CARTESIUS: Jurnal Pendidikan Matematika
ISSN: 2828-6847 (online)
Vol. 9, No. 1, Juni 2026, pp. 100-106
https://ejournal.ust.ac.id/index.php/CARTESIUS

VALIDITY TEST OF SEVENTH-GRADE DAILY TEST ITEMS MTsN 1 PASAMAN

Chelsy Naila Friska[1], Pameladia[2], Fitri Ulfa Aktafia[3], Fatihah Risqi Hasani[4], Bunga Anjelia[5], Popy Khofifah[6], Zulfitri Aima[7], Anna Cesaria[8]
[12345678]Program Studi Pendidikan Matematika, Fakultas Sains dan Teknologi, University PGRI Sumatera Barat

Informasi Artikel

Sejarah Artikel:

Diterima, April 24, 2026
Revisi, Mei 25, 2026
Disetujui, Juni 26, 2026

Keyword:

Item validity, Daily test, Learning evaluation

ABSTRACT

This study aims to determine the validity level of daily test questions for seventh-grade students MTsN 1 Pasaman, in order to find out whether the questions used are appropriate and reliable for measuring students' abilities. Question validity is an important aspect to analyze because invalid questions can produce data that does not accurately reflect students' actual abilities. The method used in this study is descriptive quantitative, analyzing data from the daily test results of seventh-grade students MTsN 1 Pasaman. The obtained data were then processed using a validity test to determine the accuracy of each item in measuring what it is intended to measure. The results show that among the tested questions, all items were found to be valid, with calculated correlation coefficient (r) values greater than the table r-value at a 5% significance level. The valid questions showed a strong and significant correlation between item scores and total scores. Based on this analysis, it can be concluded that the daily test items used have met the validity requirements, indicating that the evaluation instrument is reliable for accurately measuring students' abilities. Teachers are nevertheless encouraged to continue conducting regular item analysis to maintain the consistency of instrument quality.

Corresponding Author:

Anna Cesaria,
Program Studi Matematika, Fakultas Sains dan Teknologi,
University PGRI Sumatera Barat.
Email: annacesaria13@gmail.com

1. INTRODUCTION

Learning evaluation is an essential component of the educational process because it serves as a measuring tool to determine the extent to which learning objectives have been achieved. One common form of evaluation used by teachers in schools is the daily test, which functions to periodically monitor students' understanding of the material that has been taught. The instruments used in daily tests, in the form of questionnaires, surveys, or test items, must be tested for validity and reliability before being used as data-collection tools (Hartanto et al., 2023). If the items used are not tested for feasibility, the results obtained have the potential to not reflect students' actual ability, which can mislead teachers in making decisions related to the subsequent learning process.

Several previous studies have discussed the importance of item analysis at various levels and subjects. Fiska et al. (2021) analyzed science daily test items using a classical test theory approach and found that not all items used by teachers met the valid criteria. In line with this, Himawan and

Nurgiyantoro (2022), who analyzed Indonesian language final-semester assessment items for eighth grade, also reported variation in item quality in terms of validity and difficulty level. Research at a more specific level, namely seventh grade, also shows similar findings. Yusuf (2021), in an item analysis of the final mathematics examination for seventh grade, found that of the items tested, only a small portion were declared valid, while the rest were invalid and required revision. Similarly, Kasanova and Sulistiyono (2023), who analyzed odd-semester daily test items, recommended that teachers continually conduct item analysis before the items are used, in order to determine the degree of instrument validity early on.

More recent studies have expanded item analysis to cover not only validity but also reliability, difficulty level, and discrimination index, thereby providing a more comprehensive picture of an instrument's quality. Anshari et al. (2024) analyzed the validity and reliability of odd-semester summative test items for Islamic Religious Education and found considerable variation in item quality, such that a number of items required revision. Dianova and Anwar (2024), who studied Arabic summative test items at the elementary school level, reported that most valid items fell into the easy difficulty category, with discrimination indices ranging from poor to good across items. In the same vein, research on literacy ability test quality found that validity, reliability, difficulty level, and discrimination index need to be examined together to determine whether an instrument is truly fit for use (Rohmah & Mauliddiyah, 2025). In the context of mathematics evaluation, research on even-semester final test items for eighth grade concluded that the items used were appropriate as a learning evaluation tool because they met the established validity and reliability requirements (Pratiwi & Marzal, 2024).

