OPEN ACCESS
Indonesian Journal of Pedagogy and Teacher Education Vol 4 No 2 (2026) https://doi.org/10.58723/ijopate.v4i2.860

Check for updates

Research Article

Four-Tier Assessment for Islamic Religious Education in Sports: Development and Preliminary Psychometric Evaluation

Received: August 07, 2026	Revised: August 29, 2026	Accepted: September 14, 2026	Published: September 18, 2026

Ari Tri Fitrianto*, Aulia Putri Yuliasti, Abdul Hafiz, Muhammad Irwaysh Abdhi

Abstract:
Background: Four-tier assessments can distinguish answer selection, reasoning, and confidence levels, but their use in Islamic Religious Education and sports education remains limited. This study therefore focuses on instrument development and preliminary psychometric evaluation using partial least squares structural equation modeling (PLS-SEM), rather than on diagnostic prevalence.
Aims: This study aimed to develop a 12-item four-tier assessment covering Akhlaq, Fiqh, Tawhid, and Qur'anic Studies in sports contexts and to evaluate preliminary first-order measurement evidence.
Methods: Using a research and development approach, the instrument was constructed through curriculum document analysis, concept mapping, open-ended pre-testing with 62 students, distractor development, and prototype assembly. A separate sample of 250 sports-education students completed the prototype version. The available analysis matrix was examined with first-order PLS-SEM in SmartPLS 4 Pro.
Results: All 12 indicators loaded significantly on their intended first-order domains. Composite reliability ranged from 0.754 to 0.777, average variance extracted from 0.507 to 0.537, and heterotrait–monotrait ratios from 0.320 to 0.648. Cronbach's alpha was modest (0.510–0.569), indicating weak within-domain covariance. Keyed option positions were also strongly unbalanced, a position cue constraining interpretation of the estimates.
Conclusion: The study provides preliminary first-order psychometric evidence for a four-tier assessment in an interdisciplinary socio-religious context. Because separate Tier 1–4 responses are not retained, observed diagnostic-category prevalence is not reported. Stronger validation requires raw tier-level responses, a documented scoring transformation, less transparent distractors, balanced key positions, and independent item response theory or Rasch analysis.

Keywords: Diagnostic Assessment, Four-tier test, Islamic Religious Education, Misconception, Sports education

1. INTRODUCTION

Islamic Religious Education (IRE) in higher education plays a crucial role in shaping students' character, ethical values, and worldview across many domains of life (Ginanjar et al., 2025; Rachmawati et al., 2024), including sport (Jani et al., 2025). From a constructivist perspective, students actively build knowledge by integrating new information with prior knowledge and experience (Do et al., 2023).

Students nevertheless enter the classroom carrying pre-existing conceptions rooted in personal belief and assumption, and where these do not align with established religious or scientific principles they may harden into misconceptions (Bryce & Blown, 2024; Taber, 2019). Previous studies have generally assessed students' understanding of religious values with conventional instruments such as Likert-scale questionnaires (Judijanto et al., 2024), multiple-choice tests, open-ended questions, and other diagnostic tools (Atika et al., 2023; Jani et al., 2025; Putra & Anhusadar, 2025). Each carries limitations. Interviews demand considerable time and resources and scale poorly. Conventional multiple-choice items frequently fail to separate genuine understanding from answers arrived at by guessing (Al-Lawama & Kumwenda, 2023; Barnard, 2025).

Multi-tier diagnostic formats were developed to overcome precisely these limitations. Two-tier tests pair an answer with a reason, three-tier tests add a single confidence rating, and the four-tier tests rates confidence separately for the answer and for the reason, a systematic review of 71 multi-tier studies found that the two-tier form still accounts for most applications while four-tier instruments remain comparatively rare (Duran & Dikmenli, 2024). That separation is what permits genuine conceptual understanding to be distinguished from correct answers reached by guessing, and insufficient knowledge to be distinguished from firmly held misconceptions (Yan & Subramaniam, 2018). Kaltakci-Gurel et al. (2017) set out the development and classification framework now most

widely used, although evidence from physics education cautions that confidence in an incorrect answer is informative but not by itself sufficient to establish a misconception (Hull et al., 2022). Subsequent applications have extended the format to chemical kinetics (Habiddin & Page, 2019), to physics using modern test theory Istiyono et al. (2023) and Rasch modeling (Aripiani et al., 2023), to biology (Andriani et al., 2021), to environmental topics such as global warming (Aksoy & Erten, 2022), and to primary science including mobile-administered formats (Astuti et al., 2023; Desstya et al., 2025), while five- and six-tier extensions have also been trialed (Imaduddin et al., 2023; Isra & Mufit, 2023).

Misconception research has repeatedly shown why this precision matters: students' alternative conceptions resist conceptual change and cannot be identified reliably through surface-level assessment formats (Soeharto & Csapó, 2022; Taber, 2019). Within physical education, a parallel literature has established that sport can serve as a meaningful vehicle for character development, sportsmanship, and moral growth when it is pedagogically connected to values-based education (Aygun et al., 2024; Bisa, 2023; Bronikowska et al., 2024; Jadwiszczak et al., 2025; Rahayu et al., 2026; Shen et al., 2022). Meta-analytic evidence indicates that structured pedagogical models in physical education improve prosocial attitudes and motivational outcomes relative to traditional instruction (Dai et al., 2024; Manninen & Campbell, 2022; Zhang, Soh, et al., 2024; Zhang, Xiao, et al., 2024). These two bodies of work have, however, developed largely in parallel.

Although scholars have increasingly examined the integration of religious values into educational settings, they have paid limited attention to how students understand the relevance of Islamic values in sport (Jani et al., 2025). Educators consequently have little evidence about how students interpret and apply Islamic principles in sports-related situations, largely because diagnostic instruments designed for this interdisciplinary domain remain scarce. To establish how far that gap extends, a structured search was carried out in Scopus, Web of Science, ERIC, and Google Scholar, together with the Indonesian national indexes SINTA and Garuda, combining the terms four-tier, multi-tier, diagnostic test, and misconception with Islamic education, religious education, moral education, character education, and sport or physical education, in English and Indonesian and without date restriction. The search was run on 15 January 2026 using the string ("four-tier" OR "multi-tier diagnostic test" OR "diagnostic assessment") AND ("Islamic education" OR "religious education" OR "moral education") AND ("sport education" OR "physical education") AND ("misconception" OR "alternative conception"). It identified 426 records; 87 duplicates were removed, 339 titles and abstracts were screened, 46 full texts were assessed, and 18 studies were retained as relevant to multi-tier diagnostic assessment in these fields. The retrieved four-tier applications were concentrated in chemistry, physics, biology, and environmental science, and none applied the format to Akhlaq (Islamic ethics), Fiqh (Islamic jurisprudence), Tawhid (Islamic theology), and Qur'anic Studies asthey bear on sports practice. This is reported as the outcome of a bounded search rather than as a claim of priority: relevant instruments may exist in local, non-indexed, or non-English outlets that the search did not reach and in reporting the complete scoring transformation so that the estimates can be recomputed from the archived response file.

