JURNAL CINTA NUSANTARA
Volume 04 Nomor 04 2026
E-ISSN: 3025-4469
https://jurnalbundaratu.org/journal/index.php/cintanusantarajournal
Lisensi Creative Commons Atribusi 4.0 Internasional

EXPLORING MULTIDIMENSIONAL TEST ITEM DEVELOPMENT AND SCORING THROUGH THE LENS OF BLOOM'S TAXONOMY

Muslem Daud[1]*, Muhammad Duskri[2]
[1] *Lecturer at Universitas Serambi Mekkah, Banda Aceh, Aceh, Indonesia, Email: muslem.daud@serambimekkah.ac.id*
[2] *Lecturer at Universitas Islam Negeri Ar-Raniry, Banda Aceh, Indonesia, Email: m.duskri@ar-raniry.ac.id*
**Corresponding author: muslem.daud@serambimekkah.ac.id*

Abstract
The development of educational tests requires more than the preparation of questions and the assignment of numerical scores. High-quality assessment instruments should represent the intended learning objectives, address different levels of cognitive complexity, and employ scoring procedures that accurately reflect students' knowledge and abilities. This conceptual article explores a multidimensional approach to test item development and scoring through the lens of Bloom's Taxonomy, with particular emphasis on the revised taxonomy developed by Anderson and Krathwohl (2001). The article examines the relationship between cognitive processes, knowledge dimensions, item formats, cognitive demand, and scoring criteria. It proposes an integrated conceptual framework in which test development proceeds through alignment of learning objectives, specification of cognitive and knowledge dimensions, construction of assessment items, development of scoring rubrics, and evaluation of assessment quality. The framework distinguishes the cognitive level intended by an item from its empirical difficulty and emphasizes that the two constructs should not be treated as interchangeable. It also highlights the importance of validity, reliability, fairness, and transparent scoring in interpreting assessment results. The article argues that multidimensional assessment can provide a more comprehensive representation of student learning than approaches that rely exclusively on factual recall or undifferentiated total scores. Nevertheless, the proposed framework requires empirical validation across subject areas, educational levels, and assessment formats. The article contributes a conceptual foundation for developing educational tests that are theoretically informed, instructionally aligned, and methodologically defensible.

Keywords: Bloom's Taxonomy, multidimensional assessment, test item development

1. INTRODUCTION

Educational assessment is a fundamental component of teaching and learning because it provides information about what students know, understand, and can do (Daud et al., 2024). Assessment results inform instructional decisions, curriculum development, student feedback, educational accountability, and judgments about learning achievement (Kuo et al., 2015). However, the quality of these decisions depends substantially on the quality of the assessment instruments used to produce the evidence. A test that measures only a narrow range of learning objectives may provide an incomplete representation of student competence, even when its scores appear precise and consistent (Daud et al., 2026; Daud et al., 2015).

Traditional educational tests frequently emphasize the recall of facts, definitions, and previously presented information. Although recall is an essential component of learning, educational objectives often extend beyond remembering information. Students may also be expected to explain concepts, apply principles, analyse relationships, evaluate evidence, and develop original solutions (Daud & Sariakin, 2025; Daud, 2017). An assessment system that S

concentrates predominantly on factual recall may therefore fail to capture important differences in students' cognitive performance. Bloom's Taxonomy provides an influential conceptual framework for classifying educational objectives according to cognitive complexity. The original taxonomy organized cognitive objectives into a hierarchy that included knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al., 1956). This framework helped educators articulate learning objectives and consider whether assessment tasks were consistent with the intellectual demands of instruction.

Previous related study was conducted by Anderson and Krathwohl (2001) subsequently revised the taxonomy by distinguishing a cognitive process dimension from a knowledge dimension. The revised framework identifies six cognitive processes—remember, understand, apply, analyse, evaluate, and create—and four knowledge categories: factual, conceptual, procedural, and metacognitive knowledge. This distinction offers a more flexible basis for examining what students are expected to know and the kinds of thinking through which they demonstrate their learning.

Other study done by Reeves and Marbach-Ad (2016) emphasized the importance of establishing validity evidence to support the interpretations and uses of assessment results. Their discussion highlights that a test represents a sample of student performance from a broader domain of knowledge or ability. Accordingly, the selection of test items should reflect the content and competencies that the assessment intends to measure. This principle is particularly relevant to multidimensional test development because it requires assessment designers to consider the representation of multiple learning domains rather than concentrating exclusively on a limited number of topics or cognitive processes. Similarly, Raymond and Grande (2019) emphasized the importance of test blueprinting as a systematic procedure for determining the content and cognitive dimensions of an assessment instrument. A test blueprint provides a structured relationship between learning objectives, content domains, and the distribution of assessment items. Through this approach, educators can ensure that the number and types of questions correspond to the relative importance of the instructional content and the competencies being assessed (Khadijah et al., 2026).