The validity of a test instrument can be determined through the calculation of the correlation between item scores and total scores using the Pearson Product Moment formula, in which an item is declared valid if the calculated r-value (r-calculated) is greater than the r-table value at a certain significance level (Anggraini et al., 2022). This approach has been widely used in educational research because it is relatively simple yet provides a clear picture of the feasibility of each item. In line with this, Amini (2023) emphasized that validating an instrument before it is used in the field is an important step to ensure that the data collected truly reflects the construct intended to be measured, particularly for instruments aimed at elementary and secondary school students. Nevertheless, research on the validity of daily test items, specifically for seventh grade at MTsN 1 Pasaman using students' actual test result data, still needs to be continually examined, given that item characteristics and student abilities vary across schools and subjects.

Based on the description above, a question arises as to whether the daily test items used for seventh-grade students at MTsN 1 Pasaman meet the validity requirements, and to what extent the items are accurate in measuring student ability. Therefore, this study aims to determine whether the seventh-grade daily test items used at MTsN 1 Pasaman are valid or invalid through a validity test based on students' test result data, so that it can serve as a reference for teachers in improving the quality of the evaluation instruments used in the classroom.

Theoretically, the quality of a test item can be analyzed through two main approaches, namely classical test theory (CTT) and item response theory (IRT). The CTT approach is more widely used in school-level educational research because its calculation procedure is relatively simple and does not require specialized software, even though its results are sample-dependent (Rohmah & Mauliddiyah, 2025). Within the CTT framework, the validity of an instrument does not stand alone, but consists of several complementary types, namely content validity, construct validity, and criterion validity, each of which addresses an aspect of instrument feasibility from a different perspective (Lestari et al., 2025). Content validity focuses on the alignment between test items and the test blueprint and the material taught, while criterion validity—including item validity through the correlation of item scores with total scores—focuses on empirical evidence from students' test results (Ariyanto et al., 2023).

Recent studies in mathematics education also reinforce the urgency of conducting item analysis on an ongoing basis. Cahyaningrum et al. (2023), who analyzed odd-semester summative mathematics test items for seventh grade with the assistance of Anates software, found that not all items composed by teachers met the valid criteria, and thus recommended routine evaluation of the item banks used in schools. Relatively similar findings were obtained at the senior high school

level; research on multiple-choice mathematics test items for twelfth grade at SMAN 1 Adiankoting reported that although the majority of items met the validity and reliability criteria, improvements were still needed for a number of items so that the test instrument would be more accurate and consistent in measuring student competence (Siregar et al., 2024). Other research in the context of algebra items at the junior high school level even showed a very high proportion of valid items, namely nine out of ten items tested, with one item requiring revision because its correlation value fell below the r-table threshold (Qohar & Fauziyah, 2024). This variation in findings shows that the level of item validity is strongly influenced by school context, grade level, and student characteristics, such that validity testing cannot be generalized from one population to another without re-verification (Welsandt et al., 2024).

In addition to validity, several researchers have emphasized that the factors influencing the accuracy of a test item can originate from both internal and external aspects of the instrument. Internal factors include the clarity of item instructions, the alignment of difficulty level with student ability, and the quality of item sentence construction, while external factors include test administration conditions, such as the time allocation for completing the test, which can affect students' psychological state when answering items (Habibi et al., 2023). This shows that the validity of a test item is not merely a statistical matter, but is also related to the quality of instrument design and the testing conditions themselves. Therefore, validity testing based on empirical data from students' test results, as conducted in this study, needs to be complemented by an examination of item construction quality so that the resulting conclusions are more comprehensive and can serve as a basis for thoroughly improving evaluation instruments.

2. RESEARCH METHOD

This study uses a descriptive quantitative approach. Descriptive quantitative research is a method aimed at producing an objective picture or description of a particular condition using numbers, starting from data collection, data interpretation, to the presentation of results (Arikunto, 2006). In line with this, Sugiyono (2012) explains that descriptive research is research conducted to determine the value of a variable, whether one or more variables, without intending to make comparisons or relate it to other variables. The quantitative approach itself is characterized by the use of numbers at every stage of the research, from collection, interpretation, to the presentation of research result data (Arikunto, 2013). This approach was chosen because the research is descriptive and analytical in nature, examining students' daily test result data objectively without providing any treatment or intervention to the research subjects.