This study addresses that gap by developing a four-tier assessment for diagnosing students' conceptual understanding of the relevance of Islamic values in sport. Each item comprises an answer tier, an answer-confidence tier, a reasoning tier, and a reason-confidence tier, an architecture that allows patterns of understanding, uncertainty, guessing, and misconception to be distinguished. Two research questions guide the work. First, how can a four-tier assessment of the relevance of Islamic values in sport be developed and psychometrically evaluated? Second, what profile of conceptual understanding does the instrument reveal in a sample of sports-education students? Correspondingly, the study aims to construct a 12-item instrument covering the four IRE domains as they apply to sports practice, to evaluate its measurement properties using partial least squares structural equation modeling (PLS-SEM) with 250 undergraduate students, and to report the resulting diagnostic profile. The novelty of the study lies in transferring a diagnostic format developed for science education into a value-laden socio-religious domain, and in reporting the complete scoring transformation so that every estimate can be recomputed from the archived response file.

2. MATERIALS AND METHODS

This study employed a research and development design. The framework guiding development was the four-tier test construction procedure of Kaltakci-Gurel et al. (2017), set out in section 2.2, which was adopted because it was formulated for this instrument type and specifies both item construction and diagnostic classification. No additional general R&D model, such as Borg and Gall, the 4-D model, or ADDIE, was adopted, because the Kaltakci-Gurel framework directly specifies the procedures required for developing and evaluating diagnostic instruments of this type. The product developed was a four-tier assessment examining students' conceptual understanding of the integration of Islamic Religious Education with sport. To make the process auditable, five stages are distinguished and reported separately below: content development,

comprising curriculum analysis, concept mapping, and blueprint construction; expert content validation by subject-matter specialists; pilot work, comprising the open-ended pre-test and the coding of alternative conceptions; item revision and prototype assembly; and field administration followed by psychometric evaluation. Formal content-validity evidence was therefore obtained and acted upon before the instrument was submitted to psychometric analysis. Because that psychometric evidence derives from one institution and one administration, the study reports the first development cycle rather than a completed validation, and the diagnostic profile obtained from the validation sample is presented on that understanding.

2.1 Participants and Data Collection

The target population comprised students enrolled in sports-related programs at Islamic higher education institutions. Two samples were drawn from the Sports Education Program of Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari. A development sample of 62 students completed the open-ended pre-test used to generate distractors. A separate validation sample of 250 students, with no overlap with the development sample, completed the final instrument. Because both samples were drawn from a single programme at a single institution and were selected purposively, the instrument is described throughout as locally developed and preliminarily evaluated. The distractors in particular reflect the terminology, teaching emphases, and doctrinal orientation of that setting, so they cannot be assumed to represent the alternative conceptions held by students elsewhere. Any broader claim about the instrument's properties awaits administration across multiple institutions.

Participants were selected purposively. Eligibility required active enrollment in the Sports Education Program, current or prior enrollment in the compulsory IRE curriculum covering the assessed domains, and completion of all four tiers of every item. Prior completion of all four domains was not required, so semester-2 students taking the relevant coursework at the time met the inclusion criterion. Students from other programs and incomplete submissions were excluded before the analytic dataset was constructed, which comprised 250 complete cases with no item-level missing values. Each case contributed 12 item-level observations, giving 3,000 item-level observations in total; because every item comprises four tiers, these correspond to 12,000 tier-level responses. All prevalence figures reported below are computed over the 3,000 item-level observations.

The validation sample was distributed across semester 2 (n = 81), semester 4 (n = 70), semester 6 (n = 54), and semester 8 (n = 45). Participants had a mean age of 20.4 years; 146 were women and 104 were men. Ethical approval was granted by the Research Ethics Committee of Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari (No. 164-KEP-UNISKA.PPJ-2026). Field administration took place in April and May 2026, all participants gave written informed consent, and responses were anonymized before analysis. Participation was voluntary and formed no part of any graded assessment: the instrument was administered outside course examinations, responses could not be linked to any student's academic record, and neither participation nor the content of a response had any bearing on course grades. Before administration, students were told in writing and orally that the instrument was diagnostic rather than evaluative, that responses departing from the taught position would carry no academic or personal consequence of any kind, and that they could decline or withdraw at any point without giving a reason.

2.2 Instrument Development Procedure

Development followed the four-tier construction framework of Kaltakci-Gurel et al. (2017) in five sequential phases, summarised in Figure 1, with expert content validation carried out before pre-testing. Phase 1 comprised qualitative content analysis of the Semester Learning Plans for IRE courses and concept mapping of the Outcome-Based Education curriculum, which yielded the test blueprint and the reference answer key. The blueprint, the draft scenarios, and the keyed answers and reasons were then submitted to the expert panel described below, and items were revised in light of that review. Phase 2 administered an open-ended pre-test to the development sample. In Phase 3 the written justifications were coded to identify recurring alternative conceptions, which then served as candidate distractors for the reasoning tier. Phase 4 assembled the prototype, comprising blueprint, administration guidelines, items, answer keys, response sheets, scoring rubric, and interpretation guidelines. Phase 5 administered the 12-item instrument to the validation sample.

The expert panel comprised specialists in Islamic Religious Education, in sports education, and in educational measurement. Five panelists (E1-E5) took part, each holding a doctorate: E1 in Islamic Religious Education, validating IRE substance; E2 in Fiqh and Islamic Studies, validating the religious concepts; E3 in Physical and Sport Education, validating the sporting contexts; E4 in Educational Evaluation, validating instrument construction; and E5 in Psychometrics and Educational Measurement, validating methodology. All five are affiliated with Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari. That the panel was drawn entirely from the host institution is a limitation: it secures alignment with the curriculum the instrument is calibrated to, but it does not provide independent scrutiny of that curriculum's doctrinal emphases, and an external panel is recommended for the next cycle. Each panellist reviewed the items from the standpoint of the role recorded above, so that the religious substance, the juristic content, the sporting context, the construction of the items, and the measurement methodology were each covered by a designated reviewer. Agreement across the

panel was quantified with Aiken's V. Across the five panelists, Aiken's V ranged from 0.88 to 0.95: AKH1 0.93, AKH2 0.91, AKH3 0.94, FIQ1 0.92, FIQ2 0.89, FIQ3 0.91, TAU1 0.95, TAU2 0.93, TAU3 0.92, PQA1 0.90, PQA2 0.88, and PQA3 0.91. Ten items were retained unchanged and the two lowest, PQA2 (V = 0.88) and FIQ2 (V = 0.89), underwent minor revision before field use. The keyed answers and keyed reasons therefore express the mainstream Sunni position taught in the compulsory IRE curriculum of the participating institution as endorsed by this panel, and not an adjudication between schools of law. The instrument was not designed to represent differences of juristic opinion between the schools of law and does not distinguish between them; where such differences bear on a scenario, the keyed response reflects the position taught locally. This is a substantive constraint on interpretation rather than a procedural detail: what the instrument measures is alignment with a specified, locally authoritative reference standard, and the diagnostic labels used below should be read in that sense throughout.

