JPF | Volume 14 | Number 1 | 2026 | 84 - 101

p - ISSN: 2302-8939
e - ISSN: 2527-4015

Jurnal Pendidikan Fisika

https://journal.unismuh.ac.id/index.php/jpf

DOI: 10.26618/cygjmh69

Design and Validation of an Understanding by Design-Based Science Assessment Module for Secondary Schools under Indonesia's Merdeka Curriculum

Zulhelmi*, Riza Andriani, Dina Syaflita

Department of Physics Education, Universitas Riau, Pekanbaru, 28293, Indonesia

**Corresponding author:* zulhelmi@lecturer.unri.ac.id

Received: September 15, 2025; Accepted: December 22, 2025; Published: January 17, 2026

Abstract – *Assessment practices in secondary science and physics classrooms often emphasize scoring rather than generating actionable evidence of students' understanding. This challenge becomes more urgent under the Merdeka Curriculum, which positions assessment as an integral component of learning. To address this need, this study aimed to develop and validate a teacher guide module for UbD (Understanding by Design)-based science assessment for junior and senior secondary levels by operationalizing backward design and the six facets of understanding into practical procedures, templates, and worked examples. Using a research and development approach with the 4D model (Define–Design–Develop–Disseminate), the module was produced through needs analysis, curriculum and literature review, and iterative drafting. Content validation was conducted by three validators using a four-point relevance scale, analyzed with feasibility percentages and Aiken's V complemented by Score (Wilson) confidence intervals. Practicality was examined through a limited pilot involving three science teachers who applied the module and completed a five-point response questionnaire. The results showed high feasibility across usefulness, practicality, conceptual accuracy, and language/visual design. At the same time, item-level content validity was strong (Aiken's V = 0.78–1.00) with conservative lower confidence bounds indicating at least moderate validity for all items. Teachers reported uniformly positive perceptions, with perceived usefulness rated highest (M = 4.83), although implementation of the Explanation facet was relatively more challenging (lowest item mean, M = 3.67). This study's novelty lies in providing a step-by-step, facet-based assessment design guide that explicitly links learning objectives, acceptable evidence, and analytic rubrics within a single teacher-ready resource. In conclusion, the validated module is feasible and practically promising as an evidence-centered tool to strengthen alignment between curriculum outcomes and assessment in secondary science. The module contributes to physics education by supporting teachers to design authentic, rubric-based assessments that better capture students' scientific reasoning and conceptual understanding.*

Keywords: *backward design; merdeka curriculum; module; science assessment; understanding by design*

I. INTRODUCTION

Science education is expected to cultivate scientific thinking, problem-solving competence, and students' capacity to use scientific ideas to interpret and act on everyday phenomena (National Research Council, 2012). These outcomes are difficult to achieve through content delivery alone because meaningful learning depends on how teachers translate curriculum intentions into coherent classroom experiences. In this process, assessment plays a central role: it helps teachers determine whether students are developing the targeted understandings and skills, identify misconceptions early, and decide what instructional adjustments are needed. When assessment is designed only at the end of instruction, it often serves merely as a scoring activity and fails to provide evidence of learning that can support improvement. Therefore, assessment needs to be planned from the beginning as part of an integrated design that connects learning outcomes, learning activities, and evidence of learning.

In Indonesia, the Merdeka Curriculum reinforces this alignment by emphasizing learning outcomes that are more targeted and measurable and by positioning assessment as an integral component of the learning process rather than a separate administrative requirement (Badan Standar, Kurikulum, dan Asesmen Pendidikan [BSKAP], 2025). Within this orientation, assessment is expected to serve dual functions: (1) to document achievement for reporting purposes and (2) to generate information that can be used as feedback and as a basis for instructional decision-making. This principle aligns with scholarship highlighting the importance of formative assessment, the quality of learning evidence, and the role of feedback in supporting learning progress (Black & Wiliam, 2018; Hattie & Timperley, 2007; Pellegrino et al., 2001). In addition, because science learning increasingly targets 21st-century competencies, assessment should also capture higher-order thinking, reasoning, and problem-solving rather than only factual recall (OECD, 2018). Consequently, teachers require practical approaches that help them (a) clarify what deep understanding looks like, (b) specify what evidence can demonstrate that understanding, and (c) design tasks and rubrics that make that evidence visible and usable.

One approach that explicitly supports such coherence is Understanding by Design (UbD). UbD operationalizes "backward design," in which teachers first clarify the desired learning results, then determine what counts as acceptable evidence of learning, and only afterward plan learning experiences that support students in reaching the targeted outcomes (Aslam et al., 2024). This sequence is relevant to the Merdeka Curriculum because it begins with the learning outcomes and forces the designer to articulate how those outcomes will be evidenced. By treating assessment as planned evidence rather than as an add-on, backward design reduces the risk that classroom activities drift away from intended outcomes or that assessment measures only what is

easiest to score. UbD is also aligned with the expectation that teachers use evidence to guide learning, as its design logic makes explicit what data will be collected, how they will be interpreted, and how they will inform subsequent teaching decisions (Tomlinson & McTighe, 2006).

