Jurnal Penelitian Politeknik Penerbangan Surabaya
Vol 11, No 2, Agustus, 2026
p-ISSN: 1978-6832
e-ISSN: 2622-5948

Retrospective Classical Item Audit of a Honda Automatic Motorcycle Tune-Up Test in Vocational Education

Moch. Chaqqul Yaqin[1], Rufi'i[2], Sabariah[3]
[1,2,3]Universitas PGRI Adi Buana Surabaya
e-mail: *[1]mochchaqqul yaqin@gmail.com.com, [2]rufi,i@unipasby.ac.id, [3]sabariah@unipasby.ac.id*
corresponding: mochchaqqul yaqin@gmail.com

DOI : https://doi.org/10.46491/jp.v20i2.2548

Abstract
This retrospective document-based audit evaluated a 25-item multiple-choice test on Honda automatic motorcycle tune-up knowledge using archived summary outputs from one administration to 35 Grade XI vocational students. The available records reported a mean score of 11.71 (SD = 5.32), Cronbach's alpha of 0.86, and an odd-even correlation of 0.75. Twenty-three items were classified as moderately difficult, item 19 as difficult, and item 23 as easy. Discrimination values ranged from 0.00 to 100.00 on the source scale; item 1 had the lowest value and item 11 the highest. Items 10 and 11 had the highest reported item-total correlations (0.820), whereas item 1 had a near-zero negative correlation (-0.039). Because the raw item-response matrix, calculation settings, and option-level frequencies were unavailable, the statistics could not be independently recalculated and distractor evidence was not used for definitive retention decisions. The findings indicate that relatively high total-score internal consistency can coexist with heterogeneous item functioning. Item 1 should receive priority review, while all item decisions should be confirmed through raw-data reanalysis, expert content review, revised-item pilot testing, and complementary practical performance assessment.
Keywords: Classical test theory, item analysis, multiple-choice assessment, vocational education, vocational testing.

Article History:
Received: May 8, 2026 ; Revised: July 24, 2026 ; Accepted: August 6, 2026

INTRODUCTION

Assessment in vocational education must provide defensible evidence about what students know and can do. Written tests can support decisions about feedback, remediation, and progression, but the meaning of their scores depends on the quality of the items and on the intended use of the results. Contemporary testing standards therefore require validity evidence to be accumulated from test content, response processes, internal structure, relationships with other variables, and the consequences of score use rather than inferred from a single coefficient (American Educational Research Association et al., 2014).

Multiple-choice tests remain attractive in vocational classrooms because they are efficient, objectively scored, and suitable for broad content coverage. Nevertheless, a written test primarily samples knowledge and reasoning; it cannot by itself establish practical competence in performing a motorcycle tune-up. In competency-based vocational education, written evidence should therefore complement direct performance assessment, observation, and other authentic evidence (Munawar, 2021; Wahyuni, 2021; Yusop et al., 2022).

Classical test theory offers accessible procedures for examining classroom tests through total-score internal consistency, item difficulty, discrimination, item-total relationships, and distractor functioning. These indices answer different questions. Internal consistency concerns

the coherence of total scores in a particular administration, whereas item statistics identify how individual questions behave within the tested sample. A relatively high alpha does not establish unidimensionality, validity, temporal stability, or uniformly strong item performance (Taber, 2018).

Evidence from vocational assessment studies shows that items that appear appropriate during test construction may perform unevenly after administration. Studies in technical drawing, chemistry, entrepreneurship, and other vocational domains have reported imbalanced difficulty, weak discrimination, and item-fit problems that require revision and retesting (Hakim et al., 2024; Salirawati & Nugraheni, 2020; Widiyanti et al., 2023). Distractor quality is also consequential because plausible incorrect options help differentiate partial understanding from well-developed knowledge (Gierl et al., 2017).

The present case concerns an existing 25-item test on Honda automatic motorcycle tune-up knowledge. The test package was complete at the documentary level, but the materials available for this study consisted of the test document and archived summary outputs rather than the original student-by-item response matrix. Accordingly, the study does not claim to reproduce or comprehensively validate the test. It conducts a transparent retrospective audit of the evidence that can be supported by the archived records.