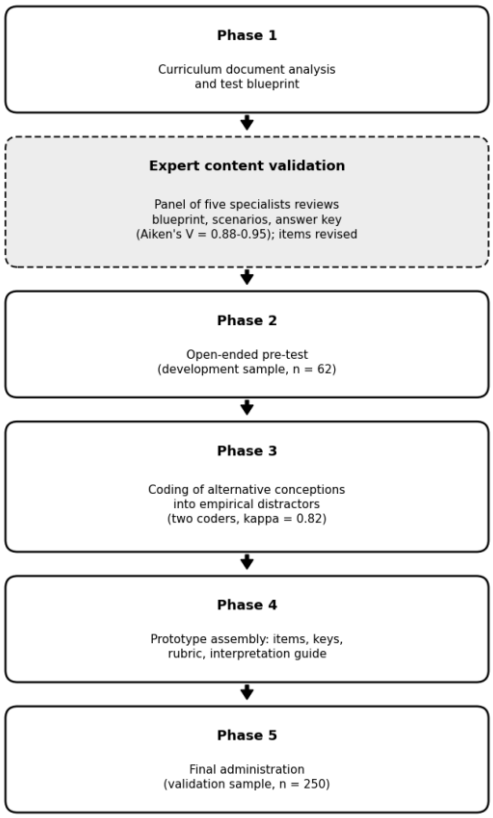

Figure 1. Five-phase development procedure

The open-ended pre-test responses were coded independently by two coders, who read the full set of written justifications and identified recurring alternative conceptions. Inter-coder agreement was quantified with Cohen's kappa, and disagreements the coders could not settle were referred to a third researcher whose decision was final. The coders agreed on 54 of the 62 coded responses and disagreed on 8, giving Cohen's kappa = 0.82, conventionally described as almost perfect agreement; three cases required adjudication by the third researcher. An alternative conception became eligible as a distractor only if it was expressed independently by at least five of the 62 respondents, approximately 8% of the development sample, and only if the expert panel judged it conceptually distinct from the keyed reason rather than a partial or clumsily worded version of it. This threshold of five responses corresponds to 8.06% of the development sample.

The blueprint links each of the four IRE domains to three applied sports concepts, giving 12 scenario-based items (Table 1). The three concepts per domain are not intended to represent those domains exhaustively. They were selected because they are the concepts that the Semester Learning Plans of the compulsory IRE courses explicitly connect to physical activity, sport, or the body, and because the expert panel, which included both Islamic Religious Education and sports-education specialists, judged them to be the applications students are most likely to meet in professional practice as sports educators. The panel reviewed and endorsed the blueprint before item writing began. Even so, three items sample only a small part of constructs as broad as Fiqh or Tawhid, so domain-level scores are evidence about the sampled concepts rather than a general measure of domain mastery, and expanding the pool to permit genuine domain-level inference is a priority for the next development cycle. All items were written at

cognitive level C3 (Apply) of the revised Bloom taxonomy, since the construct of interest is the application of religious principles to concrete sporting situations rather than the recall of doctrine.

Table 1. Test blueprint of the 12 four-tier items

Domain	Concept assessed in sports context	Item
Akhlaq	Adab and intention during sport	AKH1
Akhlaq	Honesty and fair play in competition	AKH2
Akhlaq	Emotion regulation and respectful competition	AKH3
Fiqh	Clothing and awrah in training and competition	FIQ1
Fiqh	Prayer obligations within training schedules	FIQ2
Fiqh	Ethics of male-female interaction in sport	FIQ3
Tawhid	Orientation of physical activity to Allah	TAU1
Tawhid	Maintaining health as gratitude and amanah	TAU2
Tawhid	Victory, defeat, tawakkul, and humility	TAU3
Qur'anic Studies	Balance and avoidance of excess in health	PQA1
Qur'anic Studies	Consistency, time, and responsibility	PQA2
Qur'anic Studies	Physical activity that promotes benefit	PQA3

Each final item comprises four tiers: an answer tier with one correct option and three distractors; an answer-confidence tier; a reasoning tier with one correct explanation and three distractors derived from the coded pre-test responses; and a reason-confidence tier. Confidence in both the answer and the reason was rated on a five-point Likert scale from 1 (not at all confident) to 5 (highly confident). Following the interpretation guide accompanying the instrument, ratings of 1 and 2 denote low confidence and ratings of 4 and 5 high confidence, while a rating of 3 forms an intermediate band assigned to neither pole. Assigning the midpoint to neither pole differs from certainty-of-response scales that allocate every rating to one of the two poles, and it is reported here because it materially affects the diagnostic classification. Two considerations motivated the five-point scale with an unassigned midpoint. First, the six-point index of Hasan and colleagues has no midpoint by construction and so obliges respondents toward one pole; the five-point format is the one students at the participating institution meet in routine coursework, and in a value-laden domain forcing a choice risks recording a commitment the respondent does not hold. Second, reading the midpoint as evidence of either confidence or doubt would attribute to it a meaning that has not been demonstrated, so it is held apart as a transition band rather than collapsed into either pole. Because this decision proves to have a large effect on the diagnostic profile, its consequences are examined directly in section 3.3.1, where the classification is recomputed with the midpoint assigned to the high band and to the low band in turn.

2.3 Scoring Transformation and Diagnostic Classification

Every administered item yields four fields: the accuracy of the selected answer, the answer-confidence rating, the accuracy of the selected reason, and the reason-confidence rating. Two distinct derivations were applied to these fields, and both are reported in full so that all results below can be reproduced from the archived response file.

For the measurement-model analysis, each item was reduced to a single ordinal indicator using the weighted composite in equation (1). Answer accuracy A and reason accuracy R are scored 1 for a keyed response and 0 otherwise, while the answer-confidence rating CA and the reason-confidence rating CR are normalised to the unit interval before weighting:

$$S = 4\,[\,w1\,A + w2\,R + w3\,(CA - 1)/4 + w4\,(CR - 1)/4\,] \quad (1)$$

The weights are w1 = w2 = 0.30 for the two accuracy components and w3 = w4 = 0.20 for the two confidence components, so that accuracy carries 60% of the item score and confidence 40%. These weights are researcher-specified rather than empirically derived, and the reasoning is as follows. Accuracy receives the larger share because the answer and reason tiers carry the substantive claim about what a student holds, whereas the confidence tiers qualify that claim; this ordering follows the logic of the four-tier literature, in which confidence is used to interpret accuracy rather than to stand in for it. The two accuracy components are weighted equally because the format treats reasoning as no less diagnostic than answer selection, and the two confidence components are weighted equally for the same reason. A 60/40 split rather than a larger accuracy share was chosen because the distinctive contribution of

the format is precisely the confidence information that conventional scoring discards; weighting accuracy much more heavily would reduce the composite toward a two-tier score. Comparable four-tier studies have adopted a range of conventions, from equal weighting of all tiers to accuracy-only scoring with confidence used solely for classification, and no consensus weighting exists, which is itself a reason to test the sensitivity of the results. The factor 4 rescales the weighted sum to the reported 0-4 range. Because every component is discrete, the resulting indicator takes one of exactly 21 values spaced 0.2 apart across the 0-4 interval; all 21 values occur in the observed data. The full scoring syntax is retained in the archived workbook as spreadsheet formulas rather than as stored constants, so the indicator matrix can be regenerated from the raw tier responses. Because this composite becomes the observed indicator for the measurement model, the stability of the psychometric conclusions under alternative weightings is examined in section 3.2.1, where equal weighting of 0.25 per component and accuracy-dominant weighting of 0.40/0.40/0.10/0.10 are applied to the same responses. The diagnostic classification reported in section 3.3 is derived independently of equation (1) and is therefore unaffected by the choice of weights.