Despite the usefulness of taxonomic frameworks, translating educational objectives into high-quality test items remains a complex task. Assigning a cognitive label to a question does not automatically establish that the item measures the intended ability. A question classified as analysis may require only superficial recognition, while a seemingly straightforward question may demand sophisticated reasoning when students must interpret unfamiliar evidence or justify a decision. The cognitive demand of an assessment item therefore depends on the knowledge required, the task presented, the information available to students, and the reasoning necessary to formulate an adequate response.

A related challenge concerns scoring. Different assessment formats generate different kinds of evidence and require appropriate scoring procedures. Selected-response items may be scored objectively according to predetermined answer keys, whereas constructed-response tasks may require explicit criteria, performance levels, and trained evaluators. When scoring procedures fail to reflect the intended learning objectives, the resulting scores may not adequately represent student competence. These issues suggest that test item development and scoring should be treated as interconnected components of a coherent assessment design rather than as separate technical activities. A multidimensional approach can help educators establish explicit relationships among learning objectives, cognitive processes, knowledge categories, assessment formats, and scoring criteria.

The purpose of this article is to explore a conceptual framework for multidimensional test item development and scoring based on Bloom's Taxonomy. Specifically, the article seeks to (a) examine the theoretical foundations of the taxonomy and its relevance to educational assessment; (b) explain how cognitive processes and knowledge dimensions can inform test

item construction; (c) develop principles for aligning item formats and scoring procedures with learning objectives; and (d) propose an integrated framework for improving the validity, reliability, fairness, and interpretability of educational assessment.

The central argument is that Bloom's Taxonomy is most useful when applied as a systematic design framework rather than as a simple labelling system. Its contribution to assessment quality depends on how effectively educators translate learning objectives into observable tasks, develop appropriate scoring criteria, and examine whether the resulting evidence supports the intended interpretations of student performance.

2. METHOD

Method

This article adopts a conceptual and narrative literature review approach. Rather than reporting the results of a new empirical investigation, it synthesizes established theoretical perspectives on educational objectives, cognitive processes, test item development, and assessment scoring. The principal conceptual foundation is Bloom's original taxonomy and its subsequent revision, supplemented by literature on assessment design, higher-order thinking, educational measurement, and test development.

The analysis proceeds through four stages and based on Anderson & Krathwohl studies (2001). First, it examines the conceptual relationship between educational objectives, cognitive processes, and knowledge dimensions. Second, it identifies the implications of these dimensions for constructing assessment items across different formats. Third, it examines scoring procedures in relation to the characteristics of the evidence produced by each item type. Fourth, it integrates these relationships into a proposed multidimensional framework for educational assessment.

The synthesis is organized around five dimensions: alignment with learning objectives, cognitive demand, knowledge representation, scoring appropriateness, and assessment quality. These dimensions are used to formulate conceptual propositions and practical principles for test development. Because the article is conceptual, the proposed framework should be understood as a theoretically informed model rather than an empirically validated instrument. Its effectiveness must be examined through subsequent research involving expert review, pilot testing, item analysis, reliability assessment, and investigations of validity evidence across relevant educational contexts.

Conceptual Definition of the Multidimensional Item Development Model

The Multidimensional Item Development Model is a systematic approach to educational test construction in which test items are developed to represent all relevant learning domains or content areas covered during instruction. Rather than concentrating on selected topics or relying primarily on a single cognitive level, the model seeks to ensure that the assessment provides comprehensive coverage of the material taught. Each learning domain is represented by an appropriate set of test items, with the cognitive demands of those items determined in accordance with the intended learning objectives and Bloom's Taxonomy.

The model rests on two fundamental dimensions: content-domain coverage and cognitive-process representation. Content-domain coverage ensures that the assessment represents the breadth of the instructional material, while cognitive-process representation ensures that students are assessed at appropriate levels of thinking, ranging from remembering and understanding to applying, analysing, evaluating, and creating. These dimensions are integrated through an assessment blueprint that specifies the distribution of items across learning domains and cognitive categories (Duskri et. al., 2014).

The model further incorporates a scoring framework that establishes how student responses are evaluated within and across learning domains. This allows assessment results to

provide information not only about overall achievement but also, where the test design and measurement evidence support such interpretations, about students' performance across particular content domains and cognitive processes. Accordingly, multidimensionality in this model refers primarily to the comprehensive representation of instructional content domains and their alignment with cognitive objectives.

3. RESULT AND DISCUSSION

RESULT

Theoretical Foundations

1) Bloom's Original Taxonomy and Its Educational Significance

Bloom's Taxonomy emerged from an effort to provide educators with a systematic classification of educational objectives. Bloom et al. (1956) organized cognitive objectives into six categories: knowledge, comprehension, application, analysis, synthesis, and evaluation. The taxonomy was intended to help educators describe learning objectives more precisely and develop assessment procedures that reflected the intended outcomes of instruction.