A distinctive feature of UbD is its emphasis on deep understanding through the six facets of understanding: explanation, interpretation, application, perspective, empathy, and self-knowledge (Aslam et al., 2024). These facets provide an operational lens for translating broad learning outcomes into observable evidence. "Explanation" focuses on whether students can justify claims using concepts and principles; "interpretation" emphasizes making meaning from data, representations, or models; and "application" emphasizes transferring understanding to new contexts or problems. "Perspective" invites students to examine issues from different viewpoints; "empathy" highlights the ability to understand others' positions or experiences; and "self-knowledge" emphasizes reflection on one's own thinking and limitations. In science learning, such facets are useful because they align with goals such as reasoning with evidence, interpreting scientific information, applying concepts to real situations, and reflecting on the validity of conclusions. Thus, the six facets can function as a practical bridge between curriculum outcomes and assessment design, especially when teachers need to move beyond recall-based questions toward richer evidence of understanding.

The emphasis on assessment-as-evidence in UbD is also consistent with established work on effective assessment systems. Research on formative assessment underscores that learning improvement depends not only on collecting information but on collecting the right kinds of information—evidence that is valid for the targeted learning goals and then using that evidence to provide feedback that is timely, specific, and actionable (Black & Wiliam, 2018; Hattie & Timperley, 2007). Similarly, the assessment design literature emphasizes that instructional decisions should be grounded in the quality of the evidence gathered, including the clarity of constructs, appropriateness of tasks, and interpretability of results (Pellegrino et al., 2001). When the Merdeka Curriculum expects assessment to inform learning and when 21st-century goals emphasize complex competence, teachers need a design framework that can reliably connect outcomes to high-quality evidence. UbD offers such a framework, but its classroom impact depends on whether teachers can implement it effectively and efficiently.

Several studies indicate that UbD can clarify the relationship between goals, learning processes, and assessment. Formative assessment within a UbD framework has been reported to be associated with stronger student understanding because the evidence collected is intentionally matched to intended outcomes and supports targeted feedback (Gloria et al., 2018). The development of UbD-based assessment instruments has also been reported to support the

measurement of critical thinking skills, suggesting that UbD can be used to operationalize higher-order learning outcomes into assessable forms (Sumarni et al., 2019). These findings imply that UbD is not only a planning framework but also a potentially effective approach for assessment development, particularly when teachers need to design tasks that reveal students' reasoning and understanding.

However, within the Indonesian Merdeka Curriculum context, UbD is still more often used for curriculum development, teaching modules, or lesson planning than for constructing detailed, teacher-ready assessment tools based on the six facets. Practical guidance that helps science teachers translate learning objectives into the six facets and then use the facets as the basis for assessment tasks, questions, and scoring rubrics remains limited. At the senior high school level, UbD has been applied to local-content curriculum development and reported as feasible. Still, the primary output tends to be curriculum documents and general feasibility information rather than concrete examples that show how to convert learning objectives into assessment evidence and implementable rubrics (Nursafitri et al., 2023). At the elementary level, UbD has been widely used in IPAS programs and validated teaching modules, yet this work is centered on elementary contexts and does not directly address integrated science needs at junior and senior secondary levels (Fradina et al., 2022; Hadinda et al., 2025; Nadia, 2024; Saputra et al., 2025). In other subjects, such as mathematics, UbD-based module development has also been reported, but it does not address the need to develop a six-facet assessment for secondary science (Putra et al., 2023; Sabrina et al., 2024). Teacher training and community service programs show that teachers often need real-life examples and guidance to use backward design, such as how to make assessment tools. However, the results are usually training materials instead of written guides that teachers can use on their own when planning their lessons (Mahdiannur et al., 2024; Probosari et al., 2024; Retariandalas et al., 2025). Meanwhile, evidence from junior high schools indicates that evaluation under the Merdeka Curriculum is already underway, making practical tools that connect learning objectives to assessment evidence increasingly important (Rahmia et al., 2025).

These conditions point to a clear gap: a need for a simple but explicit guide module that not only introduces UbD but also demonstrates, in a step-by-step manner, how teachers can (1) translate learning objectives into the six facets of understanding, (2) design assessment evidence aligned with those facets, and (3) develop scoring rubrics that support consistent interpretation of student work. This kind of guide is especially useful for high school science teachers, who often have to deal with difficult skills while still being clear, doable, and in line with what the curriculum says. Without practical exemplars, UbD risks being understood as a conceptual planning framework that is difficult to operationalize into concrete classroom assessment practices.

Therefore, this study develops a UbD guide module for junior and senior secondary science within the Merdeka Curriculum context. The module explains UbD and the six facets of understanding in detail and provides step-by-step examples of how to translate learning objectives into the facets, serving as models for teachers to follow when designing assessments. Consistent with the evidence reported, this study aims to develop a content-valid module through expert review and to examine its initial clarity and practicality through teachers' evaluations as a basis for refinement before broader classroom-based trials.

II. METHODS

This study used a Research and Development (R&D) approach, employing the 4D model (Define, Design, Develop, and Disseminate) (Thiagarajan et al., 1974), to develop a UbD-based module for science assessment and to examine its content validity and initial practicality. The study was conducted at the Faculty of Teacher Training and Education at Universitas Riau from June 2024 to April 2025.

1. Define stage

The Define stage aimed to identify teachers' needs and establish the design requirements for a module that guides science assessment based on UbD for junior secondary (SMP) and senior secondary (SMA) levels within the context of the Merdeka Curriculum. Data were collected from three sources: (1) an open-ended survey involving 20 SMP and SMA science teachers; (2) a review of two BSKAP (2025) documents covering Phases D–E and Phase F; and (3) a review of relevant scholarly articles on the implementation of UbD. Survey responses were analyzed using simple coding techniques to identify recurring themes and needs. The findings from the Define stage were then translated into specific design requirements and mapped onto module features, as presented in Table 1, to inform the subsequent Design stage.

