

QUALITY ANALYSIS OF TEACHER-MADE TESTS IN FINANCIAL ACCOUNTING SUBJECT AT VOCATIONAL HIGH SCHOOLS

Heni Mulyani^{1*}, Heraeni Tanuatmodjo¹, Rangga Iskandar¹

¹Universitas Pendidikan Indonesia

Jl. Dr. Setiabudi No.229, Isola, Sukasari, Kota Bandung, Jawa Barat 40154, Indonesia

Abstract

Assessment of student learning outcomes needs to be done using tests that meet the criteria for quality tests. This study aims to determine the quality of teacher-made tests on financial accounting subjects in Vocational High Schools. This research is descriptive research with a quantitative approach. Data collected are questions made by 32 teachers, answer sheets from 689 Accounting students. The validity of objective and essay tests using product-moment correlation. Reliability of the objective test using the KR20 formula, while the essay test using Alpha formula. Difficulty level and distinguishing power of objective tests using Anates 4. Difficulty level and distinguishing power of essay tests used Microsoft Excel 2013. The research results obtained are as follows: (1) validity of teacher-made test items cannot accurately measure learning outcomes; (2) reliability of teacher-made tests cannot show stable results despite repeated testing of the same subject; (3) teacher-made tests do not have a proportion of degree of difficulty that is suitable for use as a Mid-Semester assessment tool; (4) distinguishing power of tests made by teachers cannot distinguish students who have mastered the test material (upper or superior group) from students who have not mastered the test material (lower group or user); (5) Multiple choice test distractors made by teachers are not evenly chosen, and the key options and deception options do not function effectively. Quality analysis of teacher-made tests through item analysis is intended to identify damaged test items and to show areas that are already mastered by students.


Keywords: *teacher-made test, test quality, financial accounting, reliability, validity*

How to cite: Mulyani, H., Tanuatmodjo, H., & Iskandar, R. (2020). Quality analysis of teacher-made tests in financial accounting subject at vocational high schools. *Jurnal Pendidikan Vokasi*, 10(1), 1-9. doi:<https://doi.org/10.21831/jpv.v10i1.29382>



***Corresponding Author:**

Heni Mulyani  henimulyani@upi.edu

 Department of Accounting Education, Faculty of Economics and Business Education, Universitas Pendidikan Indonesia

Jl. Dr. Setiabudi No.229, Isola, Sukasari, Kota Bandung, Jawa Barat 40154, Indonesia

INTRODUCTION

Based on a review of research results on teacher-made tests (Ashtiani & Babaii, 2007; Carroll & Moody, 2006; Marso & Pigge, 1991), several problems were found: (1) teachers view tests designed by the teacher as positively affecting teaching and learning; (2) most of the tests developed by teachers contained many errors; and (3) teachers usually do not use test improvement strategies such as test blueprints or item analysis. Teacher-made tests are usually tests that refer to the teacher's criteria to assess and evaluate student mastery of certain knowledge (Wiggins, 1989). Research on teacher-made tests has been carried out by researchers in various countries, including Notar et al. (2004), DiDonato-Barnes et al. (2014), and Ing et al. (2015) who examined the use of specification tables to present the validity of teacher-made tests. Meanwhile, Kinyua and Okunya (2014) investigated not only the validity of teacher-made tests but also their reliability. In comparison, Quaigrain and Arhin (2017) evaluate tests developed by teachers using reliability and item analysis. These studies are basically conducted to assess the quality of teacher-made tests, as stated by Walker et al. (2004) that well-made and well-managed teacher-made tests can provide evidence of quality learning and teaching. Given the test's prevalence as a very common means of determining student learning, it is necessary to focus on the characteristics and basic principles to build a good test for the class (Grant & Gareis, 2015).