Table 2. Decision rules for the five diagnostic categories

Category (code)	Decision rule
Target conception (TC)	Answer and reason both correct, both confidence ratings high
Correct but low-confidence response (CLC)	Answer and reason both correct, at least one confidence rating low
Genuine misconception (MC)	Exactly one tier incorrect, held at high confidence (R4a); or both tiers incorrect with at least one held at high confidence (R4b)
Lack of knowledge (LK)	Exactly one tier incorrect, held at low confidence (R5a); or both tiers incorrect with both held at low confidence (R5b)
Transition (TR)	Either confidence rating equals 3, the intermediate band (rule evaluated first)

Where both tiers are incorrect, the rule is that a high confidence rating on either tier is sufficient for classification as a genuine misconception, and lack of knowledge is recorded only when both incorrect tiers are held with low confidence; this asymmetry follows the same reasoning, since a confidently held error requires conceptual change wherever it occurs. The complete mapping is given as Supplementary Table S1, which enumerates all 36 patterns with the rule applied to each. Of the 36, twenty contain a midpoint rating and are assigned to the transition band, seven map to genuine misconception, five to lack of knowledge, three to a correct but low-confidence response, and one to a target conception. That a single pattern out of 36 yields a target conception, while twenty yield the transition band, is a structural property of the rule set rather than a finding about students, and it is a further reason to read the prevalence figures in section 3.3 comparatively rather than absolutely.

Note. The rules are applied in the order listed, and the transition rule takes precedence. Two category labels depart from those used in science-education applications of this format. The category usually termed scientific conception is here called target conception, because the criterion of correctness in this content area is the curricular reference standard described in section 2.2 rather than an empirically established scientific fact. The category usually termed correct guessing is here called correct but low-confidence response, because low confidence is consistent with guessing but does not by itself establish that guessing occurred. The codes MC, LK, and TR are unchanged.

2.4 Data Analysis

Measurement quality was examined with partial least squares structural equation modeling in SmartPLS 4 Pro, specifying the four IRE domains as first-order composites rather than as reflective common factors. The specification needs comment, because the choice of method is itself a substantive claim about the constructs. Akhlaq, Fiqh, Tawhid, and Qur'anic Studies are categories of a syllabus, not psychological traits, and there is no theoretical warrant for expecting a single latent disposition to generate a student's standing on every concept within one of them. PLS-SEM was preferred to covariance-based confirmatory factor analysis for exactly this reason: it estimates constructs as weighted composites of their indicators and does not require the common-factor assumption that a shared latent cause produces the observed covariation. Reporting follows established guidance (J. Hair & Alamer, 2022; J. F. Hair et al., 2019, 2022; Sarstedt et al., 2022). Outer loadings, Cronbach's alpha, composite reliability, average variance extracted, and heterotrait-monotrait ratios were examined; bootstrapping used 5,000 subsamples with significance evaluated at an alpha level of .05. These indices are reported as descriptive evidence about how tightly the three indicators of a domain cohere, not as confirmation of a reflective latent structure. Where they are low, the appropriate inference is that the domain is a heterogeneous collection of concepts, which is a

substantive finding about curricular categories rather than a measurement fault to be repaired by deleting indicators. Item response theory and Rasch modeling were considered and were not adopted at this stage. Both estimate item parameters from the response matrix, and applying them to an item bank with the structural defect in keyed positions reported in section 3.1 would yield parameters that could not be interpreted; they are recommended for the next development cycle in section 3.8, once the bank has been corrected. Diagnostic prevalence was computed as the proportion of the 3,000 item-level observations falling in each category, disaggregated by domain, by item, and by semester. Because curricular exposure varied across the sample, a supplementary analysis repeated the measurement-model estimation on participants in semesters 4 to 8 only. Two further robustness checks were specified: the measurement model was re-estimated under alternative scoring weights (section 3.2.1), and the diagnostic classification was recomputed under alternative treatments of the confidence midpoint (section 3.3.1).

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Descriptive Characteristics of the Responses

The final instrument comprised 12 four-tier items distributed evenly across the four domains, each represented by three items: Akhlaq (AKH1-AKH3), Fiqh (FIQ1-FIQ3), Tawhid (TAU1-TAU3), and Qur'anic Studies (PQA1-PQA3). Across the 3,000 item-level observations the mean composite score was 2.50 (SD = 1.36) on the 0-4 scale, and domain means were closely grouped: Qur'anic Studies 2.56, Tawhid 2.51, Akhlaq 2.50, and Fiqh 2.44.

Answer-tier accuracy ranged from 0.612 to 0.724 across items and reason-tier accuracy from 0.596 to 0.732, indicating that the items were of moderate difficulty and none approached a floor or ceiling. Confidence was distributed almost identically across the two confidence tiers: 50.0% of answer-confidence ratings and 49.5% of reason-confidence ratings fell in the high band, 30.9% and 31.8% respectively, in the low band, and 19.1% and 18.7% in the intermediate band. Mean confidence was 3.32 (SD = 1.42) for answers and 3.29 (SD = 1.41) for reasons.

A descriptive audit of the item bank identified one structural weakness, and it is the most serious limitation of the study. Keyed positions are strongly unbalanced: option B is correct in 11 of the 12 answer tiers and option A in 11 of the 12 reasoning tiers, and options C and D are never keyed. A regularity of this magnitude is detectable by test-takers, so a respondent who notices it can select the keyed option without engaging the content, and answer-tier accuracy may therefore partly index sensitivity to a formatting pattern rather than understanding. Because every indicator, diagnostic classification, reliability estimate, and measurement-model result reported below derives from this bank, the defect constrains all of them. Rebalancing the keys and re-administering the revised instrument is the strongest available response and is the first correction recommended in section 3.8. The analyses that follow are reported on the explicit understanding that they describe a prototype with a known structural flaw: they establish that the format yields interpretable diagnostic information in this content area, and they do not establish that this particular item bank is validated.

3.1.2 Measurement Model Evaluation

All four domain composites met the conventional criteria applied to composite models (Table 3). Composite reliability ranged from 0.754 to 0.777 and average variance extracted from 0.507 to 0.537, all above the 0.50 threshold, indicating that each composite accounts for more than half the variance of its indicators. Because the domains are specified as composites rather than as reflective factors, these figures are reported as descriptive indices of indicator coherence and should not be read as confirming that a latent trait underlies each domain.

Table 3. Internal consistency and convergent validity

Construct	alpha	CR	AVE
Akhlaq	0.520	0.757	0.509
Fiqh	0.569	0.777	0.537
Tawhid	0.533	0.762	0.516
Qur'anic Studies	0.510	0.754	0.507

Note. CR = composite reliability; AVE = average variance extracted.

Cronbach's alpha was nevertheless modest, ranging from 0.510 to 0.569, which corresponds to average inter-item correlations of approximately 0.27 for Akhlaq, 0.31 for Fiqh, 0.27 for Tawhid, and 0.26 for Qur'anic Studies. These values indicate weak covariance among items within the same domain, a finding interpreted in section 3.5. Read alongside the composite reliability and average variance extracted figures, they show why the summary phrase acceptable measurement properties would be misleading here: the indices that do not assume tau-equivalence are satisfactory, the index that does assume it is not, and the most parsimonious reading is that each domain is a set of related but distinct concepts rather than a coherent unidimensional construct. That reading is pursued in section 3.5 and is reflected in the wording used throughout this article.

Table 4. Heterotrait-monotrait ratios (n = 250)

Construct pair	HTMT
Akhlaq - Fiqh	0.362
Akhlaq - Qur'anic Studies	0.648
Akhlaq - Tawhid	0.449
Fiqh - Qur'anic Studies	0.595
Fiqh - Tawhid	0.352
Tawhid - Qur'anic Studies	0.320

Discriminant validity, assessed through the heterotrait-monotrait ratio, ranged from 0.320 to 0.648 (Table 4). All values fall well below the 0.85 threshold, indicating that the four domains are empirically distinguishable rather than redundant (Hair, Hult, et al., 2022; Roemer et al., 2021). The highest ratio, between Akhlaq and Qur'anic Studies, is consistent with the substantial conceptual overlap between Qur'anic ethical injunctions and Islamic character education

Bootstrapping with 5,000 subsamples confirmed that all 12 indicators loaded significantly on their intended constructs (Table 5).