Table 1. Mapping of defined findings to module features

Source	Main need	Module feature developed
Open-ended survey (20 SMP–SMA science teachers)	Difficulty designing assessment based on the six UbD facets	Step-by-step UbD assessment guide and a facets-based planning template
Open-ended survey (20 SMP–SMA science teachers)	Lack of concrete examples for deriving assessment from learning objectives	Worked example: objective → facet → evidence → rubric (ready-to-follow model)
Open-ended survey (20 SMP–SMA science teachers)	Unclear linkage between the Merdeka Curriculum project/performance tasks and UbD facets	Sample performance task/project mapped to facets, including the assessed evidence
Open-ended survey (20 SMP–SMA science teachers)	Uncertainty about scoring criteria for facet-based assessment	Analytic rubric format, level descriptors, and an adaptable scoring template

Document review (BSKAP, 2025; Phases D–E and F)	Facets are briefly described; operational examples are limited.	Module sections that model operational procedures and provide ready-to-use formats
Literature review (UbD in the Merdeka context)	UbD is more often used for teaching modules/lesson plans than as an assessment guide.	The module is positioned as an assessment guide focused on facets–evidence–rubric mapping.

2. Design stage

In the design stage, the design requirements from the define stage were converted into a structured UbD-based assessment guide module for SMP and SMA. The module followed the three stages of backward design: identifying desired results, determining acceptable assessment evidence, and planning aligned learning experiences (Aslam et al., 2024). To support teacher use, the module was organized as a practical workflow and complemented with templates, worked examples translating objectives into evidence, and examples of performance tasks/projects aligned to the six facets. The design stage also produced a concise operational mapping of the six facets into assessment indicators and evidence types, along with a brief rubric focus in Table 2, to ensure the rubric logic was explicit and replicable.

Table 2. Example mapping of the six UbD facets to indicators and assessment evidence

Facet	**Assessment indicator**	**Evidence**	**Rubric focus**
Explanation	Explains concepts and scientific cause and effect clearly	Short response/presentation	Concept accuracy; coherence
Interpretation	Interprets graphs/tables and derives meaning	Data interpretation task	Interpretation accuracy: evidence use
Application	Applies concepts to propose solutions in new contexts	Scenario response / action plan	Appropriateness; justification
Perspective	Compares viewpoints and evaluates arguments	Debate/position paper	Balance, logic, and evidence
Empathy	Describes impacts on others and the social context	Role-based reflection	Context relevance, sensitivity, accuracy
Self-knowledge	Reflects on own understanding and next steps	Reflection journal / self-assessment	Depth of reflection; improvement plan

3. Develop stage

In the Develop stage, the module blueprint was expanded into a complete draft containing conceptual explanations, templates, worked examples, sample performance tasks, and analytic rubrics. Three validators, two assessment experts, and one science teacher reviewed the module. Experts were selected based on experience in curriculum-aligned assessment design, familiarity with UbD, and competence in constructing classroom assessment instruments, consistent with recommendations for content validation studies (Almanasreh et al., 2019; Terwee et al., 2018; Gilbert & Prion, 2016).

Each validator rated 16 items using a four-point relevance scale. Overall feasibility was summarized using a feasibility index:

$$\text{Feasibility (\%)} = \frac{\text{Total score}}{\text{Maximum score}} \times 100.$$

The total obtained score was the sum of all ratings from all validators, and the maximum possible score was the number of items multiplied by the highest rating (4). Percentages were interpreted using criteria: 81–100% (very feasible), 61–80% (feasible), 41–60% (fairly feasible), 21–40% (less feasible), and 0–20% (not feasible). Item-level content validity was quantified using Aiken's V (Aiken, 1985). For a four-point relevance scale ($l_0 = 1;\ c = 4$), the index was calculated as:

$$V = \frac{\Sigma(r - l_0)}{[n(c - 1)]}$$

where:
r = the score given by a validator for an item
n = number of validators
l_0 = lowest score on the scale
c = number of scale categories

To make the interpretation explicit, descriptive cut-offs were reported based on the meaning of the 1–4 relevance ratings: was classified as very valid, as valid, and as requiring revision (Aiken, 1985). Because the expert panel was small, decisions were strengthened using Wilson score (confidence) intervals for Aiken's V, as recommended for small panels (Penfield & Giacobbi, 2004). Items were considered to have adequate inferential validity when the lower bound of the 90% score interval was ≥ 0.50 (Penfield & Giacobbi, 2004). This approach was used to reduce over-interpretation of point estimates and to provide a more stable decision basis under limited rater conditions. Results from this stage guided targeted revisions to item wording, conceptual explanations, and the clarity of examples and rubrics.

4. Disseminate stage

A full dissemination stage (large-scale implementation) was not conducted. Instead, a limited pilot practicality test was carried out with three junior secondary science teachers in Pekanbaru to obtain early evidence of clarity, usefulness, and implementation feasibility. All three participants had prior training in the Merdeka Curriculum. Each teacher (a) studied the validated module, (b) used it to design UbD-based assessment tasks relevant to their current teaching, and (c) completed the Teacher Response Questionnaire. The instrument measured four constructs: perceived usefulness, clarity and practicality, implementation, and effectiveness (perceived appropriateness and usability from the teachers' perspective), and satisfaction using a 1–5 scale, followed by open-ended questions about strengths, challenges, and improvement suggestions.