The original framework contributed to educational assessment by emphasizing that learning objectives vary in their cognitive demands. Remembering information is different from interpreting it, applying it in a new situation, or evaluating the quality of an argument. Recognizing these differences allows teachers to design assessments that represent a broader range of intellectual activities (Daud et al., 2022; Daud & Ansari; 2026).

The taxonomy also provides a basis for examining alignment among curriculum objectives, instructional activities, and assessment tasks. If a course emphasizes analytical reasoning but its tests primarily require factual recall, the assessment may not adequately represent the intended curriculum. Similarly, an assessment designed to measure basic conceptual understanding should not depend excessively on reading complexity or specialized knowledge that was not part of the learning objectives.

Nevertheless, the taxonomy should not be interpreted as an inflexible sequence through which every learning task must progress. The cognitive demands of educational activities depend on the discipline, the knowledge involved, the learning context, and the experience of the learner. Remembering essential information can be necessary for complex reasoning, while evaluation and creation may require substantial factual and conceptual knowledge.

Accordingly, Bloom's original taxonomy remains valuable as a framework for describing educational objectives, but its categories should be applied in relation to the actual demands of the assessment task rather than solely through the verbs used in a question.

2) The Revised Bloom's Taxonomy

Anderson and Krathwohl (2001) revised the original taxonomy to provide a more differentiated framework for educational planning and assessment. The revised taxonomy distinguishes two dimensions: the knowledge students are expected to acquire and the cognitive processes through which they work with that knowledge. The cognitive process dimension includes remembering, understanding, applying, analyzing, evaluating, and creating. The knowledge dimension includes factual, conceptual, procedural, and metacognitive knowledge. Their intersection creates a framework that enables educators to specify learning objectives more precisely.

For example, remembering factual knowledge may involve identifying a term or recalling a definition. Understanding conceptual knowledge may require explaining the relationship between two theories. Applying procedural knowledge may involve using an established method to solve a problem. Evaluating metacognitive knowledge may require students to examine the appropriateness of their own learning strategies.

These examples demonstrate why cognitive processes and knowledge types should be considered jointly. Two test items may address the same cognitive process while requiring

different kinds of knowledge. Conversely, two items may address the same knowledge category while requiring different levels or types of cognitive processing.

The revised taxonomy therefore provides a useful foundation for constructing a test specification or assessment blueprint. Such a blueprint can identify which learning objectives should be assessed, the knowledge categories involved, the intended cognitive processes, and the proportion of assessment evidence allocated to each area. However, a taxonomy matrix should not be treated as a guarantee of assessment quality. An item's actual cognitive demand must be examined in relation to its wording, context, response requirements, and expected reasoning. The classification should guide item development and review rather than replace substantive analysis of the task.

3) Multidimensionality in Educational Assessment

The term *multidimensional* can refer to several related but distinct concepts in assessment. First, it may describe the deliberate representation of multiple cognitive processes and knowledge categories in a test. Second, it may refer to the inclusion of multiple item formats, such as selected-response, short-answer, extended-response, and performance tasks. Third, in psychometric measurement, multidimensionality may refer to the presence of more than one underlying latent trait contributing to observed responses (Kuo et al., 2015).

These meanings are related, but they should not be conflated. A test can contain items classified across several Bloom categories while still measuring a predominantly general academic ability. Conversely, a test may exhibit multiple statistical dimensions even when its items appear similar in format or content. From a test-development perspective, multidimensionality begins with a clear definition of the intended learning domain. Educators must determine whether the assessment is designed to measure several distinguishable competencies, a broad achievement construct, or a combination of both.

If the purpose is to assess a broad domain of achievement, an overall score may be useful, provided that the interpretation is supported by appropriate evidence. If the purpose is to distinguish performance across cognitive processes or knowledge domains, separate subscale scores may be more informative. Such scores, however, should not be reported as reliable measures of distinct abilities unless the assessment design and empirical evidence support that interpretation. Downing and Haladyna (2006) emphasize the importance of systematic test development and evaluation. Applied to multidimensional assessment, this principle implies that the intended structure of the test must be specified before item construction and examined again during validation.

4) Cognitive Complexity and Empirical Item Difficulty

An important distinction in educational measurement concerns cognitive complexity and item difficulty. Cognitive complexity describes the type and depth of thinking intended or required by an item. Item difficulty describes how challenging the item is for a particular group of examinees under specified testing conditions.

These constructs are related but not equivalent. A question requiring analysis may be easy for students who have practiced similar tasks extensively. A question requiring recall may be difficult if it concerns unfamiliar terminology or material that was inadequately taught.