Table 5. Outer loadings and bootstrap results

Indicator	Loading	t	p
AKH1	0.698	11.574	< .001
AKH2	0.694	11.466	< .001
AKH3	0.748	16.193	< .001
FIQ1	0.722	13.344	< .001
FIQ2	0.703	12.375	< .001
FIQ3	0.772	16.987	< .001
TAU1	0.702	8.976	< .001
TAU2	0.724	10.485	< .001
TAU3	0.729	11.274	< .001
PQA1	0.743	16.064	< .001
PQA2	0.640	9.651	< .001
PQA3	0.748	14.257	< .001

Outer loadings ranged from 0.640 to 0.772 and every t-statistic exceeded the critical value of 1.96. Statistical significance is a weak criterion at this sample size, however: with n = 250 even modest loadings are reliably distinguishable from zero, so significance alone is not a sufficient basis for retaining an indicator. Retention decisions in an assessment instrument should also rest on content coverage, item difficulty, discrimination, distractor functioning, and construct relevance. No indicator was removed here because each is the sole representative of a blueprinted concept and removing it would narrow content coverage, not because its loading reached significance. PQA2, whose loading of 0.640 is the lowest in the set, and PQA3 and FIQ2, which carry the highest rates of alternative conception in Table 7, are flagged for item-level review in the next development cycle. Figure 2 presents the estimated first-order measurement model.

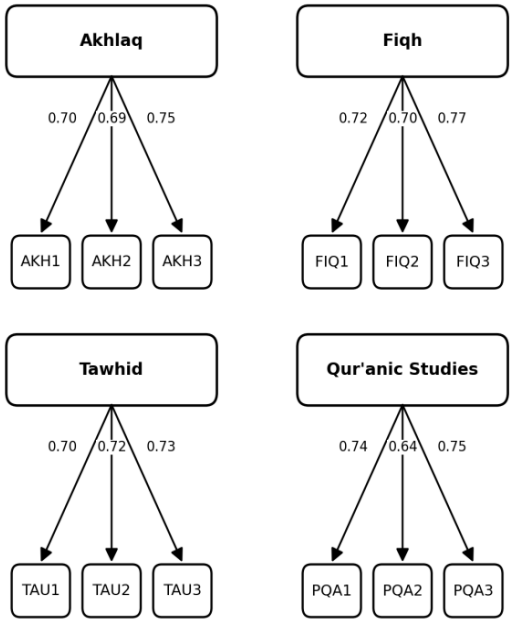

Figure 2. First-order measurement model

3.1.2.1 Robustness of the Measurement Model to the Scoring Weights

Because the composite defined in equation (1) is researcher-specified, the item indicators were reconstructed from the original tier-level response matrix under two alternative weightings, equal weighting in which each of the four components contributes 0.25, and accuracy-dominant weighting with 0.40 for each accuracy component and 0.10 for each confidence component. Reliability, convergent validity, discriminant validity, and indicator loadings were recalculated under each scheme; the complete values are reported in Supplementary Table S2.

The estimates are close to invariant. Across the three weightings Cronbach's alpha moved by at most 0.033 for any domain, composite reliability by at most 0.013, average variance extracted by at most 0.017, individual outer loadings by at most 0.029, and heterotrait-monotrait ratios by at most 0.025. Every loading remained between 0.658 and 0.766 and every heterotrait-monotrait ratio between 0.309 and 0.631, well below the 0.85 threshold, and the rank order of the domains on each index was unchanged. The measurement-model conclusions reported in section 3.2 are therefore properties of the response data rather than artefacts of the scoring formula, and the choice among these weightings can be regarded as presentational.

One qualification deserves statement. Under accuracy-dominant weighting the average variance extracted for Akhlaq falls to 0.497, marginally below the conventional 0.50 threshold, against 0.510 under the reported weighting and 0.514 under equal weighting. Akhlaq is the domain most sensitive to the weighting on every index, which is consistent with the within-domain heterogeneity discussed in section 3.5, and the margin by which it meets the convergent-validity convention is thin enough to be decided by a scoring choice. This does not overturn the finding of near-invariance, but it does mean that convergent validity for Akhlaq should not be described as securely established.

Two points concern the reconstruction itself. First, it was carried out independently of the SmartPLS estimation reported in Tables 3 to 5, and the two do not agree to the last decimal: under the reported weighting the reconstruction reproduces alpha, composite reliability, and average variance extracted to within 0.012 and heterotrait-monotrait ratios to within 0.030, but individual outer loadings differ by up to 0.053, and the within-domain ordering of loadings differs for Akhlaq and Tawhid. Discrepancies of this size are expected between implementations differing in standardisation and convergence settings, but they are large enough that the loadings in Table 5 should be read as approximate rather than exact, and the comparisons above are therefore made within the reconstruction, where the pipeline is held constant. Second, the diagnostic classification in section 3.3 does not use equation (1), so the diagnostic profile is unaffected by the weighting by construction; the sensitivity that matters for the profile is the treatment of the confidence midpoint, examined in section 3.3.1.

3.1.3 Diagnostic Profile of Students' Conceptions

Applying the rules in Table 2 to all 3,000 item-level observations produced the profile shown in Table 6. A target conception was the single most frequent substantive category at 25.1%, followed by correct but low-confidence responses at 15.8%, lack of knowledge at 13.2%, and genuine misconception at 12.6%. A further 33.3% fell in the transition band because at least one confidence rating occupied the intermediate value.

Table 6. Diagnostic category prevalence by domain (%)

Domain	TC	CLC	MC	LK	TR
Akhlaq	26.0	15.9	12.0	12.8	33.3
Fiqh	22.9	16.3	14.5	13.6	32.7
Tawhid	23.7	15.5	11.7	13.6	35.5
Qur'anic St.	27.9	15.6	12.0	12.7	31.9
All domains	25.1	15.8	12.6	13.2	33.3

Note. TC = target conception; CLC = correct but low-confidence response; MC = genuine misconception; LK = lack of knowledge; TR = transition. Category labels are defined in Table 2 and section 2.3. n = 750 item-level observations per domain, 3,000 overall.

Two features of this distribution deserve emphasis. First, correct but low-confidence responses account for 15.8% of observations, which means that approximately three of every eight fully correct answer-and-reason pairs were produced without confident endorsement. A conventional scoring key would have credited all of them as mastery, and this is the clearest empirical justification for the four-tier format in this content area. The category is deliberately not labeled correct guessing. Low confidence is consistent with guessing, but it is equally consistent with a student who understands the concept and remains hesitant, which is a plausible state in a normative domain where a respondent may know the taught ruling yet be unsure how it applies to the scenario as described; a general tendency to avoid the extremes of a rating scale may also contribute. Separating these possibilities would require think-aloud protocols or response-process evidence that the present design did not collect, so no inference about guessing behavior is drawn. Second, the transition band absorbs a third of all observations. That proportion follows directly from the decision to hold a confidence rating of 3 apart from both poles, and it is too large to be treated as a naturally occurring psychological state without evidence; section 3.3.1 therefore examines what happens when the midpoint is assigned differently.

Fiqh produced the lowest rate of target conception (22.9%) together with the highest rate of genuine misconception (14.5%), while Qur'anic Studies produced the highest rate of target conception (27.9%). This pattern is intelligible: Fiqh items require normative rulings to be applied to concrete situations such as awrah during competition or the timing of obligatory prayer within a training schedule, where students must weigh competing obligations rather than recognise a general principle.