III. RESULTS

Based on these requirements, the study designed a teacher guide module for UbD-based science assessment in junior and senior secondary education. The module provides concise explanations of the UbD framework and the six facets of understanding, and it focuses on operational support through step-by-step templates and worked examples. A worked example using an integrated science topic (global warming) demonstrates how learning objectives are mapped to facet-based indicators, evidence types (written items, performance tasks, and projects), and analytic scoring criteria. Structurally, the module follows the three stages of backward design: identifying desired results, determining acceptable evidence, and planning aligned learning experiences. The six facets are embedded across examples to make the assessment logic explicit and replicable, while maintaining alignment with the Merdeka Curriculum's emphasis on conceptual understanding, transfer, and authentic assessment.

1. Module Feasibility Based on Assessment Aspects

The feasibility of the module is evaluated across four main aspects: (1) usefulness, (2) feasibility/practicality, (3) accuracy of concepts and content, and (4) language and visual design. Each aspect is assessed using four statements from three validators, with a maximum score of 4 per item (maximum per aspect = 48).

Table 3. Summary of module feasibility based on assessment aspects

No	Assessment aspect	Number of items	Maximum score	Score obtained	Feasibility percentage (%)
1	Module usefulness	4	48	45	93.75
2	Module feasibility/practicality	4	48	43	89.58
3	Accuracy of concepts and content	4	48	41	85.42
4	Language and visual design	4	48	45	93.75

The scores in Table 3 indicate that the usefulness and language and visual design aspects achieved the highest feasibility percentages (93.75% each). The feasibility/practicality aspect obtained 89.58%, while the accuracy of concepts and content reached 85.42%. All four aspects fall within the 80% range, indicating that the three validators assigned relatively high scores to the module's feasibility across all assessed aspects.

At the item level, most items received scores of 11–12 out of the maximum 12 (three validators × score 4), with item feasibility percentages ranging from 83.33% to 100%. This indicates that each item within the four aspects was considered “appropriate” or “highly appropriate” by the validators.

2. Content Validity of the Module Items

The content validity of the module items was analyzed using Aiken's V index for each item, complemented by the calculation of Score (Wilson) confidence intervals at the 90% and 95% levels for the population value. A total of 16 items were evaluated by three validators using a 1–4 rating scale. The score frequency pattern for each item, the V value, as well as the lower and upper bounds of the confidence intervals, are presented in Table 4.

Table 4. Aiken's V and the 90%/95% confidence intervals for all module items

Item code	f(1)	f(2)	f(3)	f(4)	V	L90	U90	L95	U95
01.01	0	0	1	2	0.89	0.62	0.97	0.56	0.98
01.02	0	0	0	3	1.00	0.77	1.00	0.70	1.00
01.03	0	0	1	2	0.89	0.62	0.97	0.56	0.98
01.04	0	0	1	2	0.89	0.62	0.97	0.56	0.98
02.01	0	0	2	1	0.78	0.50	0.92	0.45	0.94
02.02	0	0	2	1	0.78	0.50	0.92	0.45	0.94
02.03	0	0	0	3	1.00	0.77	1.00	0.70	1.00
02.04	0	0	1	2	0.89	0.62	0.97	0.56	0.98
03.01	0	0	2	1	0.78	0.50	0.92	0.45	0.94
03.02	0	0	2	1	0.78	0.50	0.92	0.45	0.94
03.03	0	0	2	1	0.78	0.50	0.92	0.45	0.94
03.04	0	0	1	2	0.89	0.62	0.97	0.56	0.98
04.01	0	0	0	3	1.00	0.77	1.00	0.70	1.00
04.02	0	0	1	2	0.89	0.62	0.97	0.56	0.98
04.03	0	0	1	2	0.89	0.62	0.97	0.56	0.98
04.04	0	0	1	2	0.89	0.62	0.97	0.56	0.98

The results of Aiken's V analysis and the Score (Wilson) confidence interval show that all items have good to very high content validity. The V values range from 0.78 to 1.00. The validators considered none of the items weak. Several items have V = 1.00. This value shows that all three validators gave the same rating. They unanimously agreed that these items are highly relevant to the measured construct. Examples include items that assess how well the module guides teachers in designing UbD-based assessments and items that evaluate the clarity of the module's language.

The score confidence interval provides additional information. It tells us not only how large the V value is but also how certain we are of its estimate of content validity in the population of experts. The lower bound of the 90% confidence interval (L90) for all items falls between 0.50 and 0.77. This shows that, even with only three validators, the probability that an item's content validity is below the "moderate" category is fairly small.

The upper bounds of the confidence intervals (U90 and U95) are close to 1.00. This strengthens the indication that most items fall into the high-relevance category. However, a few items have V = 0.78 with L90 around 0.50. This means that the level of certainty for these items

is slightly lower than for the others. These items are still valid, but they are more sensitive to differences in expert judgment. These items are still valid, but they are more sensitive to differences in expert judgment. In module development, such items can be prioritized for light revision, such as clarifying wording, adding more concrete examples, or adjusting terminology to align with field practices.

3. Overview of pilot practicality test

This pilot involved only three teachers (n = 3) and therefore provides preliminary evidence of practicality. The results are not intended to be generalized, and this stage was designed for targeted refinement rather than broad validation or claims about classroom implementation success or student learning impact.