The assessment of students' success in mastering learning material carried out in class is done by the teacher using tests. Wiggins (1989) says that teacher-made tests usually refer to criteria designed by the teacher to assess and evaluate student mastery of certain knowledge. Before the test that has been made by the teacher is used to assess students, several criteria must be met so the test meets the criteria for quality tests. Validity is an attribute to deduce the validity of a test based on a score and requires the use of a test score. On the other hand, an instrument-based approach states that the test is either inherently valid or invalid (Kinyua & Okunya, 2014). Formative validity seeks to determine the extent to which a test's ability can provide information that can help improve the way to achieve the goals of a program. For example, in an assessment for learning, the aim is to gather the information that will improve teaching methods that benefit students (Clark, 2008).

Reliability is one of a series of test scores that shows the number of measurement errors associated with the score. Teachers should know about reliability so they can use test scores to make the right decisions about their students. Frisbie (1988) stated that the level of consistency of a set of scores can be predicted using internal analysis methods to calculate the reliability coefficient. Meanwhile, Meshkani and Abadie (2005) declared that test reliability refers to the conditions to which the instrument can produce the same results in repeated trials or the tendency towards consistency found in repeated measurements is called reliability.

However, Heyneman and Fägerlind (1988) explained that the requirements that must be met to make a quality test are not only limited to the validity and reliability of the questions, but other requirements that must be met are also difficulty level, distinguishing power, and effectiveness of distractors. The main purpose of item analysis in a teacher-made test is to identify deficiencies in the test or in learning. Teacher-made tests that have not been analyzed can reduce the quality of the tests themselves, because deficiencies in teacher-made tests have not been detected before use. The deficiencies in teacher-made tests can obscure information on the level of student learning progress. This information should not be used to make decisions related to learning. Thus, the impact of teacher-made tests that have not been analyzed should be eliminated. Nevertheless, studies of the quality of teacher-made tests still examine the scope of validity, reliability of the questions, and difficulty levels, such as the study of Notar et al. (2004), DiDonato-Barnes et al. (2014), Cooper et al. (2014), Kinyua and Okunya (2014), Khairani and Shamsuddin (2016), and Quaigrain and Arhin (2017). Thus, this study is to examine the quality of teacher-made tests by analyzing the five criteria: validity, reliability, difficulty level, distinguishing power, and effectiveness of distractors.

RESEARCH METHOD

This research was a descriptive study with a quantitative approach. This study's population was a test made by financial accounting teachers in 32 Vocational High Schools majoring in Ac-

counting. The sample in this study was the same as the population, so this study uses a census. The technique used in collecting research data was the documentation technique. The data were generated from 32 teachers' tests, answer sheets from 689 students in class XI Accounting, assessment guidelines, and answer keys for midterm assessment in financial accounting subjects for the academic year 2017/2018. There were 302 items analyzed, consisting of 129 items (42.72%) of multiple-choice tests made by teachers and 173 items (57.28%) essay tests with teachers' limited answers.

Testing the validity of items was done using the Anates 4 application. The formula used to test the validity of objective test items and essay test items is the product-moment correlation formula. The reliability test was carried out using Microsoft Excel 2013. The formula used to test the reliability of the objective test was the KR20 formula. The formula used to test the reliability of essay tests was the Alpha formula. Difficulty level testing on objective tests was done using Anates 4. Difficulty level testing on essay tests was done using Microsoft Excel 2013. The formula used was adjusted according to the form of the test.

Distinguishing power testing on objective tests was done using the Anates 4 application. Distinguishing power testing on essay tests was carried out using Microsoft Excel 2013. The formula used was adjusted according to the form of the test. The deception of quality testing is carried out using the Anates 4 application. To determine the deception quality used the deception index formula. To increase the accuracy of the test results, the effectiveness of the option function was tested. The effectiveness of the option function is analyzed using Microsoft Excel 2013.

The provisions used to determine the effectiveness of key options are as follows: (1) The number of voters in the upper and lower groups is between 25% - 75%. The test was carried out using the following formula:

$$\frac{\sum PKA + \sum PKB}{n_1 + n_2} \times 100\%$$

Annotation:

$\sum PKA$ = the number of top group voters; $\sum PKB$ = the number of voters in the lower class; n_1 = the number of sample groups above (27%); n_2 = the number of sample groups below (27%).