For a dichotomously scored item, empirical difficulty is often represented by the proportion of examinees answering correctly. Under this convention, a higher proportion correct indicates an easier item, whereas a lower proportion correct indicates a more difficult item. For polytomously scored items, difficulty can be examined through score distributions, item response models, or other appropriate statistical approaches. Consequently, test developers should not infer empirical difficulty solely from a Bloom category. Instead, they should establish the intended cognitive demand through expert review and examine actual item

performance through pilot testing and appropriate item analysis. This distinction is essential because a well-designed assessment should represent the intended learning objectives while providing sufficient variation in item difficulty to support meaningful interpretations of student performance.

Conceptual Implications for Test Item Development

1) Alignment Between Learning Objectives and Test Items

The first principle of multidimensional item development is alignment. Every item should be linked to an explicit learning objective, and the intended cognitive process should be consistent with the knowledge and reasoning required by the task.

Alignment can be established through an assessment blueprint that specifies the content domain, learning objectives, cognitive processes, knowledge categories, item formats, scoring procedures, and relative weighting of items.

For example, if students are expected to analyze the causes of an economic problem, an item that merely asks them to recall the definition of that problem may provide useful background evidence but cannot, by itself, demonstrate analytical competence. A more appropriate task would require students to examine evidence, identify relationships among factors, and explain how those relationships contribute to the problem.

Alignment also involves reviewing the language and contextual demands of each item. Unnecessarily complex wording, unfamiliar cultural references, or irrelevant reading demands may prevent an item from measuring the intended learning objective. Item developers should therefore distinguish the knowledge and reasoning central to the task from incidental demands that may disadvantage particular students.

2) Constructing Items Across Cognitive Processes

Items can be developed to represent different cognitive processes, but their classification should depend on the response students must produce rather than on isolated command verbs. Remembering items may ask students to identify, recognize, or recall relevant information. Understanding items may require students to explain, summarize, classify, or interpret concepts. Applying items may ask students to use a principle or procedure in a specified situation. Analyzing items can require students to distinguish relevant from irrelevant information, examine relationships, or identify the structure of an argument. Evaluating items may ask students to judge the quality of evidence or defend a decision using explicit criteria. Creating items can require students to develop a plan, formulate a hypothesis, construct an argument, or propose a solution to a problem.

These categories provide useful design guidance, but they do not function as automatic labels. The verb *explain*, for example, may indicate understanding in one task and more complex reasoning in another. Likewise, a question using the verb *evaluate* may require only recognition of a previously taught judgment if the answer is explicitly supplied in the prompt.

Item writers should therefore articulate the intended reasoning process, identify the evidence needed to demonstrate it, and check whether the task genuinely elicits that evidence.

3) Integrating Knowledge Dimensions

Cognitive processes should also be considered in relation to factual, conceptual, procedural, and metacognitive knowledge. Factual knowledge includes terminology, details, and specific information required within a subject area. Conceptual knowledge includes principles, classifications, theories, and relationships among concepts. Procedural knowledge concerns methods, techniques, and criteria for determining when procedures are appropriate. Metacognitive knowledge concerns awareness of cognition and the strategies used to monitor and regulate learning.

A multidimensional assessment blueprint can identify which combinations of knowledge and cognitive processing are essential to the learning objectives. Such an approach helps prevent a test from becoming overly concentrated on isolated facts when the curriculum also emphasizes conceptual understanding, procedural competence, or reflection on learning.

Not every assessment must cover all four knowledge categories or all six cognitive processes. Coverage should reflect the purpose of the test, the instructional objectives, the developmental level of the students, and the time available for assessment. The aim is appropriate representation, not mechanical equality among categories.

Domains	Remember	Understand	Apply	Analyze	Evaluate	Create
Domain 1	✓	✓	✓	✓	✓	✓
Domain 2	✓	✓	✓	✓	✓	✓
Domain 3	✓	✓	✓	✓	✓	✓

DISCUSSION

Conceptual Framework and Proposed Model

1) Conceptualizing Multidimensional Test Item Development

The proposed Multidimensional Test Item Development Model is based on the principle that an educational assessment should represent the full range of relevant content taught during instruction. In classroom practice, teachers sometimes construct tests by selecting questions from familiar topics, emphasizing particular chapters, or using a convenient collection of items. Although such practices may produce a workable examination, they do not necessarily ensure that the resulting scores adequately represent students' learning across the instructional curriculum (Daud & Mariati, 2021).

In this article, *multidimensionality* primarily refers to the systematic representation of multiple learning domains or content areas covered during instruction. The model seeks to ensure that every relevant domain receives appropriate representation in the assessment. Bloom's Taxonomy is then used as a complementary framework for classifying the cognitive processes that students are expected to demonstrate within those domains.