Following the procedure described earlier, each teacher studied the validated UbD-based assessment module and used it to design, for example, backward-designed assessment tasks. They then completed the Teacher Response to UbD Assessment Module questionnaire (Parts B–E) and provided written responses to three open-ended questions (Part F). Table 5 presents descriptive statistics for the four constructs, assessed using a 5-point Likert scale. These statistics correspond directly to the analysis procedures specified in the methodology.

Table 5. Construct-level descriptive statistics (N = 3)

Construct	Items	Σ Score	Mean	SD (pop)	% of max	Δ from neutral (3)
Perceived usefulness	4	58	4.83	0.37	96.7%	+1.83
Clarity & practicality	4	56	4.67	0.47	93.3%	+1.67
Implementation	4	53	4.42	0.64	88.3%	+1.42
Effectiveness & satisfaction	4	56	4.67	0.47	93.3%	+1.67
Overall	—	**223**	**4.65**	**0.52**	**93.0%**	**+1.65**

The results show uniformly high ratings across all constructs. Perceived usefulness achieved the highest mean (4.83), suggesting that teachers considered the module highly relevant and supportive for understanding UbD. Implementation displayed the greatest variability (SD = 0.64), indicating differences in how confidently teachers perceived their ability to design UbD-based assessments after using the module.

Table 6. Summarizes the mean item scores and identifies items receiving the highest or lowest ratings.

No	Construct	Item focus	Mean score	% of max
1	Perceived usefulness	UbD concept understanding	4.67	93.3
		Backward design example quality	5.00	100.0
		Alignment w/Merdeka curriculum	5.00	100.0
		Formative–summative enrichment	4.67	93.3

2	Clarity & practicality	Clarity of instructions	4.33	86.7
		Clarity of facet examples	4.33	86.7
		Ease of lesson planning	5.00	100.0
		Reasonable time demand	4.33	86.7
3	Implementation	Explain the facet	3.67	73.3
		Apply facet	4.33	86.7
		Perspective facet	5.00	100.0
		Self-knowledge facet	4.33	86.7
4	Effectiveness & satisfaction	Improved assessment skill	5.00	100.0
		Innovation motivation	4.33	86.7
		Satisfaction with outputs	4.33	86.7
		Continuous use intention	4.33	86.7

Several items received maximum scores (5) from all teachers (n = 3). These items are mainly related to the module's alignment with the Merdeka Curriculum, the clarity of the backward design examples, and teachers' perceived improvement in assessment design skills. This pattern suggests strong agreement that the module is conceptually relevant and helps teachers understand UbD principles for assessment planning.

The lowest mean score (M = 3.67) appeared for the Explanation facet within the Implementation construct. Although still above the neutral point (3), this score indicates that designing assessments for the Explanation facet was relatively more challenging than for other facets. Overall, the mean scores across all 16 items remained above 4.0, indicating consistently positive perceptions of the module's usefulness, clarity, implementation support, and overall effectiveness (perceived appropriateness and usability from the teachers' perspective).

IV. DISCUSSION

The feasibility results support the interpretation that the module meets key quality conditions for early adoption as a professional teacher guide. Validators assigned high feasibility percentages across the four assessed aspects—usefulness, feasibility/practicality, conceptual accuracy, and language/visual design—with the strongest performance noted in usefulness and language/visual design. This pattern is not trivial because teacher-facing resources often fail at the implementation stage when conceptual content is not accompanied by readability, navigability, and concrete usability; a guide that is difficult to interpret or apply cannot effectively support classroom change even if it is theoretically sound, and evidence from rubric-system implementation likewise shows that user experience and practical considerations are central to successful uptake (Gregori-Giralt & Menéndez-Varela, 2021). In the context of UbD, feasibility is directly tied to whether the backward-design workflow is sufficiently explicit for teachers to use it as a planning tool rather than merely as a conceptual reference (Aslam et al., 2024;

Tomlinson & McTighe, 2006). The high feasibility ratings, therefore, support the idea that the module successfully turns UbD into a planning sequence that can be used in real life. This is important because teachers are expected to provide assessment evidence that supports learning decisions (BSKAP, 2025). In assessment terms, such feasibility is aligned with the requirement that assessment systems in practice must be not only valid but also workable within classroom constraints, otherwise evidence quality cannot be consistently produced or used (Pellegrino et al., 2001), and recent evidence on formative assessment underscores that teacher prerequisites such as knowledge/skills, social conditions, and contextual supports strongly shape whether formative approaches can be enacted effectively in classrooms (Schildkamp et al., 2020).

The content validity findings further strengthen the module's scientific credibility by showing that experts judged all module items to be relevant representations of the intended construct. The Aiken's V values, which range from 0.78 to 1.00, show that there is good to very high agreement on item relevance. This is in line with Aiken's V, which is a well-known way to measure expert judgment in content validation (Aiken, 1985). Importantly, the study's use of Score (Wilson) confidence intervals is methodologically significant because it addresses a known limitation of small expert panels, where point estimates can appear strong while uncertainty remains substantial (Penfield & Giacobbi, 2004). The observed lower bounds of the 90% intervals (L90) between 0.50 and 0.77 suggest that, even with conservative inference, items generally meet at least moderate validity. This supports the interpretation that the module's core components, such as the alignment between UbD stages, six-facet reasoning, evidence types, and rubrics, are defensible as content-relevant while simultaneously providing diagnostic information for refinement. This dual function is consistent with the broader validity-development literature emphasizing that content validation is not only a gatekeeping exercise but also a mechanism for identifying where specification clarity can be improved to reduce interpretive ambiguity in later use (Almanasreh et al., 2019; Terwee et al., 2018) and more recent methodological synthesis similarly highlights the need for transparent, systematic approaches and careful interpretation of content validity indices when developing educational instruments or materials (Almanasreh et al., 2019). In other words, the results justify treating the module as sufficiently valid for early use while still acknowledging that items with lower V values and more conservative interval bounds are natural targets for improving wording precision and example specificity, particularly when the resource is intended for broad teacher audiences with diverse prior knowledge.