(2) The number of voters in the upper group must be greater than the number of voters in the lower group. The provisions used to determine the effectiveness of fraud options are as follows: (a) The number of voters in the upper and lower groups is not more than $= 25\% \times \frac{1}{2(\sum a)} \times (Ka + Kb)$. (b) The number of voters in the lower class must be greater than the number of voters in the upper group.

Annotation:

d= number of deception options; Ka = top group; Kb = bottom group.

RESULTS AND DISCUSSION

Results

The description of the results of the analysis of the quality of tests made by accounting teachers are as follows.

Item Validity

The results of the analysis of the validity of multiple-choice test items show that 129 multiple choice test items made by teachers in financial accounting subjects, there were 40 items (31.01%) declared valid, while the remaining 89 items (68.99%) were declared invalid. The results of the analysis of the validity of essay test items with limited answers showed that 173 items essay test questions with limited answers made by teachers on financial accounting subjects, there were 124 items (71.68%) was declared valid, while the remaining 49 items (28.32%) were declared invalid.

Reliability

The results of the analysis of the reliability of multiple-choice tests show that among the 129 multiple choice test items made by teachers in financial accounting subjects, there were 21 items (16.28%) declared reliable while the remaining 108 items (83.72%) were declared unreliable. The results of the analysis of the reliability of essay tests with limited answers showed that 173 items essay test questions with limited answers made by teachers in financial accounting subjects, there were 108 items (62.43%) declared reliable, while the remaining 65 items (37.57%) were declared unreliable.

Difficulty Level

Difficulty level analysis should be done on teacher-made tests. If it is related to this research object, the mid-semester assessment should be built from items with a moderate degree of difficulty. Thus a quality teacher-made test for mid-semester assessment needs to be constructed from items with a moderate degree of difficulty or at least the proportion of items with a moderate degree of difficulty than the proportion of difficult and easy items. Based on the analysis of the level of difficulty, it can be seen that in general, the teacher-made tests on financial accounting subjects do not have a proportion of degree of difficulty that is feasible to be used as a mid-semester assessment. The results of the analysis of the difficulty level of multiple-choice tests show that 129 items of multiple-choice test questions made by teachers in financial accounting subjects, there are 28 items (21.71%) declared difficult, 39 items (30.23%) were stated to have a moderate level of difficulty, while the remaining 62 items (48.06%) were declared easy. Based on the results of the analysis, it can be seen that the proportion of items with a moderate degree of difficulty on teacher-made compound choice tests on financial accounting subjects is not greater than the proportion of difficult and easy items. The results of the analysis of the difficulty level of essay tests with limited answers showed that 173 items essay test questions with limited answers made by teachers in financial accounting subjects, there were 17 items (9.83%) declared difficult, 39 items (22.54%) were stated to have a moderate level of difficulty, while the remaining 117 items (67.63%) were declared easy. Based on the results of this analysis, it can be seen that the proportion of items with a moderate degree of difficulty on essay tests with limited answers made by teachers on financial accounting subjects is no greater than the proportion of difficult and easy items.

Distinguishing Power

Based on the results of the analysis of distinguishing power, it can be seen that in general, teacher-made tests on financial accounting subjects cannot distinguish students who have mastered the test material (upper or superior) and students who have not mastered the test material (lower group).

The results of the analysis of the differentiation power of multiple-choice tests showed that of 129 multiple choice test items made by teachers in financial accounting subjects, there were 54 items (41.86%) declared to have power an adequate differentiator, while the remaining 75 items (58.14%) were declared not to have adequate distinguishing power. The results of the analysis of distinguishing essay tests with limited answers showed that of the 173 items essay test questions with limited answers made by teachers in financial accounting subjects, 108 items (62.43 %) was stated to have adequate distinguishing power, while the remaining 65 items (37.57%) were stated not to have sufficient differentiating power.

Effectiveness of Distractor

When referring to the results of the analysis of deception quality, it can be seen that the proportion of distractors who have poor quality is 224 deception (43.41%), which is the largest proportion of deception quality. In addition, 52 outfits (10.08%) were also found of unknown quality. If all test takers choose a key option, and no one chooses the deception provided, the deception option's quality cannot be known. This is due to the ease of the questions so that test-takers can easily choose the key options and ignore the deception options provided.

