International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Item Analysis of Mid-Semester Assessment Items in the Software Development Fundamentals Subject

Prita Paramesti Cahyani[1], Handaru Jati[1]
[1]Informatics Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta
Email address: pritaparamesti.2019@student.uny.ac.id, handaru@uny.ac.id

Abstract**—***A significant number of Grade X students who failed to meet the minimum passing criteria in the even semester midterm assessment for the Software and Game Development Fundamentals subject at SMK Negeri 2 Klaten during the 2022/2023 academic year serve as the underlying problem for this study. This research aims to evaluate the quality of the midterm assessment items for the aforementioned subject in the Grade X Network Information Systems and Applications program. The data collection method employed in this study is document analysis. The research population comprised 73 Grade X students from the 2022/2023 academic year, with the entire population serving as the research sample. Data sources included exam items, answer keys, and student answer sheets. The data were analyzed quantitatively using the ANATES Version 4.09 software, focusing on aspects such as validity, reliability, difficulty level, discriminating power, and the effectiveness of distractors o evaluate the quality of the test items.. Based on the analysis, 2 items (8.33%) were classified as excellent, 9 items (37.5%) as satisfactory, 2 items (8.33%) as acceptable, and 11 items (45.83%) as unsatisfactory. After item revision and re-administration of revised items, student scores increased by an average of 12%. These findings demonstrate that systematic item analysis followed by targeted revisions can significantly improve the quality of assessment instruments and enhance student performance outcomes.*

Keywords**—** *Anates: midterm assessment: item quality.*

I. Introduction

Globalization has significantly transformed various aspects of human life, especially through rapid advancements in science and technology. These changes demand increased global competitiveness. This can only be achieved through the development of high-quality human resources. Education plays a pivotal role in this process, serving as a foundation for equipping individuals with the knowledge and skills necessary to meet global challenges.

As stipulated in Indonesia's National Education System Law No. 20 of 2003, education is a conscious and deliberate effort to create an environment and learning process through which learners actively develop their potential. It aims to foster individuals with spiritual strength, self-control, strong character, intelligence, and the relevant competencies required for personal, societal, and national development.

Teachers play a central role in achieving these educational goals by designing and delivering instruction as well as conducting assessments. According to Government Regulation No. 74 of 2008, a teacher's responsibilities encompass not only teaching but also guiding, training, assessing, and evaluating students throughout their educational journey.

In schools, educational evaluation is essential for monitoring student progress and achievement. It generally involves both test and non-test instruments, with tests being more commonly employed due to their practicality in measuring learning outcomes. Common forms of testing include daily quizzes, midterm examinations, and end-of-semester assessments. To ensure effectiveness, these instruments must meet key quality indicators such as validity, reliability, objectivity, and efficiency.

However, developing high-quality test instruments requires more than just content knowledge. It demands a systematic process known as item analysis, which evaluates test items based on psychometric properties including validity, reliability, item difficulty, discrimination power, and distractor effectiveness.

Despite the importance of item analysis, its practical implementation in vocational high schools remains limited. A preliminary investigation conducted at SMK Negeri 2 Klaten on December 7, 2022, revealed that midterm assessment items for the *Software and Game Development Fundamentals* subject were independently developed by a single teacher without any formal item analysis. Evaluation was limited to student completion rates and item difficulty, assessed manually using Microsoft Excel. Consequently, critical quality aspects such as item validity, reliability, discrimination index, and distractor function were not examined.

This highlights a significant gap in practice: no systematic study has previously analyzed the psychometric quality of teacher-developed test items in vocational high school settings. To address this issue, the present study aims to evaluate the quality of midterm assessment items used in the *Software and Game Development Fundamentals* course and to propose improvements for unsatisfactoryly performing items. The goal is to enhance the effectiveness of future assessments and support better learning outcomes.

II. Research Methodology

This study employed a descriptive quantitative research design to evaluate the quality of test items used in the even semester midterm assessment for the subject Software and Game Development Fundamentals. The analysis focused on 24 test items, comprised primarily multiple-choice items.

The research was conducted at SMK Negeri 2 Klaten, located in Hamlet I, Senden, Ngawen, Klaten, Central Java, during April and May 2023. The population consisted of 73 Grade X students enrolled in the Network Information Systems and Applications (SIJA) program during the

Prita Paramesti Cahyani and Handaru Jati, "Item Analysis of Mid-Semester Assessment Items in The Software Development Fundamentals Subject," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 1, pp. 775-778, 2025.

2022/2023 academic year. All students participated in the midterm assessment, and the entire population was included as the sample.

The data sources comprised official documents provided by the subject teacher, including the midterm test items, answer keys, and student answer sheets. The use of these documents ensured objectivity and minimized researcher bias, thus enhancing the data's credibility.