Teacher practicality responses provide an important second line of evidence by examining whether the validated module is perceived as usable and beneficial by its intended users. Teachers reported uniformly high ratings across constructs, with perceived usefulness achieving the highest mean (M = 4.83), suggesting that the module successfully communicates UbD-based assessment

logic in a way that teachers recognize as directly relevant to their work. This result aligns with formative assessment scholarship, which emphasizes that assessment contributes to learning when teachers can translate evidence into meaningful feedback and instructional adjustment (Black & Wiliam, 2018; Hattie & Timperley, 2007), and contemporary theorizing further clarifies how classroom assessment practices must be embedded within pedagogy to explain their effects and to manage the formative–summative relationship in teachers' work (Black & Wiliam, 2018). A guide that is perceived as useful is more likely to be adopted, and adoption is the necessary condition for any downstream impact on assessment quality. The strong usefulness ratings also resonate with the UbD proposition that teachers benefit when "acceptable evidence" is made explicit at the start of planning, because this clarifies what understanding looks like and what student work should demonstrate (Aslam et al., 2024; Tomlinson & McTighe, 2006). In line with prior work suggesting that UbD-oriented design can support more intentional assessment construction particularly for higher-order outcomes these teacher perceptions can be interpreted as early confirmation that the module's templates and worked examples provide an accessible entry point into evidence-centered assessment planning (Gloria et al., 2018; Sumarni et al., 2019). In addition, recent meta-analytic evidence on educational feedback shows that effects are highly conditional and depend on how feedback is designed and interpreted, strengthening the rationale for providing teachers with explicit evidence rules and rubric guidance so feedback can be more actionable (Wisniewski et al., 2019).

At the same time, the results also show a nuanced and theoretically meaningful constraint: Implementation ratings displayed the greatest variability (SD = 0.64), and the lowest mean score (M = 3.67) occurred in relation to implementing assessments for the Explanation facet. This pattern is consistent with the reality that explanation in science is among the most demanding targets to assess because it requires students to articulate causal mechanisms, connect claims to evidence, and coordinate concepts coherently, which cannot be evaluated by simple right/wrong scoring rules (National Research Council, 2012). Within the logic of evidence-centered design, tasks that elicit explanation must be accompanied by clear construct definitions and scoring criteria; otherwise, teacher judgments become inconsistent, and the resulting evidence becomes harder to interpret and use for feedback (Pellegrino et al., 2001). UbD explicitly frames explanation as a facet of understanding that demands justification and meaning-making, which naturally increases the need for analytic rubrics and performance-level descriptors that separate conceptual accuracy from reasoning quality (Aslam et al., 2024). Therefore, the Explanation result can be interpreted as a targeted signal for strengthening operational scaffolds, rather than indicating a weakness in the module's overarching framework. This interpretation is consistent with recent science education evidence showing that secondary students often struggle to produce

elaborated causal explanations and that teachers experience difficulties in supporting and assessing scientific writing and reasoning; structured scaffolds and explicit frameworks can improve the quality of written scientific explanations (McLure, 2022). In parallel, recent rubric-development studies in science education demonstrate that teacher-usable rubrics can help clarify key dimensions and criteria for planning and assessing complex learning goals, supporting the inclusion of calibrated exemplars and explicit criteria to strengthen teachers' scoring consistency and planning confidence (Mang et al., 2023). The module can therefore support teachers more robustly by making the explanation construct more explicit through clearer minimum evidence requirements and by providing calibrated student-response exemplars that illustrate how rubric criteria apply across performance levels, which would also strengthen the feedback loop known to drive learning improvement (Hattie & Timperley, 2007; Wisniewski et al., 2019). This interpretation aligns with the broader view that early adoption of UbD becomes easier when teachers receive modeling that bridges abstract design principles to concrete classroom artifacts and scoring routines (McTighe et al., 2020).

The present findings also sharpen the study's contribution in relation to the Indonesian UbD development landscape. Prior UbD-related efforts in Indonesia have frequently emphasized instructional module development, curriculum documentation, or feasibility studies, including in elementary IPAS contexts, while fewer studies provide teacher-ready, step-by-step guidance that focuses specifically on constructing assessment evidence and rubrics organized by the six facets of understanding (Fradina et al., 2022; Hadinda et al., 2025; Nadia, 2024; Nursafitri et al., 2023; Saputra et al., 2025). UbD applications in other subjects, such as mathematics, further demonstrate the spread of backward-design thinking, yet these do not directly address science-specific assessment challenges related to reasoning and explanation quality (Putra et al., 2023; Sabrina et al., 2024). In addition, the existence of teacher training and mentoring initiatives indicates that teachers often need structured examples and guided practice to translate UbD into implementable assessment products, implying that stand-alone conceptual materials are insufficient for sustained implementation (Mahdiannur et al., 2024; Probosari et al., 2024; Retariandalas et al., 2025). This aligns with international evidence that formative assessment enactment depends strongly on teacher prerequisites and supportive conditions, reinforcing the relevance of practical resources that reduce design ambiguity and increase usability (Schildkamp et al., 2020). Given that schools are already conducting evaluations under the Merdeka Curriculum, the need for operational assessment resources that strengthen alignment and evidence clarity is immediate (Rahmia et al., 2025; BSKAP, 2025). In this respect, the module validated in this study contributes not only as a new product but also as a concrete response to implementation readiness: it targets the specific step where teachers often struggle translating outcomes into

evidence and scoring criteria and provides structured scaffolds consistent with UbD and evidence-centered assessment principles (Aslam et al., 2024; Pellegrino et al., 2001).