This distinction is important. Content-domain coverage and cognitive complexity are related but different aspects of test design. Anderson and Krathwohl (2001) explain the revised taxonomy through the relationship between a knowledge dimension and a cognitive-process dimension. The proposed model incorporates this distinction while placing particular emphasis on comprehensive coverage of the content taught.

The model therefore has two principal dimensions:

1) Content-domain dimension: The range of topics, units, competencies, or learning domains included in instruction and represented in the test.
2) Cognitive-process dimension: The types of thinking required to answer the questions, classified according to the revised Bloom's Taxonomy: Remember, Understand, Apply, Analyze, Evaluate, and Create.

The intersection of these dimensions provides a practical foundation for constructing test specifications. For example, a teacher assessing economics may identify cooperatives, cooperative principles, organizational structures, capital, and member participation as relevant content areas. The teacher can then determine which cognitive processes are appropriate for assessing each area. The model does not assume that every content domain must be assessed at all six cognitive levels. Instead, item allocation should reflect instructional objectives, the

depth of instruction, the importance of the content, and the learning outcomes expected of students.

2. The Proposed Conceptual Model

The model comprises six interconnected stages.

Stage 1: Identification of learning objectives and instructional content. Teachers identify the learning outcomes and content actually taught. Assessment coverage should be based on documented instructional objectives and curriculum requirements rather than on arbitrary item selection.

Stage 2: Mapping of content domains. The instructional content is organized into identifiable domains. Each domain is assigned a relative weight based on considerations such as instructional time, curricular importance, competency requirements, and the breadth of learning expected.

Stage 3: Alignment with Bloom's Taxonomy. Teachers classify the intended cognitive demands of the assessment items. The classification should be based on the thinking required to answer a question, not merely on the action verb used in its wording.

Stage 4: Multidimensional item construction. Questions are developed for each relevant content domain and aligned with appropriate cognitive objectives. The number and type of questions are determined by the assessment blueprint.

Stage 5: Scoring and domain-level analysis. Before administering the test, teachers establish scoring rules, item weights, partial-credit criteria where appropriate, and procedures for calculating total and domain-level scores.

Stage 6: Interpretation and reporting of learning outcomes. Scores are interpreted in relation to the intended learning outcomes. In addition to a total score, teachers may report students' performance in individual content domains to identify strengths and learning needs.

This sequence is consistent with established assessment guidance emphasizing explicit test specifications, alignment between intended content and test items, and documented scoring procedures (American Educational Research Association et al., 2014).

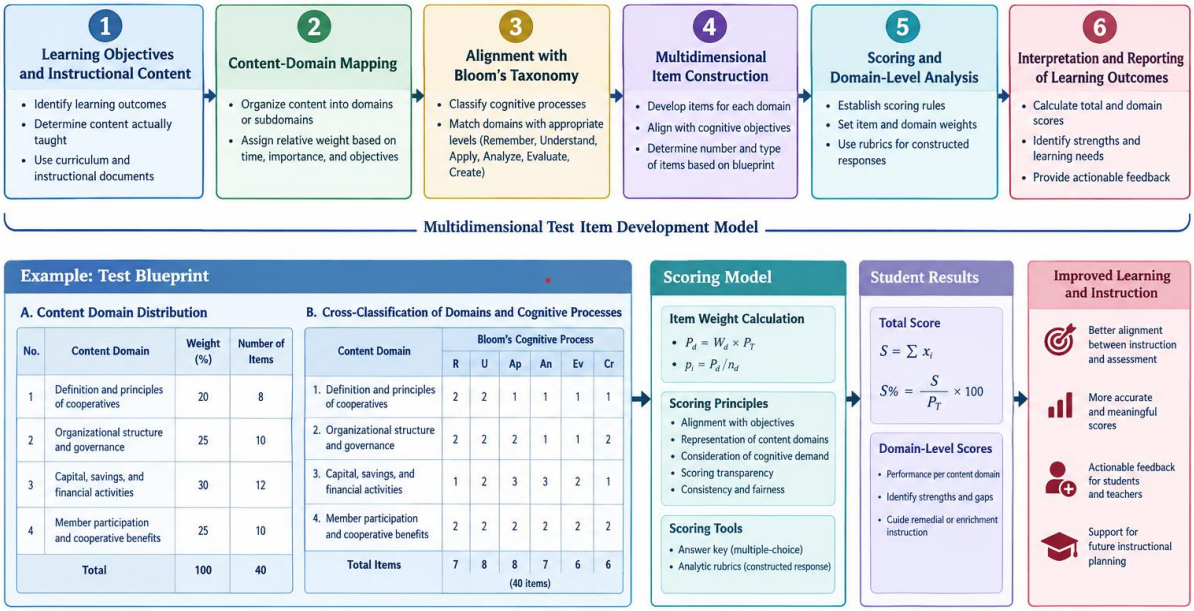


Figure 1. Conceptual Framework and Proposed Model of Multidimensional Test Item Development and Scoring

3) The Assessment Blueprint

The assessment blueprint is the operational foundation of the proposed model. It translates instructional objectives into a planned distribution of questions and scores.