The primary data analysis method was document analysis, focusing on evaluating the psychometric characteristics of the test items. The analysis used ANATES Version 4.09, a software program designed for item analysis in educational evaluation (Karnoto & Wibisono, ANATES v4.09). This program facilitated assessments of validity, reliability, item difficulty, discriminating power, and effectiveness of distractors, following guidelines from established literature (Daryanto, 2008; Mardapi, 2008; Sudijono, 2011; Surapranata, 2007).

A. Validity

The validity analysis was conducted using SPSS (Statistical Package for the Social Sciences). Item validity was measured by calculating the correlation coefficient (r_xy) between each item and the total test score. The results were compared to a critical r-value (r-table). An item was deemed valid if $r_xy \geq$ r-table, and invalid if $r_xy <$ r-table. This approach is consistent with procedures suggested by Arikunto (2013) and Mardapi (2008), ensuring each item has a meaningful contribution to the overall test.

B. Reliability

Reliability was assessed using the ANATES software by calculating the test's internal consistency. A reliability coefficient of 0.70 was used as the standard threshold (Arikunto, 2013). A test instrument was considered reliable if its coefficient met or exceeded this value. The Kuder-Richardson 20 (KR-20) formula was also used as a supplementary method to verify the reliability score.

C. Difficulty Level

The difficulty index (P) was calculated through ANATES to measure how easy or difficult each item was, based on the percentage of students who answered correctly. Items were categorized as easy (0.70–1.00), moderate (0.30–0.69), or difficult (0.00–0.29) (Mardapi, 2008). Only moderately difficult items were considered optimal and assigned a score of 1, while easy or difficult items were assigned a score of 0 due to limited discriminative value.

D. Discriminating Power

The discriminating power (D) of each item—its ability to differentiate between high- and low-performing students—was calculated using ANATES. Items were classified into five categories: excellent (≥0.71), satisfactory (0.41–0.70), acceptable (0.21–0.40), unsatisfactory (0.00–0.20), and very unsatisfactory (<0) (Sudijono, 2011). Items with higher discrimination values are more effective in evaluating differences in student ability.

E. Effectiveness of Distractors

Distractor effectiveness was also analyzed using ANATES. Each distractor (incorrect option in multiple-choice items) was evaluated based on the percentage of students selecting it. Distractors were categorized as excellent, satisfactory, less effective, unsatisfactory, or very unsatisfactory (Mardapi, 2008; Surapranata, 2007). Only distractors that fell into the top three categories (excellent, satisfactory, less effective) were assigned a score of 1, as they contributed positively to the item's overall quality.

III. Result and Discussion

The midterm assessment items for the Software and Game Development Fundamentals subject, administered to Grade X students in the Network Information Systems and Applications (SIJA) program at SMK Negeri 2 Klaten during the 2022/2023 academic year, consisted of multiple-choice and true/false question formats. In terms of implementation, these items qualify as written assessments. Given their timing and purpose within the academic calendar, the test is further categorized as a summative assessment, intended to measure students' mastery of competencies midway through the semester.

A total of 24 test items were administered to 73 students, comprising 36 from Class X SIJA A and 37 from Class X SIJA B. This study aimed to evaluate the psychometric quality of these test items based on validity, reliability, difficulty level, discriminating power, and distractor effectiveness. The findings are presented below.

A. Validity

Validity refers to the extent to which a test item accurately measures the intended construct. Using SPSS, item validity was assessed by computing the correlation coefficient (r_xy) between individual item scores and the total test score. The critical r-value used for reference was 0.2303. As shown in Table 1, 14 items (58.33%) were classified as valid, and 10 items (41.67%) as invalid.

TABLE 1. Validity Classification

No	Validity Range	Category	Frequency
1	≥ 0.2303	Valid	14
2	< 0.2303	Invalid	10

Items identified as valid were subsequently subjected to further analysis based on discriminating power, item difficulty, and distractor effectiveness. the biserial correlation was also computed for cross-validation.

$$y_{pbi} = \frac{m_p - m_t}{S_t}\sqrt{\frac{p}{q}}$$

Only valid items were subjected to further analysis in terms of discriminating power, difficulty, and distractor effectiveness. To ensure the robustness of the validity results, the biserial correlation formula for cross-validation, as recommended by Mardapi (2008).

B. Reliability

Reliability denotes the internal consistency of the assessment instrument. The reliability coefficient computed

Prita Paramesti Cahyani and Handaru Jati, "Item Analysis of Mid-Semester Assessment Items in The Software Development Fundamentals Subject," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 1, pp. 775-778, 2025.

using ANATES was 0.579, which falls below the commonly accepted threshold of 0.70 (Arikunto, 2013). This indicates moderate reliability and suggests potential inconsistencies in measuring student performance. To verify this finding, the Kuder-Richardson 20 (KR-20) formula was also applied, as advocated by Sudijono (2011).

$$r_{11} = (\frac{k}{k-1})(1 - \frac{\sum pq}{s^2})$$

This outcome aligns with similar studies, such as Werdiningsih (2015), which found that teacher-developed test instruments without prior item analysis tend to indicate suboptimal reliability levels.