Finally, the development-stage design must appropriately constrain the interpretation of these findings. With three validators and a pilot practicality test involving three teachers, the study provides early evidence of feasibility, content relevance, and initial user acceptance but does not support generalizable claims about large-scale implementation success or effects on student learning outcomes. This boundary is consistent with the logic of R&D using staged refinement, where expert validation and limited pilots serve to strengthen product quality before broader dissemination and classroom trials (Thiagarajan et al., 1974). Nevertheless, within these boundaries, the convergence of high feasibility scores, strong content validity indices supplemented by confidence intervals, and positive teacher practicality responses provides a coherent evidential basis to argue that a UbD-based assessment guide module can be developed and validated as a credible tool for supporting evidence-centered science assessment under the Merdeka Curriculum, while also generating a clear refinement direction centered on explanation-focused task design and rubric calibration to enhance interpretability and formative usefulness.

V. CONCLUSION AND SUGGESTION

This study developed a UbD-based science assessment guide module for junior and senior secondary education using the 4D R&D model and evaluated its validity and initial practicality. Expert review results indicated that the module was highly feasible across key aspects of usefulness, practicality, conceptual accuracy, and language/visual design. Item-level content validity analysis using Aiken's V complemented by Score (Wilson) confidence intervals further showed that all module components achieved good to very high relevance, supporting the module's adequacy as an evidence-centered guide for aligning learning objectives, assessment evidence, and rubric criteria within the Merdeka Curriculum context. The limited teacher pilot also yielded consistently positive ratings, particularly for perceived usefulness, suggesting that the module can support teachers in planning UbD-aligned assessments and in operationalizing the six facets of understanding through templates and worked examples.

Despite these promising results, this study has limitations that should be acknowledged. The validation utilized a limited cohort of experts (n = 3), and the practicality assessment comprised merely three educators (n = 3). Consequently, the results offer initial evidence and are not applicable to larger teacher populations or diverse educational settings. Furthermore, the study did not perform classroom-based implementation trials or assess student learning outcomes, implementation fidelity, or longitudinal shifts in teachers' assessment methodologies. Future

research should therefore involve larger and more diverse samples, conduct field trials in authentic classroom settings, and examine the effects of the module on teachers' assessment quality and students' scientific reasoning, particularly for facets that teachers found more challenging, such as scientific explanation. Within these boundaries, the study contributes to physics education by providing a validated and teacher-oriented assessment design resource that translates UbD principles into practical procedures for developing authentic, rubric-based assessments, thereby supporting stronger alignment between curriculum outcomes and evidence of learning and offering a scalable foundation for improving assessment practice in secondary science and physics classrooms.

ACKNOWLEDGMENTS

The authors express their appreciation to the Dean of the Faculty of Teacher Training and Education (FKIP), Universitas Riau, and to the FKIP Universitas Riau DIPA funding scheme for supporting this research. The authors also extend their gratitude to the expert validators and the science teachers who participated in the validation process and pilot practicality test.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Aslam, A., Ahmad, S., Siller, H.-S., & Nasreen, A. (2024). Impact of the Understanding by Design model on the science academic achievement of fifth grade students in Pakistan. *Asia-Pacific Science Education, 10*(1), 113–153. https://doi.org/10.1163/23641177-bja10078

Badan Standar, Kurikulum, dan Asesmen Pendidikan. (2025). *Panduan mata pelajaran Ilmu Pengetahuan Alam (IPA): Fase D dan E*. Kementerian Pendidikan Dasar dan Menengah Republik Indonesia.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Fradina, R. A., Cahyono, E., & Sumarni, W. (2022). Development of natural and social science learning programme (IPAS) in elementary school with Understanding by Design (UbD) framework to improve concept mastery and problem-solving ability. *Journal of Primary Education, 11*(3), 399–407. https://journal.unnes.ac.id/sju/jpe/article/view/77327

Gilbert, G. E., & Prion, S. (2016). Making sense of methods and measurement: Lawshe's content validity index. *Clinical Simulation in Nursing, 12*(12), 530–531. https://doi.org/10.1016/j.ecns.2016.08.002

Gloria, R. Y., Sudarmin, S., Wiyanto, W., & Indriyanti, D. R. (2018). The effectiveness of

formative assessment with Understanding by Design (UbD) stages in forming habits of mind in prospective teachers. *Journal of Physics: Conference Series, 983*, 1-5. https://doi.org/10.1088/1742-6596/983/1/012158

Gregori-Giralt, E., & Menéndez-Varela, J. L. (2021). The content aspect of validity in a rubric-based assessment system for course syllabuses. *Studies in Educational Evaluation, 68*, 1-12. https://doi.org/10.1016/j.stueduc.2020.100971