Item analysis in this study is based on the classical test theory (CTT) framework, which views a student's total score as a representation of true ability (true score) plus a measurement error component. This approach was chosen because its procedure is relatively simple and is widely applied in school-level educational evaluation research, particularly for small-scale test instruments such as daily tests (Lestari et al., 2025). The validity tested in this study falls into the category of criterion validity, namely validity determined through the correlation between each item's score and the student's total score, as is commonly used for instruments that have already been empirically administered in the field (Ariyanto et al., 2023). The use of this approach is consistent with a number of previous studies that have also applied the product moment correlation technique to assess item validity in learning achievement test instruments across various levels and subjects (Cahyaningrum et al., 2023; Siregar et al., 2024).

The instrument analyzed in this study consists of 5 seventh-grade mathematics daily test items that had been administered to students on material previously delivered by the teacher. Prior to the empirical validity test, the items had been constructed based on a test blueprint aligned with the learning outcomes, so that they conceptually satisfied the element of content validity. However, this study specifically focuses on establishing statistical validity through students' test result data, rather than on content validation by experts (expert judgment), given that the research objective is to evaluate items that have already been used directly in the classroom. The selection of the Pearson Product Moment correlation technique as the analytical tool was based on the interval nature of the data and its distribution across students' total scores, making this technique

appropriate for measuring the degree of linear relationship between item scores and total scores (Habibi et al., 2023).

The subjects of this study were seventh-grade students at MTsN 1 Pasaman, totaling 32 students. The object of this study was the daily test items that had been administered to these students. The data used in this study were secondary data, namely students' answer sheets and daily test scores that had been collected by the teacher.

Data collection was carried out using the documentation method, namely collecting the answer sheets of seventh-grade students at MTsN 1 Pasaman together with the answer key. Students' scores on each item were then tabulated for further analysis using the Pearson Product Moment correlation technique to determine the validity level of each item.

Validity analysis was carried out by calculating the correlation coefficient between the score of each item and the student's total score using the following Pearson Product Moment formula:

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Information:

- r_{xy} = represents the correlation coefficient between the item score and the total score
- n = Sum of students
- x^2= the sums of the squares of scores x
- y^2= the sums of the squares of scores y
- $\sum x$ = the score of the test item
- $\sum y$ = the students total score
- $\sum xy$ = the sum of the products of scores X and Y

The resulting $r_{calculated}$ value was then compared with the r_{table} value at a significance level of 5% with a degree of freedom (df) = n − 2 = 30. A test item was declared valid if $r_{calculated} > r_{table}$, whereas an item was declared invalid if $r_{calculated} \leq r_{table}$.

3. RESULT AND DISCUSSION

Based on the results of the data analysis conducted on the 5 seventh-grade daily test items at MTsN 1 Pasaman using the Pearson Product Moment correlation technique involving 32 students, the $r_{calculated}$ value for each item was obtained and then compared with the r_{table} value at a 5% significance level with df = 30, which is 0.349. The results of the item validity calculation are presented in Table 1 below.

Table 1. Validity Test Results of Seventh-Grade Daily Test Items

Item Number	r_{table}	$r_{calculated}$	Result
1	0,349	0,619	Valid
2	0,349	0,694	Valid
3	0,349	0,627	Valid
4	0,349	0,724	Valid
5	0,349	0,682	Valid

Based on Table 1, it can be seen that all of the items tested, namely items 1 through 5, have an $r_{calculated}$ value greater than the r_{table} value (0.349). This shows that all five items are declared statistically valid, because the validity requirement for an item is met when $r_{calculated}$ is greater than r_{table}.

The research results show that the five seventh-grade daily test items at MTsN 1 Pasaman that were analyzed have varying correlation coefficient values, ranging from 0.619 to 0.724. The highest $r_{calculated}$ value was obtained for item number 4, at 0.724, while the lowest value was found in item number 1, at 0.619. Although there is variation in the values, all items remain well above the r_{table} value of 0.349, meaning that each item has a strong and significant correlation with the students' total scores.

The high correlation values across all items indicate that each item is able to effectively distinguish between students with high ability and those with low ability in understanding the material being tested. Items with high correlations, such as item number 4 (0.724) and item number 2 (0.694), show that these items were consistently answered correctly by students with high total scores and incorrectly by students with low total scores. This is consistent with the principle that a valid item must be able to measure student ability accurately and consistently with the overall ability measured by the test instrument (Anggraini et al., 2022).

