating between students with high and low abilities. This is in line with the findings of Crocker and Algina (1986), who stated that questions with very different strengths can provide an accurate picture of variations in students' abilities.

The lowest item discrimination index is 0.30 from item 1 and the highest item discrimination is 0.55 from item 7. The results of the item discrimination index by category in the IPAS achievement test are presented in Table 4.

Table 4: The Item Discrimination Result of the IPAS

Item Discrimination	Category	Items	Total	Percentage
0.40-1.00	Very good	3,4,6,7,10	5	50
0.30-0.39	Good	1,2,5,8,9	5	50
0.20-0.29	Fair	-	0	0
0.00-0.19	Poor	-	0	0
Negative r_{pbis}	Low-performing students	-	0	0

Based on table 4, it can be concluded that there are 5 questions that are included in the very good category in the discrimination test items. The percentage of item discrimination that is included in the very good category is 5/10 x 100%= 50%. There are 5 items that are included in the good category. The percentage of item discrimination that is included in the good category is 5/10 x 100%= 50%.

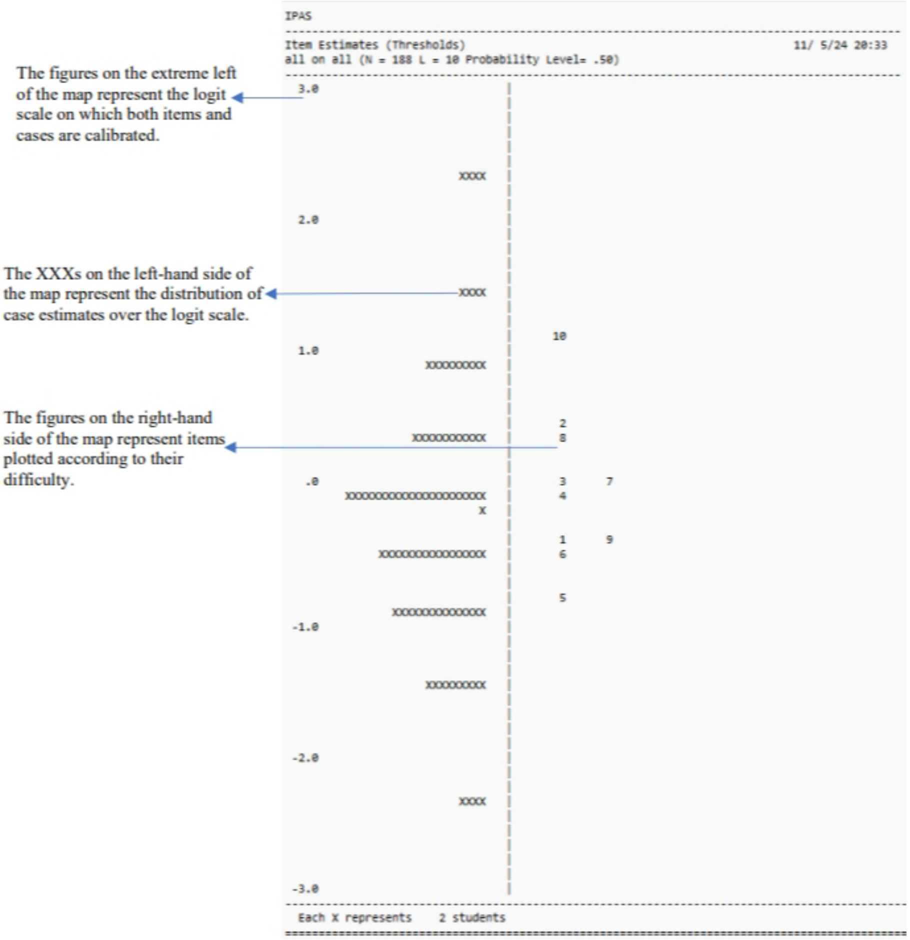
The item discrimination of the questions is the ability of the questions to differentiate between students who have high and low abilities. This research found that there were 50% of questions that were very good and 50% of questions that were good, which means that the questions were effective in assessing variations in ability between students. This is in line with the view of Ebel (1972), who stated that questions with high discrimination can provide an accurate picture of student abilities, making them effective for evaluation purposes.

Distractor

The IPAS test is an objective test in the form of multiple choices. The IPAS test consists of 10 questions. Each question item has four answer choices. One correct answer choice is the answer key, while three wrong answer choices are called distractors. So, the IPAS test has 30 distractors. The results of the analysis using the Quest program obtained 29 effective distractors and 1 less effective distractor. The less effective distractor is distractor D in question item 2. An effective distractor means that the distractor can deceive respondents who have low ability, while a less effective distractor means that the distractor cannot deceive respondents who have low ability. An effective distractor indicates that the distractor was chosen by more than 5% of respondents, while an ineffective distractor indicates that the distractor was chosen by less than 5% of respondents (Suwarto at al., 2023).