Consider a hypothetical economics examination on cooperatives consisting of 40 items. The following table illustrates how a teacher might distribute the items across content domains.

Content domain	Weight	Number of items
Definition and principles of cooperatives	20%	8
Organizational structure and governance	25%	10
Capital, savings, and financial activities	30%	12
Member participation and cooperative benefits	25%	10
Total	**100%**	**40**

The percentages are illustrative, not empirically established standards. Actual weights should be justified by curriculum requirements, instructional objectives, and the relative importance of the content.

The blueprint can be extended by cross-classifying each domain with the cognitive processes in Bloom's revised taxonomy. For example, questions about cooperative capital may require students to recall definitions, explain financial concepts, calculate a simple financial amount, analyze a financial situation, or evaluate a proposed financial decision. The appropriate cognitive level depends on the intended learning objective. The advantage of this approach is that teachers can identify content that has been underrepresented before the examination is administered. It also makes the rationale for item allocation and scoring more transparent.

Proposed Scoring Model for Multidimensional Assessment

1) Why Equal Scoring Can Be Problematic

A common classroom practice is to assign the same number of points to every test item. This procedure is not inherently incorrect. Equal weighting can be appropriate when items have comparable purposes, scoring demands, and importance within the assessment design (Duskri et al., 2017).

The problem arises when equal weighting is applied automatically, without considering the assessment blueprint, the learning objectives, or the amount and type of evidence required to demonstrate mastery. A short factual question and an extended problem-solving task may demand different amounts of work and provide different kinds of evidence. Treating them as equivalent without a defensible rationale may result in scores that do not adequately represent the intended learning outcomes.

However, a second mistake must also be avoided: automatically assigning more points to a question simply because it appears difficult. Difficulty and cognitive complexity are not interchangeable. A recall question can be difficult because its content is unfamiliar or poorly learned, while an analytical question can be relatively easy when the problem is familiar or straightforward. Brookhart (2010) emphasizes the importance of distinguishing the level of difficulty from the level of thinking when developing assessments of higher-order thinking.

Consequently, the proposed model recommends criterion-based, blueprint-aligned scoring rather than automatic equal weighting or arbitrary difficulty-based weighting.

2) Principles for Assigning Item Weights

The model proposes five principles for determining item weights. First, alignment with learning objectives. Each item should measure an identified learning outcome. Its weight should reflect the importance of the outcome and the role of the item in providing evidence of achievement. Second, representation of content domains. The total points assigned to each domain should correspond to its justified weight in the assessment blueprint. This prevents a

large number of questions from one topic from unintentionally dominating the total score. Third, cognitive demand. Items assessing application, analysis, evaluation, or creation may require more elaborate scoring criteria than straightforward recall questions. For constructed-response items, additional points may be warranted when students must demonstrate multiple essential elements of reasoning. However, cognitive level alone should not determine the weight. Fourth, scoring transparency.

Students and teachers should be able to understand how points are allocated. For complex responses, analytic rubrics can identify the criteria for correctness, reasoning, use of evidence, and completeness. Fifth, consistency and fairness. Scoring procedures should be applied consistently across students. For open-ended responses, explicit rubrics and, where feasible, moderation or double-scoring procedures can help reduce unnecessary differences among scorers. These principles align with broader standards for test development, scoring, interpretation, and fairness (American Educational Research Association et al., 2014).

3) Calculating Item Weights

The calculation should provide a transparent method for distributing points. They do not require all items across the entire test to receive equal weights. Nor do they require every item within a domain to receive equal weights if the assessment design justifies differentiated weights. For example, suppose a teacher allocates 25% of a 100-point examination to a domain on cooperative governance. The domain receives 25 points. If it contains five items of equal weight, each item receives five points. If the domain instead contains four multiple-choice questions and one extended-response question, the teacher may distribute the same 25 points differently, provided that the allocation is justified by the objectives and scoring criteria.

The important principle is that item weights should be established during test planning, not improvised after students have completed the examination.

4) An Illustrative Example of Differentiated Scoring

Consider a hypothetical 10-item test with a maximum score of 100 points. The following distribution illustrates how scoring may reflect differences in assessment demands while preserving the intended content coverage.

Item type	Number of items	Points per item	Total points
Basic conceptual understanding	4	5	20
Application of concepts	3	8	24
Analysis of a case or problem	2	13	26
Evaluation supported by reasoning	1	30	30
Total	**10**	—	**100**

This is a hypothetical example, not a recommended universal weighting scheme. The 30-point evaluation item would be defensible only if it assessed an important learning outcome and used a rubric capable of supporting that allocation. A teacher should not assign it 30 points merely because it is considered the hardest question. The example also illustrates a limitation: a single extended-response item may have a disproportionate influence on the total score. If that item is ambiguous, poorly aligned, or scored inconsistently, the overall result may be distorted. Teachers should therefore consider the number of items, the breadth of evidence, and the reliability of scoring when determining weights.