Thus, the seventh-grade daily test items at MTsN 1 Pasaman analyzed in this study can be said to have met the feasibility standard as a learning evaluation measurement tool, because all items were proven to be statistically valid. This indicates that the teacher who composed the items had considered the alignment between the items and the ability intended to be measured, so that the test results obtained by students can be trusted to reflect their actual ability. Although all items were declared valid, teachers are still advised to conduct item analysis periodically on subsequent evaluation instruments in order to maintain the consistency of item quality used.

The finding that all items were declared valid in this study differs from several similar studies that generally report variation in item quality, with some items declared invalid. For instance, Cahyaningrum et al. (2023), in an analysis of summative mathematics test items for seventh grade, found that not all items tested met the valid criteria, such that a number of items required revision before being reused. Similarly, research on algebra items at the junior high school level reported that of ten items tested, one item was declared invalid because its $r_{calculated}$ value fell below the r_{table} value (Qohar & Fauziyah, 2024). This difference can be interpreted as indicating that the quality of the seventh-grade daily test items at MTsN 1 Pasaman analyzed in this study is relatively better than several similar instruments in other studies, while also showing that validity test results are indeed highly contextual and cannot be generalized across schools or subjects (Welsandt et al., 2024).

On the other hand, the results of this study are also consistent with findings at the senior high school level, which reported that most mathematics test items met the validity and reliability criteria, making them suitable as a measure of student competence, although that study also found several items that needed improvement (Siregar et al., 2024). This consistency of findings reinforces the argument that instruments constructed based on a clear test blueprint and administered to a relatively homogeneous group of students tend to produce items with good validity levels. In addition, the high correlation values across all items in this study also indicate that the instrument used has good discriminating power, since one characteristic of a valid item is its ability to consistently distinguish between high- and low-ability students (Habibi et al., 2023).

Nevertheless, it should be noted that the validity established in this study only covers criterion validity based on the correlation between item scores and total scores, and has not yet addressed content validity through expert judgment or a comprehensive analysis of reliability, difficulty level, and discrimination index. A number of previous studies show that these four aspects need to be examined together to obtain a more complete picture of instrument quality, because an item that is statistically valid does not necessarily have an ideal difficulty level and discrimination index (Rohmah & Mauliddiyah, 2025). Therefore, the results of this study should be understood as initial evidence of instrument feasibility from the standpoint of criterion validity, which needs to be supplemented with further studies so that conclusions regarding the quality of the seventh-grade daily test items at MTsN 1 Pasaman become more comprehensive and methodologically accountable.

The findings of this study also have practical implications for teachers in compiling and managing item banks at schools. Given that validity test results are contextual and may change

with differences in student characteristics across cohorts, teachers are advised not to rely solely on items that were previously declared valid, but rather to re-analyze them each time they are reused with a different group of students (Lestari et al., 2025). Furthermore, since this study only involved 5 items from a single daily test, the results obtained cannot be generalized to all evaluation instruments used by the school as a whole. Teachers and schools need to build a habit of conducting systematic and well-documented item analysis, either manually using the Pearson Product Moment formula or with the assistance of software such as Anates or SPSS, so that the learning evaluation process becomes more efficient and accurate (Cahyaningrum et al., 2023; Habibi et al., 2023). In this way, a culture of evidence-based evaluation can continue to be developed as part of efforts to improve the quality of mathematics learning in schools.

4. CONCLUSION

This study aimed to determine whether the daily test items used for seventh-grade students MTsN 1 Pasaman were valid or not, through a validity test based on students' test result data. Based on the analysis conducted using the Pearson Product Moment correlation technique involving 32 students, it can be concluded that all of the tested daily test items were declared valid, as the $r_{calculated}$ value for each item was greater than the r_{table} value at a significance level of 5%. Thus, the daily test items used have met the requirements of a good and reliable evaluation instrument capable of accurately measuring seventh-grade students MTsN 1 Pasaman abilities in accordance with the intended learning objectives.

Based on these findings, it is recommended that teachers continue to conduct item analysis regularly and consistently for every evaluation instrument to be used, not only for daily tests but also for other forms of assessment such as midterm and final semester examinations, so that the quality of valid items can be maintained and improved over time. Furthermore, future researchers are encouraged to expand this study by increasing the number of student samples and test items analyzed, as well as incorporating reliability testing, difficulty level, and discriminating power analysis simultaneously, in order to obtain a more comprehensive picture of the quality of evaluation instruments used in schools.

REFERENCE

Amini, R. P. (2023). Validation and reliability analysis of the critical thinking ability instrument for elementary school students. *Edukatika*, *1*(1), 1–10.