The study addresses five questions: (1) what total-score and descriptive information was reported for the administration; (2) how were items distributed by reported difficulty; (3) what variation was reported in discrimination; (4) which item-total relationships can be identified from the archived output; and (5) which items can be prioritized for further review without making unsupported retention or deletion decisions?

The contribution is practical rather than methodologically novel. By distinguishing verifiable findings from unavailable evidence, the study demonstrates how a vocational school can audit an existing classroom test, identify priority concerns, and specify the additional data required before the instrument is reused for consequential decisions.

LITERATURE REVIEW

1. Instrument Evaluation in Educational Assessment

Instrument evaluation concerns the adequacy of evidence supporting the interpretation and use of test scores. Under classical test theory, an observed score is understood as a combination of a true-score component and measurement error. Classroom item analysis does not by itself establish comprehensive validity, but it can identify response patterns that warrant item review and further evidence collection (American Educational Research Association et al., 2014; Díaz et al., 2023).

For vocational assessment, the interpretive boundary is especially important. A written test can provide evidence about technical knowledge, procedural sequencing, diagnosis, safety principles, or tool selection, but practical competence requires direct evidence of performance. Instrument evaluation should therefore relate each item to the test blueprint and should position written scores as one component of a broader competency-assessment system (Ramadani et al., 2024; Soeprijanto et al., 2022; Yusop et al., 2022).

a. Test Reliability and Validity

Internal consistency describes the degree to which item scores covary in a specific administration. Cronbach's alpha is frequently reported, and for dichotomously scored items it is interpreted as an internal-consistency estimate rather than proof that the instrument is valid, stable over time, or unidimensional. Alpha is influenced by test length, item covariance, score variability, and sample characteristics; consequently, it should be read together with item-level evidence (Taber, 2018).

Validity concerns the justification of score interpretations and uses. Evidence about content representation, response processes, internal structure, relationships with external criteria, and consequences is cumulative. In the present audit, only limited evidence about internal item functioning was available. No claims are therefore made about comprehensive construct validity, criterion validity, fairness, or practical competence (American Educational Research Association et al., 2014; Navarro-González et al., 2024).

b. Item Difficulty and Discrimination

Item difficulty for a dichotomously scored item is ordinarily represented by the proportion of examinees answering correctly. Difficulty is sample dependent and should be interpreted in relation to the intended purpose and content coverage. Very easy or very difficult items may still be educationally important, but they generally provide less differentiation within a particular group than items near the middle of the score range (Díaz et al., 2023; Widiyanti et al., 2023).

Discrimination indicates whether an item distinguishes higher- and lower-scoring examinees. It may be estimated through an upper-lower group index or through correlation-based statistics. Positive values are generally desirable, whereas values near zero indicate little differentiation. Negative values require diagnostic review of the key, wording, content alignment, scoring, and response data; they do not, by themselves, prove construct misalignment (Díaz et al., 2023; Koskey et al., 2020).

c. Item-Total Correlation and Distractor Quality

An item-total correlation indicates the relationship between performance on one item and the total test score. For item evaluation, a corrected relationship that excludes the focal item from the total is preferable because it avoids part-whole inflation. Interpretation remains sample dependent, and a low coefficient may result from limited score variation, content heterogeneity, a weak key or option set, or sampling fluctuation.

Distractors should be plausible to examinees who have incomplete understanding while remaining clearly incorrect to knowledgeable examinees. Their evaluation normally requires option-level response frequencies and, ideally, selection patterns by score group. Without these data, statements about distractor quality must remain descriptive and should not be used to make definitive item-retention decisions (Gierl et al., 2017).

2. Position of the Present Study

The literature supports an iterative assessment cycle: construct a blueprint, review content and language, pilot the items, analyze responses, revise problematic components, and retest the revised form. Vocational teachers can use classical indices as an accessible first-stage quality-control tool, provided that the calculations and their limitations are reported transparently (Setyaedhi et al., 2024; Widiyanti et al., 2023).