5) Scoring Constructed-Response Items

For open-ended questions, the proposed model recommends analytic scoring rubrics when different aspects of performance need to be evaluated separately. A rubric might assign points to conceptual accuracy, application of relevant principles, quality of reasoning, use of supporting evidence, and completeness of the response.

For instance, an item requiring students to evaluate a cooperative's financial decision could be scored using the following illustrative rubric:

Criterion	Maximum points
Accuracy of economic concepts	4
Application to the given case	4
Quality of reasoning and analysis	6
Evidence supporting the conclusion	4
Total	**18**

The criteria and their weights should be specified before scoring begins. This approach makes the basis of the score more visible than a single holistic judgment without an explicit explanation. Brookhart (2010) similarly emphasizes the need to determine how students' responses will be interpreted and scored when designing assessments of higher-order thinking.

Importantly, a rubric should reward demonstrated learning rather than the length of a response. A concise answer that provides accurate analysis and sufficient evidence should not automatically receive a lower score than a longer answer containing irrelevant information.

6) Calculating and Interpreting the Total Score

Although the arithmetic is straightforward, the educational interpretation requires more care. A score of 76 for example should be interpreted in relation to the assessment's content coverage, scoring rules, intended learning outcomes, and applicable achievement criteria. A numerically precise score does not automatically guarantee a valid interpretation of what the student knows or can do.

The *Standards for Educational and Psychological Testing* emphasize the need for evidence and theory to support the interpretation of test scores for their intended uses (American Educational Research Association et al., 2014). Thus, the proposed model treats scoring as part of the overall assessment design rather than as a mathematical procedure added after item construction.

7) Reporting Scores by Content Domain

A distinctive feature of the proposed model is the recommendation to report both overall achievement and, when appropriate, performance by content domain. The total score provides a summary of performance across the examination, whereas domain-level scores can indicate where students have demonstrated mastery and where further instruction may be needed.

For example, a student may obtain a satisfactory total score while performing poorly in financial calculations but strongly in conceptual understanding. Reporting only the total score could conceal this difference. Domain-level reporting provides more actionable information for feedback, remedial instruction, enrichment, and subsequent lesson planning. Domain-level results should nevertheless be interpreted cautiously when a domain contains very few items. A score based on two questions may be less stable and less comprehensive than one based on a broader set of well-designed items. Teachers should not treat a small number of responses as definitive evidence of a student's overall ability in a domain.

Furthermore, content-domain reporting does not, by itself, establish that a test measures statistically distinct latent traits. Such a psychometric claim would require appropriate empirical evidence. In the present conceptual model, multidimensionality refers primarily to planned representation of multiple instructional content domains.

CONCLUSION AND RECOMMENDATION

Conclusion

This article highlights the importance of integrating multidimensional test item development and scoring through the framework of Bloom's Taxonomy. The proposed model emphasizes comprehensive representation of instructional content domains and alignment with the six cognitive processes of Revised Bloom's Taxonomy. It also demonstrates that scoring should not rely solely on equal point allocation or perceived item difficulty, but should consider learning objectives, content representation, cognitive demands, and transparent scoring criteria. By connecting item development, assessment blueprints, and evidence-based scoring, the model offers a conceptual foundation for improving the validity, fairness, and meaningful interpretation of student learning outcomes. Future empirical research is needed to validate its applicability and examine its effectiveness across different educational levels and subject areas. Theese would be some implications, particularly for teachers: First, teachers should prepare an assessment blueprint before writing the final set of questions. This blueprint should identify the content domains, learning objectives, cognitive processes, number of items, and allocation of points. Second, teachers should review the relationship between item difficulty, cognitive demand, and scoring weight. A difficult item should not automatically receive more points, and a higher-order item should not automatically be assigned a large weight. The weight should be justified by the intended learning outcome and the quality and breadth of evidence required.

Third, teachers should prepare scoring keys or rubrics before administering the test. For multiple-choice items, the scoring key should specify the correct response and the maximum points. For constructed-response items, rubrics should define the expected evidence and criteria for awarding full or partial credit. Fourth, teachers should examine the distribution of scores after the test has been administered. This may include checking whether some content domains were too heavily represented, whether items functioned as intended, and whether scoring rules were applied consistently. Where sufficient data are available, item analysis can help teachers identify ambiguous questions or items that do not distinguish meaningfully among levels of student performance. Finally, teachers should use assessment results to improve instruction, not merely to assign grades. Domain-level findings can inform the choice of follow-up activities and help teachers identify gaps between intended learning outcomes and demonstrated achievement.