Table 7. Item-level accuracy and diagnostic outcomes

Item	Answer	Reason	TC %	MC %
AKH1	0.696	0.712	27.6	12.0
AKH2	0.624	0.644	28.8	10.8
AKH3	0.648	0.612	21.6	13.2
FIQ1	0.624	0.668	24.4	14.8
FIQ2	0.620	0.616	20.8	16.0
FIQ3	0.668	0.652	23.6	12.8
TAU1	0.692	0.648	21.6	8.4
TAU2	0.612	0.640	24.0	14.8
TAU3	0.708	0.668	25.6	12.0
PQA1	0.708	0.732	27.6	8.0
PQA2	0.724	0.668	30.4	9.6
PQA3	0.612	0.596	25.6	18.4

Note. Answer and Reason columns give the proportion of correct responses in the respective tier.

The item-level profile in Table 7 localises these effects further. PQA3, concerning physical activity that promotes benefit, generated the highest rate of genuine misconception in the instrument at 18.4%, and FIQ2, concerning prayer obligations during training, the second highest at 16.0%. By contrast PQA1 and TAU1 generated rates of only 8.0% and 8.4%. These four items indicate where concept-specific remediation would be

most useful. Item-level rates must nonetheless be read with caution, because they are affected by the design defects described in section 3.1 as well as by student understanding. A distractor that is transparently undesirable will depress the apparent misconception rate for its item, while an item whose keyed position departs from the prevailing pattern will raise it, and the present data cannot separate these contributions. Differences of the size observed here may therefore reflect item construction as much as conceptual difficulty, and the ranking of items should be re-established after the bank has been revised.

3.1.3.1 Sensitivity of the Diagnostic Profile to the Confidence Midpoint

To establish how far the profile depends on the scoring rule rather than on the responses, the classification was recomputed twice on the same 3,000 item-level observations: once with a confidence rating of 3 assigned to the high band, and once with it assigned to the low band. Table 8 reports the outcome, and the dependence is substantial. Because the transition band is emptied under both reassignments, all four substantive categories necessarily gain, so the direction of movement carries no information and only the magnitudes are informative. Genuine misconception more than doubles, from 12.6% to 26.0%, when the midpoint is read as confidence, while correct but low-confidence responses more than double, from 15.8% to 34.9%, when it is read as doubt. Target conception ranges from 25.1% to 41.0% and is unchanged under the low-confidence reassignment, since that rule cannot create a pattern in which both tiers are correct and both confidence ratings high.

Table 8. Prevalence by 3 Midpoint Treatments (%)

Midpoint treatment	TC	CLC	MC	LK	TR
Held apart (reported)	25.1	15.8	12.6	13.2	33.3
Assigned to high band	41.0	19.1	26.0	14.0	0.0
Assigned to low band	25.1	34.9	15.3	24.6	0.0

Note. n = 3,000 item-level observations. TC = target conception; CLC = correct but low-confidence response; MC = genuine misconception; LK = lack of knowledge; TR = transition. The first row is the rule used throughout this article. Under both reassignments the transition band is empty by construction

3.1.4 Sensitivity to Curricular Exposure

Because prior completion of all four IRE domains was not required, the model was re-estimated on participants in semesters 4 to 8 (n = 169), who had covered more of the curriculum. Cronbach's alpha in this subgroup ranged from 0.467 to 0.569 and composite reliability from 0.739 to 0.776, compared with 0.510 to 0.569 and 0.754 to 0.777 in the full sample. The restricted estimates did not improve and were slightly lower overall, particularly for Qur'anic Studies.

The diagnostic profile was similarly stable across semester groups. The proportion of observations classified as a target conception was 24.7% in semester 2, 24.6% in semester 4, 25.3% in semester 6, and 26.5% in semester 8. Genuine misconception did not decline monotonically either, standing at 11.3%, 12.6%, 14.7%, and 12.2% respectively. Because semesters 2, 4, 6, and 8 are separate cohorts assessed on a single occasion rather than the same students followed over time, these figures cannot support any claim about the effect of curricular exposure: differences between the groups may reflect cohort composition, admission year, attrition, or intervening experience as readily as instruction received. What the data support is the narrower statement that the prevalence of alternative conceptions did not differ substantially across semester groups in this sample. Whether additional IRE instruction reduces alternative conceptions about the application of Islamic values to sport is a question this design cannot answer and would require a longitudinal or pre-post study, although the absence of a cross-sectional gradient is at least consistent with evidence that alternative conceptions resist conventional instruction (Soeharto & Csapó, 2022; Taber, 2019).

3.2 Discussion

The first research question asked how a four-tier assessment could be developed and evaluated for an interdisciplinary socio-religious context. Curriculum-based concept mapping, expert content validation, open-ended pre-testing, empirically grounded distractor development, prototype assembly, and composite-model evaluation together yielded a workable measurement structure: all 12 indicators related significantly to their intended domains, the composite-model indices met conventional thresholds, and the four domains proved empirically distinct. Whether that structure is also a valid one is a separate question, and the rest of this section addresses it.

The prior question is what kind of variable each domain represents. Akhlaq, Fiqh, Tawhid, and Qur'anic Studies are content domains of a curriculum rather than psychological traits, whereas a reflective measurement model assumes a latent variable generating covariation among its indicators. That assumption is defensible only if students who understand one concept within a domain reliably understand the others. The average inter-item correlations observed here, between 0.26 and 0.31, suggest it holds only weakly. A student may grasp the rules governing awrah during competition while holding an alternative conception about prayer timing during

training, although both belong to Fiqh. This is why the domains were specified as composites rather than as reflective factors in section 2.4, and why the reliability and convergent-validity indices are reported descriptively rather than as evidence that a latent trait was recovered. The modeling decision follows from the nature of the constructs; it is not a post-hoc gloss on a disappointing alpha.

The weak within-domain covariance admits at least three readings, and the present design cannot fully separate them. The first is substantive: understanding may be organized concept by concept rather than as a domain-general trait, which is what constructivist accounts of conceptual change would predict (Bryce & Blown, 2024; Taber, 2019). The item-level results in Table 7 support this reading, since misconception rates vary more than twofold between items within the same domain. The second is methodological: several distractors are transparent in their normative desirability, creating a risk that respondents choose the socially acceptable option without demonstrating nuanced understanding, and social-desirability bias is a recognized threat to validity in value-laden assessment (Teh et al., 2023). The third arises from the item bank itself: because keyed positions are strongly unbalanced, respondents who detect the regularity can answer correctly without engaging the content while those who do not cannot, and such heterogeneity in response strategy introduces construct-irrelevant variance that would itself depress inter-item covariance.

Modest internal-consistency coefficients are not unusual in four-tier instruments (Andriani et al., 2021), for instance, reported a Cronbach's alpha of 0.51 for a four-tier biology instrument whose item-level indices were otherwise satisfactory. With only three indicators per domain, alpha is additionally sensitive to small changes in inter-item covariance. The alpha values are therefore reported as a limitation rather than treated as evidence of strong internal consistency, and the composite reliability values, which do not assume tau-equivalence, are the more appropriate index for this model.

The position imbalance and the transparency of several distractors are the two design defects that most constrain what can be concluded. Both are set out in section 3.1 and their remedies in section 3.8, and they are not restated here. The point that matters for the present argument is that both plausibly account for part of the weak covariance observed, and that both are correctable by design rather than by collecting more data.