This study is positioned as a retrospective audit within that cycle. It examines an existing classroom test using only the documentary and summary evidence that was archived. The analysis identifies provisional strengths and priority concerns, while explicitly separating those observations from decisions that require raw responses, expert review, or practical-performance evidence.

METHODOLOGY

1. Research Design

This study used a retrospective, cross-sectional, document-based audit to examine an existing classroom achievement test. The object was a 25-item multiple-choice test on Honda automatic motorcycle tune-up knowledge administered once to 35 Grade XI vocational students.

The design was descriptive and did not involve an intervention, comparison group, or causal inference.

The audit differed from a de novo psychometric reanalysis. The original student-by-item response matrix was not included in the available source package. Consequently, the study extracted and critically interpreted archived summary statistics but did not independently reproduce the calculations. This limitation is incorporated into the title, abstract, analysis, and conclusions.

The unit of analysis was the assessment instrument and its reported item-level indicators rather than individual students. The audit was guided by the five questions stated in the Introduction and used classical test theory as an interpretive framework, not as evidence of comprehensive test validation.

2. Data Sources and Research Object

Two data sources were examined. The first was the test package, which contained the instructional topic, learning objectives, blueprint, 25 multiple-choice items, answer key, and item explanations. The second was an archived analysis output reporting score summaries, total-score reliability, an odd-even correlation, item difficulty categories, discrimination values, selected item-total correlations, and qualitative distractor classifications.

The archived output represented 35 test takers. Demographic variables, sampling information, testing time, administration conditions, missing-response codes, and subgroup identifiers were not retained in the supplied materials. The findings therefore describe only this administration and are not generalized to other cohorts or institutions.

The raw 35 x 25 response matrix, complete option-level frequencies, item-level calculation output, and software settings were unavailable. The audit consequently distinguishes between values explicitly present in the source and procedures that could not be verified. No absent item statistics were estimated or reconstructed.

3. Data Collection Procedure

Data collection consisted of document retrieval, source matching, and structured extraction. The test form and archived output were checked for agreement in the number of items, instructional topic, and administration summary. Values were then transcribed into a verification worksheet and organized by test-level and item-level evidence.

Extraction covered the number of examinees and items, minimum and maximum scores, mean, standard deviation, Cronbach's alpha, the reported odd-even correlation, difficulty categories, discrimination range, selected item-total correlations, and the availability of distractor information. Values not tied to identifiable item numbers were retained as aggregate observations only.

Tables 1 and 2 describe the evidence available for the audit. They are source-inventory tables rather than newly calculated results. This distinction prevents the reconstructed presentation from being mistaken for an independent reanalysis of the original student responses.

Table 1. Scope and Evidence Available for the Retrospective Audit

Component	Value
Study type	Retrospective document-based item audit
Test administration	One archived classroom administration
Number of test takers	35
Number of items	25
Observed score range	3-22
Mean (SD)	11.71 (5.32)
Available test-level indices	Cronbach's alpha = 0.86; reported odd-even correlation = 0.75
Raw item-response matrix	Not available

Table 1 summarizes the design and the evidentiary limits of the archived administration. The absence of the raw response matrix is central to interpretation because it prevents independent verification of scoring, item-level coefficients, missing responses, and reliability if an item is deleted.

Table 2. Archived item-analysis indicators available for review

Indicator	Archived summary
Difficulty classification	23 moderate; item 19 difficult; item 23 easy
Discrimination scale	0.00-100.00, as reported by the source
Lowest discrimination	Item 1 = 0.00
Highest discrimination	Item 11 = 100.00
Higher item-total correlations	Items 10 and 11 = 0.820; item 6 = 0.671; item 17 = 0.650; item 23 = 0.593
Low item-total correlations	0.167; 0.165; 0.156; item numbers not retained
Negative item-total correlation	Item 1 = -0.039
Distractor evidence	Qualitative labels only; option frequencies unavailable

Table 2 records the item-level summaries recoverable from the archived output. Values associated with identifiable items are reported by item number; unlinked low and moderate correlations are not assigned to items because doing so would require unsupported reconstruction.