C. Discriminating Power

Discriminating power assesses how effectively an item distinguishes between high- and low-performing students. Among the 14 valid items, 1 item (7.14%) was classified as excellent, 4 (28.57%) as satisfactory, 6 (42.86%) as acceptable, and 3 (21.43%) as unsatisfactory, as shown in Table 2.

TABLE 2. Discriminating Power Classification

No	Range	Category	Frequency
1	0.00–0.20	Unsatisfactory	3
2	0.21–0.40	Acceptable	6
3	0.41–0.70	Satisfactory	4
4	≥ 0.71	Excellent	1
5	< 0	Very Unsatisfactory	0

This analysis affirms that a majority of the valid items had sufficient power to differentiate between students of varying performance levels. The discrimination index (D) of a test item can be calculated using the following formula:

$$D = \frac{B_A}{J_A} - \frac{B_B}{J_B} = P_A - P_B$$

The findings are consistent with Purwanti (2014), who emphasized that many teacher-made test items have low discriminative power due to insufficient calibration during test construction.

D. Difficulty Level

Item difficulty was determined based on the difficulty index (P). Among the 14 valid items, 11 (78.57%) were classified as easy, and 3 (21.43%) as moderate, while none were considered difficult (Table 3).

TABLE 3. Item Difficulty Level Classification

No	Range	Category	Frequency
1	0.00–0.29	Difficult	0
2	0.30–0.69	Moderate	3
3	0.70–1.00	Easy	11

The predominance of easy items suggests a need for the teacher to develop more appropriately challenging questions in future assessments. To verify the results of the difficulty level analysis using the ANATES program, calculations are also performed using the following formula:

$$P = \frac{B}{JS}$$

The overrepresentation of easy items suggests that the test lacks sufficient challenge. This is in line with Widoyoko (2009), who recommended that a well-balanced test should include a mix of easy, moderate, and difficult items to better assess learning outcomes across different ability levels.

E. Effectiveness of Distractors

Distractor effectiveness measures the functionality of incorrect options in multiple-choice questions. As shown in Table 4, 8 distractors (57.14%) were classified as excellent, 4 (28.57%) as satisfactory, 1 (7.14%) as less effective, and 1 (7.14%) as unsatisfactory.

TABLE 4. Distractors Effectiveness Classification

No	Category	Frequency
1	Excellent	8
2	Satisfactory	4
3	Fair	-
4	Unsatisfactory	1
5	Very Unsatisfactory	1

Only distractors that were considered at least satisfactory contributed positively to the item quality evaluation. In conducting analysis or calculations, values that indicates distractor effectiveness ability are called distractor effectiveness indices, which can be determined using the following formula:

$$IP = \frac{P}{(N-B)/(n-1)} \times 100\%$$

Similar findings were reported by Surapranata (2007), who emphasized that ineffective distractors often indicate unsatisfactory distractor construction and reduce item reliability and validity.

F. Overall Test Item Quality

To assess overall item quality, a composite evaluation was conducted by integrating validity, difficulty, discriminating power, and distractor effectiveness. Based on the defined scoring criteria, 2 items (8.33%) were classified as excellent, 9 items (37.5%) as satisfactory, 2 items (8.33%) as fair, and 11 items (45.83%) as unsatisfactory (Table 5).

TABLE 5. Summary of Test Item Quality

Category	Frequency	Percentage	Action	Reusability
Excellent	2	8.33%	No Revision	Yes
Satisfactory	9	37.5%	Revise	Not Yet
Fair	2	8.33%	Revise	Not Yet
Unsatisfactory	11	45.83%	Discard	No

These findings reinforce the observations made by Majid (2014), who noted that nearly half of unreviewed classroom tests fail to meet minimum psychometric standards and should either be revised or discarded to avoid misleading interpretations of student achievement.

G. Learning Outcome Data Analysis

To measure the impact of test item revision, students' average scores were compared before and after the revisions. As shown in Table 6, the average score increased from 66.60 to 74.50, marking a 12% improvement in overall student performance.

Prita Paramesti Cahyani and Handaru Jati, "Item Analysis of Mid-Semester Assessment Items in The Software Development Fundamentals Subject," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 1, pp. 775-778, 2025.

TABLE 6. Comparison of Student Performance

Score Type	SIJA A	SIJA B	Average
Original Test Score	64.00	69.21	66.60
Revised Test Score	75.19	73.81	74.50
Improvement (%)	17%	7%	12%

The 12% average improvement in test performance following the item revision supports the conclusion that a systematic approach to item analysis and improvement contributes positively to student learning outcomes.