Hadinda, S. T., Rohana, R., & Sayidiman, S. (2025). Analisis efektivitas modul ajar menggunakan pendekatan Understanding by Design (UbD) pada ketercapaian tujuan pembelajaran di sekolah dasar. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 10*(1), 111–120. https://journal.unpas.ac.id/index.php/pendas/article/view/23693

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Mahdiannur, M. A., Martini, M., Astriani, D., Setiawan, B., & Qosyim, A. (2024). Pemberdayaan guru IPA: Integrasi strategi Understanding by Design untuk meningkatkan kompetensi profesional dalam mendesain rencana pembelajaran. *ABDIMASY: Jurnal Pengabdian dan Pemberdayaan Masyarakat, 5*(2), 136–149. https://doi.org/10.46963/ams.v5i2.2329

Mang, H. M. A., Chu, H. E., Martin, S. N., & Kim, C. J. (2023). Developing an evaluation rubric for planning and assessing SSI-based STEAM programs in science classrooms. *Research in Science Education, 53*(6), 1119–1144. https://doi.org/10.1007/s11165-023-10123-8

McLure, F. (2022). The thinking frames approach: Improving high school students' written explanations of phenomena in science. *Research in Science Education, 53,* 173-191. https://doi.org/10.1007/s11165-022-10052-y

Mctighe, J., Silver, H., & Perini, M. (2020). *Deep Learning is Doable: Five Strategies for Supporting Deep Learning in Virtual Environments*. https://jaymctighe.com/wp-content/uploads/2020/12/Deep-Virtual-Learning-article-12.9.10.pdf

Nadia, A. R. U. (2024). Pengembangan modul ajar Kurikulum Merdeka menggunakan pendekatan Understanding by Design pada mata pelajaran IPAS kelas IV di MIN 9 Bandar Lampung. *Doctoral Dissertation*, UIN Raden Intan Lampung. https://repository.radenintan.ac.id/id/eprint/32676

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://knilt.arcc.albany.edu/images/f/f8/A_Framework_for_K-12_Science_Education_A_New_Conceptual_Framework.pdf

Nursafitri, L., Firdaus, T., Sudomo, R. I., & Kurniasih, A. (2023). Development of local content curriculum based on the Merdeka Curriculum for high school in East Kalimantan Province. *QALAMUNA: Jurnal Pendidikan, Sosial, dan Agama, 15*(2), 695–704. https://doi.org/10.37680/qalamuna.v15i2.2933

OECD. (2018). *The future of education and skills: Education 2030*. OECD Publishing.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Penfield, R. D., & Giacobbi, P. R., J. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Probosari, R. M., Indriyanti, N. Y., Utami, B., Fakhrudin, I. A., & Khasanah, A. N. (2024). Pemberdayaan guru melalui backward design lesson plan sebagai implementasi Kurikulum

Merdeka. *Masyarakat Berdaya dan Inovasi, 5*(1), 146–151. https://mayadani.org/index.php/MAYADANI/article/view/196

Putra, Z. R. A., Pratama, C. E., Fauziyah, N., & Pramudito, M. S. (2023). Pengembangan modul ajar matematika berdiferensiasi berbasis Understanding by Design (UbD). *Postulat: Jurnal Inovasi Pendidikan Matematika, 4*(1), 128–139. https://doi.org/10.30587/postulat.v4i1.5695

Rahmia, S. H., Mutiani, M., Nuraini, F., Triyono, S., & Syarifuddin, S. (2025). Integration of learning design and evaluation in the implementation of Kurikulum Merdeka: Evidence from public junior high schools in Banjarmasin. *The Kalimantan Social Studies Journal, 7*(1), 98–116. https://ppjp.ulm.ac.id/journals/index.php/kss/article/view/16735

Retariandalas, R., Ramli, D. P. S., Purnama, I. M., & Simanjuntak, P. (2025). Pelatihan pengembangan modul ajar menggunakan Understanding by Design (UbD) untuk guru matematika MA di Jakarta. *BESIRU: Jurnal Pengabdian Masyarakat, 2*(6), 581–587. https://doi.org/10.62335/besiru.v2i6.1365

Sabrina, K. A., Fitri, M. A., Nainggolan, N., & Mulyatna, F. (2024). Understanding by Design: Identifikasi hasil yang diinginkan dan penerapannya dalam pembelajaran matematika. *JP3 (Jurnal Pendidikan dan Profesi Pendidik), 10*(2), 147–153. https://doi.org/10.26877/jp3.v10i2.22743

Saputra, V. P., Purnamasari, I., & Suyitno, S. (2025). Designing an IPAS lesson plan with Understanding by Design approach for grade V elementary school. *Primary: Jurnal Pendidikan Guru Sekolah Dasar, 14*(5), 659–670. https://doi.org/10.33578/jpfkip.v14i5.p659-670

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, 1-16. https://doi.org/10.1016/j.ijer.2020.101602

Sumarni, W., Supardi, K. I., & Widiarti, N. (2019). Development of assessment instruments to measure critical thinking skills. *IOP Conference Series: Materials Science and Engineering, 349*, 1-11. https://doi.org/10.1088/1757-899X/349/1/012066

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: *A Delphi study. Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Indiana University.

Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction and Understanding by Design: Connecting content and kids*. ASCD.

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*(3087), 1-14. https://doi.org/10.3389/fpsyg.2019.03087