4. Data Analysis Techniques

The study retained the terminology used in the source output. The reported internal-consistency coefficient was labelled Cronbach's alpha, and the second coefficient was labelled an odd-even correlation. The source did not state whether the latter was the raw correlation between test halves or a Spearman-Brown corrected estimate. Accordingly, it is reported without reclassification. Alpha is interpreted only as internal consistency for this administration (Taber, 2018).

Difficulty was interpreted from the source categories of easy, moderate, and difficult. The underlying proportion-correct values and classification cutoffs were not retained. The audit therefore reports the original categories without treating them as independently verified thresholds.

Discrimination was reported on a 0.00-100.00 scale. The regular increments in the archived output were preserved, but the formula, upper-lower grouping rule, and group sizes were not documented. Item-total relationships were also reported as archived; the source did not specify whether they were corrected for part-whole overlap. These values are therefore used to identify provisional priorities rather than definitive item quality.

Distractor evidence was limited to qualitative labels. Because per-option frequencies, keyed-option counts, and selection patterns by score group were unavailable, no operational definition of a non-functioning distractor could be applied. Distractor information was summarized as a data limitation and was not used to retain, discard, or rank individual items (Gierl et al., 2017).

No differential item functioning, invariance, factor analysis, item-response modelling, confidence intervals, or bootstrap estimates were calculated. Such analyses require raw responses, subgroup variables, and larger samples. The present audit is intentionally restricted to transparent interpretation of the archived evidence.

5. Validity, Ethics, and Methodological Limitations

The evidence supports a limited evaluation of reported internal consistency and item functioning, not comprehensive validation. Content representation was not independently judged by experts, construct structure was not tested, and no relationship with a practical motorcycle tune-up assessment or external criterion was available. Claims about practical competence,

certification readiness, fairness, and consequential validity are therefore outside the scope of the study (American Educational Research Association et al., 2014).

Only anonymized aggregate classroom-assessment outputs were available to the audit, and no student names or individual profiles were included in the analysis or reporting. The audit involved no intervention, participant contact, or access to identifiable personal data. Ethical governance for any reuse of the underlying educational records remains the responsibility of the originating institution.

The study has five principal limitations: one administration with 35 examinees; absence of the raw response matrix; unverified calculation settings and thresholds; incomplete item-number mapping for several coefficients; and lack of expert, qualitative, subgroup, and practical-performance evidence. These limitations make all item decisions provisional and prevent estimation of sampling uncertainty.

Within these constraints, the audit provides a reproducible account of what the archived materials do and do not support. Future work should recover the raw responses, verify the answer key, report complete item and option statistics, obtain expert content review, pilot revised items in an independent cohort, and examine relationships with direct practical-performance assessment.

RESULTS

1. General Profile of Student Scores

The archived administration included 35 examinees and 25 items. Scores ranged from 3 to 22, with a mean of 11.71 and a standard deviation of 5.32. These statistics establish that scores varied within the observed group, but they do not by themselves demonstrate mastery, validity, or representativeness.

Table 3 presents the descriptive values exactly as retained in the source output. Because individual response records were unavailable, the distribution could not be checked for missing values, scoring errors, skewness, outliers, or confidence intervals.

Table 3. Descriptive statistics of the archived test administration

Statistic	Value
Number of test takers	35
Number of items	25
Minimum score	3
Maximum score	22
Mean score	11.71
Standard deviation	5.32
Score range	19

The score range provides a descriptive context for the reported item indices. All subsequent findings remain conditional on the accuracy of the archived scoring and analysis output.

2. Whole-Test Reliability

The source reported Cronbach's alpha of 0.86 and an odd-even correlation of 0.75 (Table 4). Alpha indicates relatively high internal consistency for the total score in this administration. It does not establish that the test is unidimensional, stable over time, valid for practical competence, or free of problematic items.

The odd-even value is reported descriptively because the archived output did not indicate whether it was the raw correlation between halves or a corrected split-half reliability estimate. Reliability if each item was deleted could not be examined without the raw response matrix.

Table 4. Reported total-score internal-consistency indicators

Indicator	Reported value
Cronbach's alpha	0.86
Odd-even correlation	0.75; correction status not documented

The available evidence therefore supports a cautious statement of relatively high total-score internal consistency, not a conclusion that the instrument is globally valid or ready for consequential use without further review.