Anggraini, F. D. P., Aprianti, A., Setyawati, V. A. V., & Hartanto, A. A. (2022). Learning statistics using SPSS software for validity and reliability testing. *Jurnal Basicedu*, *6*(4), 6491–6504. https://doi.org/10.31004/basicedu.v6i4.3206

Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Validity and reliability analysis of the odd-semester summative test items for Islamic Religious Education. *Edukatif: Jurnal Ilmu Pendidikan*, *6*(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931

Arikunto, S. (2006). *Research procedure: A practical approach*. Rineka Cipta.

Arikunto, S. (2013). *Research procedure: A practical approach* (Revised ed.). Rineka Cipta.

Ariyanto, T., Herwin, & Sujati, H. (2023). Construct validity and reliability testing of an integer arithmetic ability test instrument using CFA. *Jurnal Pendidikan Matematika (AJPM)*, *12*(3), 2977–2987. https://doi.org/10.24127/ajpm.v12i3.7482

Cahyaningrum, I. Y., Fuady, A., & Sunismi. (2023). Item analysis of the odd-semester summative mathematics test for seventh grade using Anates software. *Mathema Journal*, *5*(2), 67–81.

Dianova, F. R., & Anwar, N. (2024). Validity, reliability, difficulty level, and discrimination index analysis of the Arabic summative test items at an Islamic elementary school. *Jurnal Bahasa Daerah Indonesia*, *1*(3), 1–13. https://doi.org/10.47134/jbdi.v1i3.2863

Fiska, J. M., Hidayati, Y., Qomaria, N., & Hadi, W. P. (2021). Item analysis of science daily test items using Anates software based on classical test theory. *Natural Science Education Research*, *4*(1), 65–76. https://doi.org/10.21107/nser.v4i1.8133

Habibi, M., Marlina, R., & Setiawan, B. (2023). Factors influencing the validity and reliability of test items in learning evaluation. *ELIPS: Jurnal Pendidikan Matematika*, *5*(2), 110–121.

Hartanto, A. A., et al. (2023). Validity and reliability testing of a lecturer performance assessment instrument using SPSS. *SAINTEK (Jurnal Sains dan Teknologi)*, *4*(2), 21–24.

Himawan, R., & Nurgiyantoro, B. (2022). Item analysis of the odd-semester final assessment exercise for Indonesian language, eighth grade SMPN 1 Bambanglipuro Bantul, using the ITEMAN program. *Kembara: Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, *8*(1), 160–180.

Kasanova, A., & Sulistiyono, S. (2023). Item analysis of the odd-semester daily test using Google Forms at SMA Ipiems Surabaya. *Syntax Admiration*, *4*(12), 2200–2210.

Lestari, D. P., Wahyuni, S., & Cappelleri, J. (2025). Evaluation of the validity and reliability of research instruments in quantitative education studies. *Jurnal Evaluasi Pendidikan*, *6*(1), 1–12.

Pratiwi, D., & Marzal, J. (2024). Quality analysis of the even-semester final mathematics test items for eighth grade.

Qohar, M. A., & Fauziyah, F. (2024). Validity and reliability analysis of algebra test items for junior high school students. *Jurnal Tarbiyatuna*, *5*(2), 88–99.

Rohmah, A., & Mauliddiyah, E. (2025). Quality analysis of literacy ability test items in terms of validity, reliability, difficulty level, and discrimination index. *GAMMA-NC*, *5*(1), 116–122.

Siregar, T. M., Saragih, S., & Surya, E. (2024). Validity and reliability analysis of multiple-choice mathematics test items for twelfth grade, even semester, at SMAN 1 Adiankoting. *AR RUMMAN: Journal of Education and Learning Evaluation*, *1*(2), 720–730.

Sudjana, N., & Ibrahim, R. (2004). *Educational research and assessment*. Sinar Baru Algensindo.

Sugiyono. (2012). *Quantitative, qualitative, and R&D research methods*. Alfabeta.

Welsandt, N. C. J., Fortunati, F., Winther, E., & Abs, H. J. (2024). Constructing and validating authentic assessments: The case of a new technology-based assessment of economic literacy. *Empirical Research in Vocational Education and Training*, *16*(1), 1–27. https://doi.org/10.1186/s40461-024-00158-0

Yusuf, R. (2021). Item analysis of the final mathematics examination for junior high school. *ELIPS Journal*.