Suppose the results of the analysis of the quality of the deception are related to the results of the analysis of the effectiveness of the deception options. In that case, it can be seen that the large proportion of the quality of the deceivers is bad. The presence of options of unknown quality has caused the proportion of distractors who are declared not to function as effectively as 342 deception options (66.28%) is greater than the deception option declared effective, that is, 174 deception options (33.72%). In general, the deception options do not function effectively because the number of deception options voters in the lower group is not greater than the number of deception options voters in the upper group. This shows that the deception option cannot outwit students who have not yet mastered the test material (lower group). Besides, generally, the deceptive option sentences are not homogeneous.

When referring to the results of the analysis of the effectiveness of the function of the key options, it can be seen that the proportion of ineffective key options is 74 key options (57.36%) greater than the declared effective key options, which are 55 key options (42.64%). This shows that the key options provided are not able to direct test participants to the correct answers. In general, the key options for multiple-choice tests made by teachers of financial accounting subjects are not well organized. This is due to the ineffective preparation of key option sentences, which results in different interpretations among test takers. In addition, the large proportion of ineffective key options is also influenced by the existence of answer keys that are not relevant to the question matter. There are 17 key options that are not relevant to the question matter.

Discussion

Learning outcomes test is declared valid if the test can measure learning outcomes appropriately, as Kinyua and Okunya (2014) stated that referring to the simplest point of view, a test can be judged valid if it measures what is meant to be measured. Nordin (2002) stated that valid tests can lead to information or grades taken to help teachers and students make judgments, conclusions, and figures of speech about achievement quality. Analysis of the validity of items should be conducted on teacher-made tests. Learning outcome assessment data must be obtained in accordance with reality. Popham (2009) stated that good evaluation data in accordance with reality are called valid data, to obtain valid data, the instrument or tool to evaluate it must be valid. Thus all items made by teacher tests must be declared valid in order to become a quality test. Based on the results of the analysis of the validity of the items, it can be seen that in general, teacher-made tests on financial accounting subjects cannot measure learning outcomes accurately. Factors that influence teacher-made tests on financial accounting subjects are not all valid, namely the item validity index is not greater than and is influenced by factors related to questions and answer keys. Weaknesses of teacher-made tests that do not meet the item validity requirements can be avoided if the teacher has carried out an item validity analysis before the test is used. In addition, teachers need to optimize the factors that affect the validity of test results. Winter et al. (2006) explained that there are several factors that affect the validity of test results, including the evaluation instrument, evaluation and scoring administration factors, and student response factors. The teacher can use data items that are not valid as a reference to correct deficiencies in the evaluation instrument. In relation to the administrative factors of evaluation, the results of the analysis of the validity of the items can be used as a reference to study the allocation of time given. In relation to student answers, the teacher should provide answer sheets. Besides, Black et al. (2010) stated that teachers can respond to problems of validity by reflecting on their values and by engaging in the joint development of portfolio assessments.

Whereas, learning outcomes tests are declared to be reliable if the tests can show stable results even though they are repeatedly tested on the same subject. As Grant and Gareis (2015) explained, a good instrument is an instrument that can consistently provide data that is in accordance with reality. Thus, the teacher-made test must be declared reliable or consistent in order to become a quality test. Parkes (2013) stated that basically, the reliability measurement principles reveal the consistency of test-takers or assessors throughout the measurement opportunity. Based on the reliability analysis results, it can be seen that in general, teacher-made tests on financial accounting subjects cannot show stable results despite repeated testing of the same subject. The

factors that influence essay tests with limited answers made by teachers on financial accounting subjects are not reliable, namely, the test reliability coefficient (r_{11}) is not greater than r_{table} . In addition, there are factors that influence the reliability of the test related to the test itself. This is relevant to the opinion of Levy and Goldstein (2014) that reliability can be influenced by matters relating to the test itself, namely the length of the test and the quality of the problem items. The length of the test relates to the number of test items. The more the number of items, the more steady a test becomes. Weaknesses of teacher-made tests that do not meet the test reliability requirements can be avoided if the teacher has conducted a test reliability analysis before the test is used. In addition, if the teacher wants to increase the number of test items in order to optimize the test reliability coefficient (r_{11}), then the addition of the number of items needs to pay attention to the quality of the items.