3. Item Difficulty Distribution

The source classified 23 items as moderately difficult, item 19 as difficult, and item 23 as easy (Table 5). The distribution indicates that the archived classification was concentrated in the middle category. Because the underlying proportion-correct values and category thresholds were unavailable, the categories cannot be independently verified. The difficult status of item 19 and easy status of item 23 should therefore be read as descriptions of this administration rather than permanent item characteristics.

Difficulty should also be interpreted with content importance and intended use. An easy item may assess essential safety knowledge, whereas a difficult item may reflect either demanding content or an item-design problem.

Table 5. Reported distribution of item difficulty categories

Difficulty category	Number of items	Item numbers
Easy	1	23
Moderate	23	1-18, 20-22, 24-25
Difficult	1	19
Total	25	All items

The difficulty profile alone does not justify retaining or removing any item. It must be considered together with discrimination, item-total relationships, content review, and practical relevance.

4. Item Discrimination

Reported discrimination values ranged from 0.00 to 100.00 on the source scale. Item 1 had the lowest value (0.00), while item 11 had the highest value (100.00). Intermediate values occurred in increments from 11.11 to 88.89 (Table 6).

This range indicates heterogeneous differentiation within the observed group. However, the formula and upper-lower grouping procedure were not retained; consequently, values should not be interpreted as precise or stable properties beyond the administration.

Item 1 warrants diagnostic review because its discrimination value was zero. Item 11 showed the most favorable reported discrimination, but its performance should be confirmed using raw data and a new cohort before it is treated as a calibrated item.

Table 6. Reported item discrimination range

Statistic	Reported value
Scale	0.00-100.00
Lowest discrimination	0.00
Item with lowest discrimination	Item 1
Highest discrimination	100.00
Item with highest discrimination	Item 11
Intermediate values and limitation	11.11-88.89; formula and grouping rule unavailable

The principal result is not that particular thresholds were definitively passed or failed, but that a test with relatively high total-score alpha contained items with markedly different reported discrimination.

5. Item-Total Correlation

The highest reported item-total correlations were 0.820 for items 10 and 11. Other relatively high values were 0.671 for item 6, 0.650 for item 17, and 0.593 for item 23. These items showed the strongest reported alignment with total-score variation in the archived output (Table 7).

Moderate values of 0.550, 0.510, 0.501, 0.484, 0.466, 0.460, and 0.421 and low values of 0.167, 0.165, and 0.156 were also retained. Their item numbers were not available, so the audit does not assign them to specific questions or propose item-specific revisions.

Item 1 had a correlation of -0.039. Because this value is only slightly below zero and the sample was small, it is best interpreted as a warning of possible item malfunction rather than proof of construct misalignment. The answer key, stem, options, content alignment, scoring, and raw responses should be checked before any decision to revise or remove the item.

Table 7. Reported item-total correlation summary

Correlation category	Reported values	Identified items
Relatively higher	0.820; 0.671; 0.650; 0.593	Items 10 and 11; 6; 17; 23
Moderate	0.550; 0.510; 0.501; 0.484; 0.466; 0.460; 0.421	Item numbers not retained
Low	0.167; 0.165; 0.156	Item numbers not retained
Negative	-0.039	Item 1

The archived output did not specify whether the item-total correlations were corrected. The values therefore support provisional prioritization only and should be recalculated from the original response matrix.

6. Distractor Quality

The source characterized distractor performance as varying from strong to very weak, but it did not retain complete option-level frequencies or identify every affected item. Table 8 therefore reports the availability of evidence rather than assigning distractor-quality labels to individual questions.

Without the number and percentage selecting each option, the keyed option, missing-response counts, and selection patterns by total-score group, it is not possible to determine which distractors were non-functioning or whether they influenced difficulty and discrimination.

Accordingly, distractor information was excluded from definitive item decisions. A future reanalysis should report A-D option frequencies for every item and apply a stated operational criterion for non-functioning distractors.