Limitations and Directions for Future Research

The present model is conceptual and should be understood as a proposed framework rather than an empirically validated scoring system. Its practical value depends on the quality of the content mapping, the appropriateness of the cognitive classifications, and the defensibility of the scoring rules. Future research should examine the model in different subjects and educational settings. Studies could compare conventional equal-weight scoring with blueprint-aligned scoring, while carefully distinguishing differences in score distributions from improvements in validity or fairness. Researchers could also investigate inter-rater agreement for constructed-response items, the stability of domain-level scores, and teachers' ability to apply the model after professional development.

REFERENCES

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

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Penerbit Longman.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Penerbit Longman.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. Penerbit David McKay.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Penerbit David McKay.

Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. Penerbit ASCD.

Daud M, Kuo BC, Duskri M, Jamaluddin, Zubair CM, Jalauddin. (2024). Assessing the Effectiveness of CAT and MCAT in Measuring Students' Understanding in Technology Education. *Journal At Ta'dib.* 2024;19(1):117-132. doi: https://doi.org/10.21111/at-tadib.v19i1.64

Daud, M. & Ansari, B.I. (2026). *Tindakan Kelas (Action for Teachers).* Penerbit Bandar Publishing.

Daud, M. & Sariakin. (2025). *Inovasi Pendidikan: Peluang dan Tantangan.* Penerbit Natural Aceh.

Daud, M. (2017). Peluang dan tantangan pendidikan wilayah pesisir Aceh. Seminar Nasional II USM 2017, *Proceeding Eksplorasi Kekayaan Maritim Aceh di Era Globalisasi dalam Mewujudkan Indonesia sebagai Poros Maritim Dunia*, 1, 32-45. Penerbit LPPM Universitas Serambi Mekkah

Daud, M., Kuo, B., Abdurahman, T., Mariati MR., Yusrizal., & Sariakin. (2022). Investigating online learning implementation in Indonesia; challenges and possible solutions. *Proceeding ICMR*, 5(2), 19-26. Penerbit LPPM Universitas Serambi Mekkah

Daud, M., Kuo, B.C., & Mok, M.M.C. (2015). The Construction of an online curriculum-based multidimensional computerized adaptive testing, *Proceedings of International Conference on Multidisciplinary Research "Improving Human Resources Quality Through Research and Development in the Era of AFTA"*, 2015, 322-328 Penerbit LPPM Universitas Serambi Mekkah

Daud, M., Khadijah, & Rahmadon. (2026). Pengembangan Tes Diagnostik Meningkatkan Capaian Hasil Belajar Pendidikan Agama Islam di MIN 7 Banda Aceh, PERISAI: Jurnal Pendidikan dan Riset Ilmu Sains, 2026/2/28, No. 5, Vol. 1, 12-21 doi: https://doi.org/10.32672/perisai.v5i1.4170

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Duskri, M., Maidiyah, E., Risnawati, R., & Ilham, S. (2017). Penerapan Model Problem Based Learning untuk Meningkatkan Kemampuan Komunikasi Matematis dalam Pemecahan Masalah Di Kelas IX-6 SMPN 8 Banda Aceh, Al Khawarizmi: Jurnal Pendidikan dan Pembelajaran Matematika, 2017/6/1, No. 1 Vol. 1, 75-101. doi:10.22373/jppm.v1i1.1734

Duskri, M., Kumaidi, K., & Suryanto S. (2014). Pengembangan tes diagnostik kesulitan Belajar matematika di SD, Jurnal Penelitian dan Evaluasi Pendidikan, 2014/6/1, No. 18, Vol. 1, 44-56 doi: 110.22373/jppm.v1i1.134

Khadijah, Daud, Rahmadon. (2026). Bentuk-Bentuk Tes dan Karakteristiknya: Kelebihan dan Kekurangannya dalam Evaluasi Pembelajaran. PERISAI: Jurnal Pendidikan dan Riset Ilmu Sains,5(1), 32-40.- Doi: https://doi.org/10.32672/perisai.v5i1.4171

Kuo, B.C., Daud, M., & Yang, C.W. (2015). Multidimensional computerized adaptive testing for testing for Indonesia junior high school biology, Eurasia Journal of Mathematics, Science and Technology Education, 2015/8/25, No. 11, Vol. 5, 1105-1118 Doi: https://doi.org/10.12973/eurasia.2015.1384a

Raymond, M. R., & Grande, J. P. (2019). A practical guide to test blueprinting. Medical Teacher, 41(8), 854–861. Link: https://doi.org/10.1080/0142159X.2019.1595556

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. CBE—Life Sciences Education, 15(1), Link: https://doi.org/10.1187/cbe.15-08-0183