The factors that influence essay test with limited answers made by teachers on financial accounting subjects are not reliable, namely, the test reliability coefficient (r_{11}) is not greater than r_{table} . In addition, some factors influence the reliability of the test related to the test itself. This is relevant to Arikunto (2012) opinion that reliability can be influenced by matters relating to the test itself, namely the length of the test and the quality of the problem items. The length of the test relates to the number of test items. The more the number of items, the more steady a test. Weaknesses of teacher-made tests that do not meet the test reliability requirements can be avoided if the teacher has conducted a test reliability analysis before the test is used (Kusaeri & Suprananto, 2012). In addition, if the teacher wants to increase the number of test items to optimize the test reliability coefficient (r_{11}), then the addition of the number of items needs to pay attention to the quality of the items. Linn and Gronlund (2000) suggested that the general definition of the reliability principle stated that reliability means the extent to which measurement tools can produce consistent readings.

Besides, the weaknesses of teacher-made tests that do not meet the difficulty level requirements can be avoided if the teacher has carried out an analysis of the difficulty level of items before the test is used. To obtain tests with a proportion of items with a moderate degree of difficulty greater than the proportion of difficult and easy items, the teacher can use the provisions of the proportion of difficulties that are normally distributed.

Wright (2007) clarified the optimal difficulty for each item depends on the teacher's assessment and testing objectives, it is known that for the purpose of selection, items used that have a high degree of difficulty, and for diagnostic purposes are usually used items that have a low level of difficulty/easy. Therefore difficult and easy items can be reused as needed. Meanwhile, according to Nordin (2002), an item's difficulty illustrates the percentage of students who can answer an item correctly.

Other than that, a distinguishing analysis should be done on teacher-made tests. Items of learning achievement test items must be able to provide test results that reflect differences in abilities found among the tessees. If the item discrimination index is interpreted as being moderate, good, and very good, then it can be concluded that it has adequate differentiation of items. On the contrary, if the item discrimination index is interpreted poorly, then it can be concluded that it does not yet have the distinguishing power of items as expected. To be able to be concluded that it has sufficient differentiation of items, the total proportion of items that are stated to have moderate, good, and excellent differentiation must reach 100%. Thus quality teacher-made tests need to be built from items that have adequate differentiation.

Factors influencing items that cannot distinguish students' abilities are (a) the key to the item answer is incorrect; (b) the item has two or more correct answer keys; and (c) the deception doesn't work. Weaknesses of teacher-made tests that do not meet the distinguishing power requirements can be avoided if the teacher has carried out a distinguishing power analysis before the test is used. In addition, the teacher can determine the appropriate action on the results of the analysis of the power of differentiation. First, for items with sufficient differentiation (having moderate, good, and excellent differentiation), the teacher can put them in the question bank for reuse or development. Second, for items with poor differentiation, the teacher can choose to discard them or explore the factors that cause the differentiation of items to be bad. If the causative factor has been found, the

item can be fixed and used for the next test. Third, for items with very poor distinguishing features, it should be discarded. This is relevant to the follow-up that needs to be done by a tester of the analysis results of distinguishing power (Sudijono, 2015).

Qualified teacher-made tests for multiple-choice need to be built from items with evenly chosen deceivers, meaning that the deceivers have very good or good quality. In addition, to increase accuracy, key options, and deception options should be declared to function effectively (Reynolds et al., 2010). Based on the results of the analysis of the distractors' effectiveness, it can be seen that in general, the multiple-choice test distractors made by financial accounting teachers were not evenly selected, and not all key options and deception options were declared to be functioning effectively.