Table 8. Availability of Distractor Evidence

Evidence element	Availability
General source classification	Strong to very weak
Per-option frequencies	Unavailable
Item-specific distractor identities	Incomplete/unavailable
Operational non-functioning threshold	Not documented
Upper/lower option-selection pattern	Unavailable
Use in item decisions	Not used for definitive retention decisions

This treatment corrects the earlier tendency to infer item-level distractor effects from incomplete summary labels. The archived evidence signals a need for further analysis but does not quantify distractor performance.

7. Relatively Strong and Problematic Items

Table 9 presents a provisional review priority rather than a final retain-revise-discard classification. Item 1 is the highest priority for diagnostic review because it combined zero reported discrimination with a near-zero negative item-total correlation. The first checks should concern the answer key, stem clarity, option plausibility, content alignment, and scoring accuracy.

Items 6, 10, 11, 17, and 23 showed relatively favorable reported item-total relationships, and item 11 also had the highest reported discrimination. These findings justify provisional retention for further review, not automatic acceptance. No item-specific decision is made for the remaining questions because complete statistics and content review evidence were unavailable.

Table 9. Provisional Item-Review Priorities

Priority	Items	Basis and required action
Priority diagnostic review	1	D = 0.00 and item-total r = -0.039; verify key, stem, options, scoring, and content alignment
Provisionally favorable	6, 10, 11, 17, 23	Relatively high reported item-total correlations; confirm with raw-data reanalysis and retesting
Insufficient item-specific evidence	Remaining items	No final decision without complete statistics, option frequencies, and expert content review

The table deliberately avoids unsupported identification of items associated with unlinked low correlations or qualitative distractor labels. Final decisions require complete raw-data output and expert review of instructional importance.

Overall, the archived evidence shows relatively high total-score internal consistency alongside heterogeneous item functioning. The strongest defensible outcome is a prioritized plan for reanalysis and revision rather than a claim that the instrument has been comprehensively validated.

DISCUSSION

1. Interpreting the Overall Pattern of Findings

The audit identifies a useful but carefully bounded pattern. The archived total score had a reported alpha of 0.86, while item-level indicators varied widely. This supports the central conclusion that relatively high internal consistency can coexist with heterogeneous item functioning. It does not establish that the test is valid, unidimensional, stable across administrations, or appropriate as the sole basis for vocational competency decisions.

This interpretation is consistent with warnings against treating alpha as a universal quality certificate. Internal consistency summarizes covariance among item scores in a particular sample and can remain high when some items require revision, particularly in a test containing multiple related content elements (Taber, 2018). Item review therefore remains necessary even when the total-score coefficient appears favorable.

The practical implication is that teachers should examine both test-level and item-level evidence. In the present case, the defensible response is not to discard the whole instrument or to accept it unchanged, but to recover the raw data, verify the analysis, review priority items, and pilot the revised form.

2. Structural Imbalance at the Item Level

The concentration of 23 items in the reported moderate category might initially suggest appropriate targeting. However, the reported discrimination range and item-total relationships show that items within a similar difficulty category did not function uniformly. Difficulty and

discrimination therefore provide complementary rather than interchangeable information (Díaz et al., 2023; Widiyanti et al., 2023).

The interpretation must remain provisional because the underlying proportion-correct values, discrimination formula, and upper-lower grouping rule were unavailable. With only 35 examinees, small changes in response patterns can materially alter item indices. Replication with a larger independent cohort is needed before any item is described as consistently easy, difficult, weak, or highly discriminating.

Content importance also matters. An item should not be removed solely because of an unfavorable statistic if it samples an essential safety or procedural outcome. Statistical evidence should trigger diagnostic review alongside the blueprint and expert judgment, followed by revision and retesting.

3. The Problematic Item and Its Interpretive Significance

Item 1 is the clearest review priority because its reported discrimination was 0.00 and its item-total correlation was -0.039. The correlation is only slightly below zero and should not be described as proof of construct misalignment or item misfit. In a small sample, it is a warning that the item did not align with total-score variation in this administration.