Item Map

Figure 1 is a variable map (sometimes referred to as an item map). The variable map shows the distribution of item difficulties and the distribution of case estimates over the variable.



```
IPAS
------------------------------------------------------------------------------
Item Estimates (Thresholds)                                     11/ 5/24 20:33
all on all (N = 188 L = 10 Probability Level= .50)
------------------------------------------------------------------------------
  3.0                          |
                               |
                               |
                               |
                               |
                               |
                          XXXX |
                               |
                               |
  2.0                          |
                               |
                               |
                               |
                               |
                          XXXX |
                               |
                               |
                               |      10
  1.0                          |
                     XXXXXXXXX |
                               |
                               |
                               |
                               |      2
                   XXXXXXXXXXX |      8
                               |
                               |
   .0                          |      3      7
         XXXXXXXXXXXXXXXXXXXXX |      4
                             X |
                               |
                               |      1      9
              XXXXXXXXXXXXXXXX |      6
                               |
                               |
                               |      5
                XXXXXXXXXXXXXX |
 -1.0                          |
                               |
                               |
                               |
                     XXXXXXXXX |
                               |
                               |
                               |
                               |
 -2.0                          |
                               |
                               |
                          XXXX |
                               |
                               |
                               |
                               |
                               |
 -3.0                          |
------------------------------------------------------------------------------
  Each X represents    2 students
==============================================================================
```

The figures on the extreme left of the map represent the logit scale on which both items and cases are calibrated.

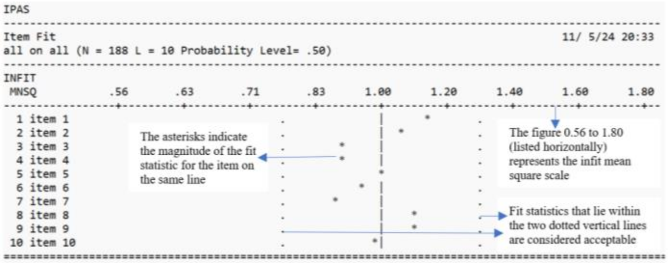
The XXXs on the left-hand side of the map represent the distribution of case estimates over the logit scale.

The figures on the right-hand side of the map represent items plotted according to their difficulty.

Figure 1: Item Map

The right side is the item number, while the left side is the subject distribution where each X represents 2 subjects. The distribution of items and subjects is arranged on the same scale so that in general we can know that the questions have a level of difficulty equivalent to the subject's ability. Item number 10 is the most difficult item. Item number 5 is the easiest item.

Item Fit Map



```
IPAS
------------------------------------------------------------------------------
Item Fit                                                        11/ 5/24 20:33
all on all (N = 188 L = 10 Probability Level= .50)
------------------------------------------------------------------------------
INFIT
 MNSQ         .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.80
----------------+---------+---------+---------+---------+---------+---------+---------+---------+--
  1 item 1                                .          |      *          .
  2 item 2                                .          |  *              .
  3 item 3                                .         *|                 .
  4 item 4                                .         *|                 .
  5 item 5                                .          *                 .
  6 item 6                                .       *  |                 .
  7 item 7                                .     *    |                 .
  8 item 8                                .          |   *             .
  9 item 9                                .          |   *             .
 10 item 10                               .         *|                 .
===================================================================================================
```

The asterisks indicate the magnitude of the fit statistic for the item on the same line

The figure 0.56 to 1.80 (listed horizontally) represents the infit mean square scale

Fit statistics that lie within the two dotted vertical lines are considered acceptable

Figure 2: Item fit map for the IPAS Test

Based on Figure 2, there are 10 IPAS achievement test items because the asterisk is between the two dotted vertical lines (Adams & Khoo, 1996). This shows that all the IPAS test questions fit or match the Rasch Model (one parameter logistic model) with acceptance limits > 0.77 to < 1.30 (Subali & Suyata, 2011). Percentage of conformity of goods= 10/10x100%= 100%. Based on these figures, it can be concluded that the characteristics of the IPAS achievement test are good.

CONCLUSION

The validity of the fifth-grade IPAS test has met the validity in terms of content. The reliability of the IPAS test is 0.910. The percentage of difficulty of the test items: easy: moderate: difficult = 0%: 90%: 10%. The percentage of item discrimination in the very good category: good = 50%: 50%. The results of the analysis obtained 29 effective distractors and 1 less effective distractor. The IPAS test questions have a level of difficulty equivalent to the ability of the test participants. All IPAS test questions fit the Rasch Model.

AUTHORS' CONTRIBUTIONS

SS conceived the idea, designed the project, and wrote the manuscript. NRH collected the data. TW participated in the design of the study and assisted in writing the manuscript. AH conducted the interviews. ZZ performed the data analysis using the Quest program.