The second research question concerned the diagnostic profile. The four-tier format extends beyond conventional correct-or-incorrect scoring precisely because it combines answer selection, reasoning, and confidence within a single item, and the present results demonstrate what that additional information buys. Of the responses in which both the answer and the reason were correct, roughly two fifths were accompanied by low confidence and would therefore be misclassified as understanding by a conventional key. Conversely, the distinction between genuine misconception (12.6%) and lack of knowledge (13.2%) separates two groups of almost equal size that call for different pedagogical responses: the former requires conceptual change, the latter requires instruction. This is the diagnostic capability that (altakci-Gurel et al. (2017) and Yan and Subramaniam (2018) describe, realised here in a socio-religious domain in what the sarch reported in section 1 suggests is one of the first applications of its kind.

3.2.1 Implications

Theoretically, the findings support contextualising Islamic Religious Education within students' lived experience, including sport, consistent with Outcome-Based Education principles and with evidence that physical education supports character development when pedagogically connected to values-based education (Aygun et al., 2024; Shen et al., 2022). A narrower implication concerns measurement itself. Because a reflective model presupposes a latent variable generating covariation among its indicators, and because that presupposition held only weakly in all four domains, treating curricular categories as unitary psychological constructs should be regarded as a hypothesis requiring test rather than a modeling convention inherited from science-education applications of the format.

Practically, the diagnostic output identifies specific targets for teaching. The concentration of alternative conceptions in items concerning beneficial physical activity and prayer timing indicates that these two topics warrant direct conceptual-change instruction, while the substantial proportion of correct but low-confidence responses indicates that apparent mastery in class assessments should not be taken at face value. These are implications of the diagnostic approach, however, and not a recommendation to adopt the present instrument. Because the keyed positions and several distractors require correction, the instrument in its current form should not be used for institutional assessment, for grading, or for decisions about individual students; adoption should follow, not precede, the revision and re-piloting described in section 3.8. One operational point does carry over immediately to any group developing a comparable instrument: because the classification depends on all four tier-level fields, administration systems should preserve those fields rather than storing only a total score.

3.2.2 Research Contribution

This study contributes to the assessment literature in four respects. First, it extends the four-tier diagnostic format beyond its established base in science education to an interdisciplinary socio-religious domain, providing what is, to the authors' knowledge, the first instrument of this type spanning Akhlaq, Fiqh, Tawhid, and Qur'anic Studies as they apply to sports practice. The

domains proved empirically distinguishable, with heterotrait-monotrait ratios well below the 0.85 threshold, indicating that the distinction between them is not merely administrative.

Second, it demonstrates that the construction sequence specified by Kaltakci-Gurel et al. (2017) is feasible in a value-laden content area, subject to adaptations that its science-education applications do not require. Curriculum blueprinting, open-ended elicitation of alternative conceptions, empirically grounded distractor development, prototype assembly, and field administration all proved workable where the criterion for a correct reason is normative as well as empirical, a condition that does not arise in the chemistry and physics applications from which the format originates. The sequence did not transfer unchanged, however. It required an explicit and independently endorsed reference standard for what counts as a correct reason, a safeguard against socially desirable responding that empirical content does not provoke to the same degree, and an interpretation of the diagnostic categories in terms of alignment with a taught position rather than with established fact. The weaknesses of the present prototype, in key balance, distractor transparency, internal consistency, and latent structure, mark precisely where those adaptations were insufficient, and they are reported here as part of the contribution rather than as incidental faults.

Third, it reports the scoring transformation and the classification rules in full, together with the archived tier-level response file, so that the estimates in this article can be recomputed from the raw tier responses. An independent reconstruction reported in section 3.2.1 recovers the reliability, convergent-validity, and discriminant-validity indices closely but reproduces individual outer loadings only to within 0.053, so full numerical replication of Table 5 would require the estimation settings as well as the data.

Fourth, and most substantively, the study identifies a measurement-theoretic problem that science-education applications do not encounter in the same form. Where domains are defined by syllabus structure rather than by conceptual dependency, the reflective measurement model that multi-tier studies conventionally assume may not hold. The weak within-domain covariance reported here is therefore not simply a psychometric shortcoming of this instrument; it is evidence bearing on how curricular domains should be modelled, and it would have been obscured had the domains been evaluated only against conventional reliability thresholds.

3.2.3 Limitations

Several limitations qualify the conclusions of this study. First, both samples were drawn from a single institution through purposive sampling, leaving generalizability to other Islamic higher education institutions, curricula, and national contexts untested. Second, each domain was represented by only three indicators, which yields imprecise internal-consistency estimates. Third, the cross-sectional design provides no test-retest evidence of temporal stability, and the semester comparison reflects differences between cohorts rather than within individual students over time. Fourth, confidence ratings were self-reported and may not accurately map onto the strength of the underlying conception. Fifth, because the instrument was calibrated to one institution's teaching of a mainstream Sunni position, the keys would require re-examination before application in settings where different juristic positions are taught.

The most consequential limitation lies in the item bank itself. Correct-answer positions were strongly unbalanced, with option B keyed in 11 of 12 answer tiers and option A in 11 of 12 reasoning tiers, which may introduce position cues. Additionally, several distractors were comparatively transparent in their normative desirability, creating a plausible risk of socially desirable responding (Teh et al., 2023). Because all indicators derive from this bank, these features constrain the interpretation of every estimate reported here. Consequently, while the evidence supports the claim that the prototype is diagnostically informative and that the construction sequence is feasible, it does not support the claim that the instrument currently possesses acceptable measurement properties. The present study is best understood as the first phase of instrument development, identifying both promising features and critical design problems that must be resolved prior to validation or practical adoption.

3.2.4 Suggestions

To address these constraints, four main priorities are recommended for future research and practice, the first two of which are correctable by design prior to further data collection. First, keyed option positions should be rebalanced so that each of the four positions is correct in approximately one-quarter of the answer and reasoning tiers, followed by a re-piloting of the reordered items. Second, distractors whose normative undesirability makes them easy to eliminate should be revised through cognitive interviewing with students and open-ended pre-testing, ensuring that each distractor reflects an alternative conception that students actually hold. (A corrected version incorporating these two design revisions is currently in preparation and will be evaluated in a subsequent study).

Third, the revised instrument should be administered to larger, multi-institution samples and evaluated using Item Response Theory (IRT) or Rasch methods, which model responses at the item level without presupposing a domain-general latent structure (Andriani et al., 2021; Aripiani et al., 2023; Isra & Mufit, 2023). Future research should also incorporate test-retest administration to establish temporal stability and examine differential item functioning across sex and year of study. Fourth, researchers should formally test—rather than assume—whether a syllabus category behaves as a reflective latent variable; where within-domain covariance proves weak, formative specifications or item-level modeling will likely represent the data more faithfully. Finally, qualitative cognitive interviewing should accompany the next administration to determine whether the transition band

corresponds to a recognizable subjective state and to evaluate whether respondents detect regularities in option placement.

4. CONCLUSION

This study developed a 12-item four-tier assessment of the relevance of Islamic Religious Education values to sports practice and evaluated it with 250 sports-education students. Answering the first research question, the construction sequence proved workable in this domain and produced an instrument whose four domains are empirically distinguishable and whose indicators all relate significantly to their intended domain. The indices that do not assume tauhid equivalence were satisfactory, while internal consistency within domains was low, which is consistent with understanding organized concept by concept rather than as a domain-general trait. Taken together, this establishes preliminary feasibility; it does not establish that the instrument is validated.