Weaknesses of teacher-made tests that do not meet the requirements of the distractor's effectiveness can be avoided if the teacher has carried out an analysis of the effectiveness of the distractor before the test is used. The teacher can also determine the right action on the results of the analysis of the effectiveness of the distractor. First, for items with evenly selected distractors and key options and deception options to function effectively, the teacher can put them in the question bank for reuse or development. Second, for items with distractors that are not evenly selected, and key options and deception options are not functioning effectively, the teacher can choose to fix them or replace them with new distractors. Hamzah and Abdullah (2011) stated that a distracter is said to be effective if the candidate, who does not know the answer, chooses the distracter as the answer.

Therefore, Lee and Lee (2013) and Young and Kim (2010) explained that teacher-made tests that have not been analyzed can reduce the tests themselves' quality because deficiencies in teacher-made tests have not been detected before use. The deficiencies in teacher-made tests can obscure information about the level of student learning progress. This information should not be used to make decisions related to learning. Thus the overall analysis of these items is a stage that must be done by the teacher as the opinion of Mitra et al. (2009), item analysis is the process of gathering, summarizing, and using information from student responses to assess the quality of test items. The resulting item statistics can be used to determine good items that need to be repaired or deleted from the question bank. Whereas according to Bichi (2015), the two objectives of the Item analysis are; firstly, to identify defective test items and secondly, to indicate subject matter that students have and have not mastered.

CONCLUSION

Based on the research findings, some conclusions are drawn as follows. (1) The validity of teacher-made test items on financial accounting subjects cannot measure learning outcomes accurately. In the multiple-choice test sample, the proportion of items that were declared valid was 31.01%, whereas, in the essay test sample with limited answers, the proportion of items that were declared valid was 71.68%. (2) The reliability of teacher-made tests on financial accounting subjects could not show stable results despite repeated testing of the same subject. In the multiple-choice test sample, the proportion of sample units that were declared reliable was 16.67%, whereas, in the essay test sample with limited answers, the proportion of sample units that were declared reliable was 62.50%. (3) Teacher-made tests on financial accounting subjects do not have a proportion of the degree of difficulty that is feasible to be used as a Mid-Semester Assessment Tool. In the multiple-choice test sample and essay test with limited answers, the proportion of items with moderate difficulty level is not greater than the proportion of difficult and easy items. (4) The distinguishing power of teacher-made tests in financial accounting subjects cannot distinguish students who have mastered the test material (upper or superior group) from students who have not mastered the test material (lower group). In the multiple-choice test sample, the proportion of items that were stated to have adequate distinguishing power was 41.86%, whereas, in the essay test sample with limited answers, the proportion of items that were stated to have adequate distinguishing power was 62.43%. (5) The multiple-choice test distractor made by teachers in financial accounting subjects was not evenly chosen, and the key options and deception options were declared ineffective. Thus, teachers need to optimize their competence in preparing

the test, taking into account the factors of quality teacher-made tests, namely item validity, reliability, difficulty level, distinguishing power, and distractor effectiveness.

ACKNOWLEDGMENT

Acknowledgments and awards are given to the colleagues of the Department of Accounting Education of Universitas Pendidikan Indonesia for providing advice, assistance, and convenience.