ACKNOWLEDGMENT

We would like to thank the Veteran Bangun Nusantara University Sukoharjo for providing research funds and the realization of this scientific article.

REFERENCES

A. A. Dichoso and M. R. J. Joy. (2020). "Test item analyzer using point-biserial correlation and p-values," *International Journal of Scientific & Technology Research*, vol. 9, no. 4, pp. 2122–2126. Retrieved from https://www.academia.edu/85503281/Test_Item_Analyzer_Using_Point_Biserial_Correlation_And_P_Values

Adams, R.J. & Kho, Seik-Tom. (1996). Acer quest version 2.1. Camberwell, Victoria: The Australian Council for Educational Research.

Afriani, R. (2019). Analisis Penilaian Hasil Belajar IPA Pada Implementasi Kurikulum 2013 SMPN 2 Sintang Kalimantan Barat. *Edumedia: Jurnal Keguruan dan Ilmu Pendidikan*, *3*(2). Retrieved from https://jurnal.unka.ac.id/index.php/fkip/article/view/365

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. 7th Edition. Upper Saddle River, NJ: Prentice Hall.

Arikunto, S. (2013). *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Brookhart, S. M. (2018). *Formative Assessment Strategies for Every Classroom: An ASCD Action Tool*. 2nd Edition. *ASCD*.

Brown, J. D. (2014). Principles of language assessment: A holistic approach to test validation. *Language Testing*, 31(1), 85-111. doi:10.1177/0265532213492014.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York, NY: Holt, Rinehart and Winston.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective Teacher Professional Development*. Palo Alto, CA: Learning Policy Institute.

DeMars, C. (2010). *Item Response Theory*. Oxford: Oxford University Press.

Dewi, J. N., Patonah, S., & Sukamto, S. (2023). Validation of Sience, Technology, Engineering, and Matemathic Based Diagnostik Assessment on Natural Resource Material for Phase B Elementary School Students using Rasch Model. *Jurnal Pendidikan Sains Indonesia*, *11*(3), 654-667. Retrieved from https://pdfs.semanticscholar.org/bc5d/02e7610b45d67ee12bc76dda61674d21fdd6.pdf

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. 5th Edition. Englewood Cliffs, NJ: Prentice Hall.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York, NY: American Council on Education/Macmillan.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 253-264. doi:10.1111/j.1745-3992.1993.tb00485.x.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1-73. doi:10.1111/jedm.12000.

Kline, P. (2000). *The Handbook of Psychological Testing*. 2nd Edition. London: Routledge.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. doi:10.1037/0003-066X.50.9.741.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. 3rd Edition. New York, NY: McGraw-Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. New York, NY: Psychology Press.

Subali, B., & Suyata, P. (2011). Panduan Analisis Data Pengukuran Pendidikan Untuk Memperoleh Bukti Empirik Kesahihan Menggunakan Program Quest. Yogyakarta: Lembaga Penelitian dan Pengabdian pada Masyarakat Universitas Negeri Yogyakarta.

Sudarwati, N., Nurhayati, D., Andayani, E., & Suwarto, S. (2023). Effects Of Using a Web Blog in Online Laboratory as A Digital Marketing Platform Towards Students' Achievement with Different Motivation Levels in Entrepreneurship Learning Practicum. *Eurasian Journal of Educational Research*, *103*(103), 173-189. Retrieved from https://ejer.com.tr/manuscript/index.php/journal/article/view/1149

Sudijono, A. (2011). *Pengantar Evaluasi Pendidikan*. Jakarta: Rajawali Pers.

Suwarto, S. (2016). Karakteristik tes Biologi kelas 7 semester gasal. *Jurnal Penelitian Humaniora*, *17*(1), 1-8. Retrieved from https://journals.ums.ac.id/index.php/humaniora/article/view/2346

Suwarto, S., Suyahman, S., Meidawati, S., Zakiyah, Z., & Arini, H. (2023). The COVID-19 Pandemic and The Characteristic Comparison of English Achievement Tests. *Перспективы науки и образования*, (2 (62)), 307-329. Retrieved from https://pnojournal.wordpress.com/2023/05/01/suwarto/

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *International Journal of Academic Research in Management*, 5(3), 28-36. Retrieved from https://hal.science/hal-02546799/

Thorndike, R. M. (1997). *Measurement and Evaluation in Psychology and Education*. 6th Edition. Upper Saddle River, NJ: Prentice Hall.

Wulandari, T., Ramli, M., & Muzzazinah, M. (2022). Analisis butir soal dynamic assessment untuk mengukur pemahaman konsep klasifikasi tumbuhan pada mahasiswa. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, *10*(1), 191-201. Retrieved from https://jurnal.usk.ac.id/JPSI/article/view/22082/0