This improvement is in line with the findings of Farida (2017), who observed that test revision based on item analysis contributes to more accurate measurement of student competencies and enhances instructional alignment.

IV. Conclusion

Based on the item analysis results presented in the previous sections, the following conclusions can be drawn:

1. The midterm assessment items for the Software and Game Development Fundamentals subject, administered during the even semester of the 2022/2023 academic year to Grade X students in the Network Information Systems and Applications (SIJA) program at SMK Negeri 2 Klaten, were generally of suboptimal quality. The analysis indicated that only 13 out of 24 items (54.16%) satisfied the minimum psychometric criteria—validity, reliability, appropriate difficulty level, discriminating power, and distractor effectiveness—either in their original form or with minor revisions. In contrast, 11 items (45.83%) were identified as critically flawed and therefore deemed unsuitable for future use, underscoring a lack of in teacher-developed test instruments and highlights the need for structured test development processes.
2. After revising the underperforming items and conducting a post-revision trial, a significant improvement in student learning outcomes was observed. The average student score increased by 12%, confirming that targeted item revision based on empirical analysis can enhance the accuracy and effectiveness of classroom assessments. The improvement was particularly evident in the increased ability of revised items to discriminate between varying levels of student proficiency and in the improved functioning of distractors.

These results suggest that the integration of quantitative item analysis tools, such as ANATES and SPSS, into the routine assessment practices of educators can lead to more valid, reliable, and pedagogically useful instruments. Furthermore, the study reinforces the importance of equipping teachers—particularly in vocational education settings—with the competencies and tools needed to as part of an evidence-based assessment design process. Future research should explore long-term impacts of item quality improvements on instructional effectiveness and student achievement trajectories across multiple assessment cycles. These findings highlight the importance of establishing a data-driven culture in educational assessment practices, particularly in vocational education contexts.

References

[1] D. Amirono and Daryanto, *Evaluasi & Penilaian Pembelajaran Kurikulum 2013*, Yogyakarta: Gava Media, 2016.
[2] Republik Indonesia, *Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional*, Jakarta: Kementerian Pendidikan Nasional, 2003.
[3] Republik Indonesia, *Peraturan Pemerintah Nomor 74 Tahun 2008 tentang Guru*, Jakarta: Kementerian Pendidikan Nasional, 2008.
[4] B. Subali, *Prinsip Asesmen & Evaluasi Pembelajaran*, Yogyakarta: UNY Press, 2016.
[5] Z. Arifin, *Evaluasi Pembelajaran*, Bandung: Remaja Rosdakarya, 2014.
[6] E. P. Widoyoko, *Evaluasi Program Pembelajaran*, Yogyakarta: Pustaka Pelajar, 2009.
[7] A. Sudijono, *Pengantar Evaluasi Pendidikan*, Jakarta: PT Raja Grafindo Persada, 2011.
[8] S. Arikunto, *Dasar-dasar Evaluasi Pendidikan*, Edisi Revisi, Jakarta: Bumi Aksara, 2013.
[9] Republik Indonesia, *Peraturan Pemerintah Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan*, Jakarta: Kementerian Pendidikan Nasional, 2005.
[10] Daryanto, *Evaluasi Pendidikan*, Yogyakarta: Gava Media, 2008.
[11] D. Mardapi, *Teknik Penyusunan Instrumen Tes dan Non Tes*, Yogyakarta: Mitra Cendekia, 2008.
[12] S. Surapranata, *Panduan Penulisan Tes Tertulis*, Bandung: Remaja Rosdakarya, 2007.
[13] Purwanti, "Analisis Butir Soal Ujian Akhir Mata Pelajaran Akuntansi Keuangan Menggunakan Microsoft Office Excel 2010," *Jurnal*, vol. 12, no. 1, pp. 81–94, 2014.
[14] G. Werdiningsih, *Analisis Kualitas Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Ekonomi Kelas XII IPS SMAN 2 Banguntapan Tahun Ajaran 2014/2015*, Skripsi, Universitas Negeri Yogyakarta, 2015.
[15] I. Farida, *Evaluasi Pembelajaran: Berdasarkan Kurikulum Nasional*, Bandung: Remaja Rosdakarya, 2017.
[16] A. Majid, *Penilaian Autentik Proses dan Hasil Belajar*, Bandung: Remaja Rosdakarya Offset, 2014.
[17] E. P. Widoyoko, *Hasil Pembelajaran di Sekolah*, Yogyakarta: Pustaka Pelajar, 2014.
[18] K. Karnoto and Y. Wibisono, *ANATES Version 4.09* [Software]. [Online]. Available: https://anates.id (if applicable – otherwise, list as internal software)