REFERENCES

- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*. Bumi Aksara.
- Ashtiani, N. S., & Babaii, E. (2007). Cooperative test construction: The last temptation of educational reform? *Studies in Educational Evaluation*, 33(3–4), 213–228. <https://doi.org/10.1016/j.stueduc.2007.07.002>
- Bichi, A. A. (2015). Item analysis using a derived science achievement test data. *International Journal of Science and Research (IJSR)*, 4(5), 1656–1662.
- Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice*, 17(2), 215–232. <https://doi.org/10.1080/09695941003696016>
- Carroll, T., & Moody, L. (2006). Teacher-made tests. *Science Scope, September*, 66. <https://www.questia.com/library/journal/1G1-153041130/teacher-made-tests>
- Clark, I. (2008). Assessment is for learning: Formative assessment and positive learning interactions. *Florida Journal of Educational Administration & Policy*, 2(1), 1–16.
- Cooper, T., Ashley, P., & Simona, W. (2014). Using reliability, validity, and item analysis to evaluate a teacher-developed test in international business. *Evaluation and Testing Research Article*, 1–11.
- DiDonato-Barnes, N., Fives, H., & Krause, E. S. (2014). Using a table of specifications to improve teacher-constructed traditional tests: an experimental design. *Assessment in Education: Principles, Policy & Practice*, 21(1), 90–108.
- Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, 7(1), 25–35.
- Grant, L., & Gareis, C. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning*. Routledge.
- Hamzah, M. S. G., & Abdullah, S. K. (2011). Test item analysis: An educator professionalism approach. *US-China Education Review*, A(3), 307–322.
- Heyneman, S. P., & Fägerlind, I. (1988). *University examinations and standardized testing: Principles, experience, and policy options*. World Bank.
- Ing, L. M., Musah, M. B., Al-Hudawi, S. H., Tahir, L. M., & Kamil, N. M. (2015). Validity of teacher-made assessment: A table of specification approach. *Asian Social Science*, 11(5). <https://doi.org/10.5539/ass.v11n5p193>
- Khairani, A. Z., & Shamsuddin, H. (2016). Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. In *Assessment for learning within and beyond the classroom* (pp. 417–426). Springer Singapore. https://doi.org/10.1007/978-981-10-0908-2_35
- Kinyua, K., & Okunya, L. O. (2014). Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu district of Kenya. *African Educational Research Journal*, 2(2), 61–71.

- Kusaeri, K., & Suprananto, S. (2012). *Pengukuran dan penilaian pendidikan*. Graha Ilmu.
- Lee, J., & Lee, Y. S. (2013). Effects of testing. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 416–418). Routledge.
- Levy, P., & Goldstein, H. (2014). *Tests in education: A book of critical reviews*. Academic Press.
- Linn, R. L., & Gronlund, N. (2000). *Measurement and assessment in teaching* (8th ed.). Prentice-Hall.
- Marso, R. N., & Pigge, F. L. (1991). An analysis of teacher-made tests: Testing practices, cognitive demands and item construction errors. *Contemporary Educational Psychology*, *16*, 279–286.
- Meshkani, Z., & Abadie, F. H. (2005). Multivariate analysis of factors influencing reliability of teacher made tests. *Journal of Medical Education*, *6*(2), e105155. <https://doi.org/10.22037/jme.v6i2.765>
- Mitra, N. K., Nagaraja, H. S., Ponnudurai, G., & Judson, J. P. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *IeJSME*, *3*(1), 2–7.
- Nordin, M. S. (2002). *Testing and interpreting in classrooms*. International Islamic University of Malaysia Publications.
- Notar, C. E., Zuelke, D. C., Wilson, J. D., & Yunker, B. D. (2004). The table of specifications: Insuring accountability in teacher made tests. *Journal of Instructional Psychology*, *31*(2). <https://www.questia.com/library/journal/1G1-119611686/the-table-of-specifications-insuring-accountability>
- Parkes, J. (2013). Reliability in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 107–123). SAGE Publications.
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, *48*(1), 4–11. <https://doi.org/10.1080/00405840802577536>
- Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. <https://doi.org/10.1080/2331186X.2017.1301013>
- Reynolds, C. R., Livingston, R. B., & Willson, V. (2010). *Measurement and assessment in education* (2nd ed.). Pearson Education.
- Sudijono, A. (2015). *Pengantar evaluasi pendidikan*. Raja Grafindo Persada.
- Walker, C. M., Schmidt, E., & Mototsune, K. (2004). *Smart tests: Teacher-made tests that help students learn*. Pembroke Publishers Limited.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, *70*, 703–710.
- Winter, P. C., Kopriva, R. J., Chen, C.-S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, *16*(4), 267–276. <https://doi.org/10.1016/j.lindif.2007.01.001>
- Wright, R. J. (2007). *Educational assessment: Tests and measurements in the age of accountability*. Sage Publications.
- Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives*, *18*(19), 1–40. <https://doi.org/10.14507/epaa.v18n19.2010>