Several explanations remain plausible: an incorrect or ambiguous key, unclear wording, implausible or multiple defensible options, scoring or data-entry error, content heterogeneity, restricted variation, or sampling fluctuation. The recommended sequence is to verify the key and scoring, compare the item with the blueprint and instructional materials, inspect option frequencies, obtain expert review, revise if necessary, and retest.

This diagnostic approach preserves the intended learning objective while correcting the unsupported causal language in the earlier manuscript. Deletion should be considered only after the source of the response pattern has been investigated and the content consequences of removal have been assessed.

4. Implications of Distractor Quality and Item Revision

The archived output suggests that distractor functioning was not uniform, but the absence of option-level data prevents item-specific conclusions. Effective distractor analysis requires the keyed response and the frequency and percentage selecting every option, preferably separated by score group (Gierl et al., 2017).

Consequently, the present study cannot demonstrate that weak distractors caused particular difficulty or discrimination values. It can only identify incomplete distractor evidence as a quality-assurance gap. Future reporting should define a non-functioning distractor criterion and document which alternatives meet it.

Revision should be followed by a new pilot administration. The reanalysis should include corrected point-biserial relationships, alpha or KR-20 if each item is deleted, complete option frequencies, explicit discrimination procedures, and uncertainty estimates where feasible. Item decisions should then integrate statistical results, content relevance, clarity, and curriculum coverage.

5. Implications for Vocational Assessment Practice

For vocational education, the study reinforces the distinction between knowledge testing and direct competency assessment. The present instrument can contribute evidence about Honda automatic motorcycle tune-up knowledge, but it cannot independently demonstrate that a student can safely inspect, diagnose, adjust, and document a motorcycle tune-up procedure. Practical performance should be assessed through observation, rubrics, or structured work tasks (Ramadani et al., 2024; Wahyuni, 2021; Yusop et al., 2022).

Scores from the current form may be useful for low-stakes formative interpretation, but consequential decisions should be made cautiously until the item calculations have been verified

and the problematic items reviewed. The audit also illustrates the importance of archiving raw response data, software settings, test blueprints, and revision histories as part of assessment quality assurance.

Teacher assessment literacy is therefore both statistical and curricular. Educators need sufficient skill to read item indices, but they must also evaluate whether an item represents an important learning outcome and whether a written response is an appropriate form of evidence for that outcome (Setyaedhi et al., 2024).

6. Contribution of the Study

The study's contribution is a transparent applied audit of an existing vocational classroom test. It does not introduce a new psychometric model. Its value lies in demonstrating how a favorable total-score coefficient can mask item-level variation and how incomplete archives constrain the decisions that can legitimately be made.

By explicitly separating observed values, provisional interpretations, and unavailable evidence, the revised study offers a practical reporting model for schools conducting retrospective assessment quality checks. The model directs attention from broad labels such as good or bad toward specific actions: verify, review, revise, retest, and triangulate with practical performance.

7. Limitations and Transition to the Conclusion

The principal limitations are the single cohort of 35 examinees, the absence of raw responses, the inability to verify formulas and thresholds, incomplete mapping of several coefficients to item numbers, and the lack of expert content review, subgroup information, qualitative response evidence, and an external practical-performance criterion. These constraints prevent confidence intervals, item-deletion reliability analysis, differential item functioning, factor analysis, and independent reproduction of the results.

Future research should recover or prospectively collect the full item-response matrix; document administration and scoring procedures; report complete item and option statistics; review content and language with automotive and measurement experts; pilot revised items in a larger independent sample; and examine how written-test scores relate to observed motorcycle tune-up performance. Until then, the findings should remain specific to the archived administration and should guide review rather than certify the instrument.

CONCLUSION

This retrospective audit found that the archived 25-item Honda automatic motorcycle tune-up test had relatively high reported total-score internal consistency (Cronbach's alpha = 0.86) while its item-level indicators were heterogeneous. Twenty-three items were reported as moderately difficult, but discrimination ranged from 0.00 to 100.00 on the source scale, and item-total correlations ranged from 0.820 to -0.039 among the values retained. Item 1 should receive priority diagnostic review because it combined zero discrimination with a near-zero negative item-total relationship. This pattern does not prove construct misalignment and should be investigated through verification of the key, wording, options, scoring, and content alignment. The absence of the raw response matrix, calculation settings, complete item-number mapping, and option-level frequencies prevents independent reanalysis and definitive retain-revise-discard decisions. Accordingly, the instrument should be treated as requiring further quality assurance rather than as comprehensively validated. Before consequential reuse, the authors should recover the raw data, conduct complete item and distractor analysis, obtain expert content review, pilot revised items in an independent cohort, and combine the written test with direct practical-performance assessment.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Díaz, N. V. M., Yoon, S. Y., Trytten, D., & Meier, R. (2023). Development and validation of the Engineering Computational Thinking Diagnostic for undergraduate students. *IEEE Access, 11*, 133099-133114. https://doi.org/10.1109/ACCESS.2023.3335931

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Hakim, S. R., Ekohariadi, Suparji, S., Harmanto, & Wiwin, E. (2024). Item test analyze of entrepreneurship readiness instrument for vocational school students in East Java Indonesia. *Studies in Philosophy of Science and Education, 5*(1), 41-47. https://doi.org/10.46627/sipose.v5i1.330

Koskey, K. L. K., Makki, N., Ahmed, W., Garafolo, N. G., & Visco, D. P. (2020). Modification and validation of the mixed-format Engineering Concept Assessment for middle school students using many-facet Rasch measurement. *School Science and Mathematics, 120*(5), 309-321. https://doi.org/10.1111/ssm.12405

Munawar, W. (2021). Assessment of competency based on dicotomic scoring to diagnose students' skills in vocational learning in vocational schools. *Journal of Vocational Education Studies, 4*(1), 89. https://doi.org/10.12928/joves.v4i1.3823

Navarro-González, M. C., Padilla, J., & Benítez, I. (2024). Analyzing measurement invariance for studying the gender gap in educational testing. *European Journal of Psychological Assessment, 40*(5), 412-426. https://doi.org/10.1027/1015-5759/a000820

Ramadani, A. H., Ekohariadi, Anifah, L., Nugroho, Y. S., & Safitri, R. (2024). Development of performance assessment instruments for measuring drawing skills in vocational students competence in mechanical engineering expertise. *International Journal of Recent Educational Research, 5*(2), 268-279. https://doi.org/10.46245/ijorer.v5i2.551

Salirawati, D., & Nugraheni, A. R. E. (2020). An analysis of validity of test items in chemistry final semester examination for vocational students majoring in computer and networking engineering in Bantul Regency. *Asian Journal of Assessment in Teaching and Learning, 10*(1), 32-40. https://doi.org/10.37134/ajatel.vol10.1.4.2020

Setyaedhi, H. S., Rusijono, R., & Khotimah, K. (2024). Training and mentoring in the development of test instruments for measuring learning outcomes of Muhammadiyah school. *International Journal of Community Service Learning, 7*(3), 272-281. https://doi.org/10.23887/ijcsl.v7i3.63063

Soeprijanto, S., Prohantoro, R., Pratama, R. A., & Ariefin, D. A. (2022). Needs analysis for developing project-based learning outcomes assessment models in electricity topic at the Center of Excellence Vocational High School. *Jurnal Penelitian & Pengembangan Pendidikan Fisika, 8*(2), 325-334. https://doi.org/10.21009/1.08214

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Wahyuni, D. S. (2021). Developing competency-based assessment at vocational high school (VHS) in Bali. *Jurnal Pendidikan Vokasi, 11*(1). https://doi.org/10.21831/jpv.v11i1.37383

Widiyanti, W., Dianawati, R., & Futachi, A. S. (2023). A cognitive assessment instrument for HOTS development in technical drawing subjects utilizing Quizziz learning platform. *Jurnal Ilmu Pendidikan, 29*(1), 1-6. https://doi.org/10.17977/um048v29i1p1-6

Yusop, S. R. M., Rasul, M. S., Yasin, R. M., Hashim, H. U., & Jalaludin, N. A. (2022). An assessment approaches and learning outcomes in technical and vocational education: A systematic review using PRISMA. *Sustainability, 14*(9), 5225. https://doi.org/10.3390/su14095225