Answering the second research question, the diagnostic profile separated states that conventional scoring cannot distinguish: responses that were correct and confidently held, responses that were correct but held with low confidence, confidently held alternative conceptions, and simple absence of knowledge. Fiqh yielded the weakest profile and Qur'anic Studies the strongest, and items concerning beneficial physical activity and prayer timing during training carried the highest rates of alternative conception. Prevalence did not differ substantially between semester groups; because those groups are separate cohorts assessed on one occasion, this says nothing about whether curricular exposure reduces alternative conceptions, and the present design cannot address that question. All prevalence figures are conditional on the treatment of the confidence midpoint, as section 3.3.1 shows.

The study is therefore best read as the first phase of instrument development. Its central message is twofold. A four-tier diagnostic can be constructed for this content area and recovers information that conventional scoring discards, since a substantial share of the fully correct responses in this sample were not confidently held. And two design corrections must be made before the instrument is used or described as validated: the keyed option positions must be rebalanced, and the more transparent distractors must be revised through cognitive interviewing and a fresh round of open-ended pre-testing. The revised instrument should then be re-piloted on a multi-institution sample and analysed with item response theory or Rasch methods. Until that work is complete, the appropriate claim is one of demonstrated feasibility and identified design problems, not of established psychometric quality.

5. ACKNOWLEDGEMENT

The authors thank the students of the Sports Education Program at Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari who took part in the open-ended pre-test and in the final administration of the instrument. We are grateful to the Faculty of Teacher Training and Education and to the Faculty of Islamic Studies for facilitating access to the Semester Learning Plans and to the participating classes, and to the Research Ethics Committee of Universitas Islam Kalimantan Muhammad Arsyad Al-Banjari for its review of the study protocol. The research received no specific grant from any funding agency, and the authors declare no competing interests.

6. AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally.

AUTHOR INFORMATION

Corresponding Authors

Ari Tri Fitrianto, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Indonesia

https://orcid.org/0000-0002-0825-2128

Email: aritri.fitrianto@uniska-bjm.ac.id

Authors

Aulia Putri Yuliasti, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Indonesia

https://orcid.org/0009-0003-3105-8678

Email: auliayuliasti09@gmail.com

Abdul Hafiz, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Indonesia

https://orcid.org/0000-0002-4850-1272

Email: abdulhafiz@uniska-bjm.ac.id

Muhammad Irwaysh Abdhi, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Indonesia

https://orcid.org/0009-0005-8908-294X

Email: muhammadirwansyah.2023@student.uny.ac.id

REFFERENCE

Aksoy, A. C. A., & Erten, S. (2022). A four-tier diagnostic test to determine pre-service science teachers' misconceptions about global warming. *Journal of Baltic Science Education*, *21*(5), 747-761. https://doi.org/10.33225/jbse/22.21.747

Al-Lawama, M., & Kumwenda, B. (2023). Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function. *BMC Medical Education*, *23*(1), 212. https://doi.org/10.1186/s12909-023-04206-3

Andriani, F., Indrowati, M., & Sugiharto, B. (2021). Analysis items of the four-tier immune system multiple choice test instrument using Rasch model. *Biosfer: Jurnal Pendidikan Biologi*, *14*(1), 99-119. https://doi.org/10.21009/biosferjpb.18020

Aripiani, S. K., Samsudin, A., Kaniawati, I., Novia, H., Aminudin, A. H., Sutrisno, A. D., & Costu, B. (2023). Diagnostic instruments of four-tier test work and energy (FORTUNE) to identify the level of students' conceptions. *Tadris: Jurnal Keguruan Dan Ilmu Tarbiyah*, *8*(1), 19-32. https://doi.org/10.24042/tadris.v8i1.13524

Astuti, I. A. D., Bhakti, Y. B., Prasetya, R., & Zulherman, Z. (2023). Android-based 4-tier physics test app to identify student misconception profiles. *International Journal of Evaluation and Research in Education*, *12*(3), 1356-1363. https://doi.org/10.11591/ijere.v12i3.25536

Atika, Y., Rahmaniar, Y., Muslim, M., Wanto, D., & Daheri, M. (2023). Formative Test Analysis of Islamic Religious Education Learning Evaluation Practices (Study at SMP Negeri 1 Muara Kemumu). *Indonesian Journal of Pedagogy and Teacher Education*, *1*(2), 38–42. https://doi.org/10.58723/ijopate.v1i2.113

Aygun, Y., Boke, H., Yagin, F. H., Tufekci, S., Murathan, T., Gencay, E., Prieto-González, P., & Ardigò, L. P. (2024). Emotional and Social Outcomes of the Teaching Personal and Social Responsibility Model in Physical Education: A Systematic Review and Meta-Analysis. *Children (Basel, Switzerland)*, *11*(4). https://doi.org/10.3390/children11040459

Barnard, J. J. (2025). A solution to guessing in multiple-choice questions: A new paradigm for (medical) assessment. *International Journal of Health and Pharmaceutical Research*, *10*(12), 1-13. https://doi.org/10.56201/ijhpr.vol.10.no12.2025.pg1.13

Bisa, M. (2023). Sports education as a means of building students' character: Values and benefits. *Al-Ishlah: Jurnal Pendidikan*, *15*(2), 1581-1590. https://doi.org/10.35445/alishlah.v15i2.3889

Bronikowska, M., Mouratidou, K., Ludwiczak, M., Karamavrou, S., McKeel, C., & Bronikowski, M. (2024). Systematic review on social/moral competence interventions in physical education. *Biomedical Human Kinetics*, *16*(1), 55-77. https://doi.org/10.2478/bhk-2024-0007

Bryce, T. G. K., & Blown, E. J. (2024). Ausubel's meaningful learning re-visited. *Current Psychology*, *43*(5), 4579–4598. https://doi.org/10.1007/s12144-023-04440-4

Dai, J., Chen, J., Huang, Z., Chen, Y., Li, Y., Sun, J., Li, D., Lu, M., Chen, J., & Tilga, H. (2024). Age-effects of sport education model on basic psychological needs and intrinsic motivation of adolescent students: A systematic review and meta-analysis. *PLOS ONE*, *19*(5), e0297878. https://doi.org/10.1371/journal.pone.0297878

Desstya, A., Sayekti, I. C., Abduh, M., & Sukartono, S. (2025). Development of a four-tier diagnostic test for misconceptions in natural science of primary school pupils. *Journal of Turkish Science Education*, *22*(2), 338-353. https://doi.org/10.36681/tused.2025.017

Do, H.-N., Do, B. N., & Nguyen, M. H. (2023). 3How do constructivism learning environments generate better motivation and learning strategies? The Design Science Approach. *Heliyon*, *9*(12). https://doi.org/10.1016/j.heliyon.2023.e22862

Duran, T., & Dikmenli, M. (2024). Use of Multi-Tier Concept Diagnostic Tests in Biology Education: A Systematic Review of the Literature. *Journal of Education in Science, Environment and Health*, *10*(4), 224–244. https://doi.org/10.55549/jeseh.755

Ginanjar, M. H., Rahman, Hidayat, R., & Halim, A. A. (2025). The implementation of Islamic religious education in efforts to shape Islamic character and develop students' talents and interests. *Edukasi Islami: Jurnal Pendidikan Islam*, *14*(2), 281–292. https://doi.org/10.30868/ei.v14i01.7732

Habiddin, H., & Page, E. M. (2019). Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK). *Indonesian Journal of Chemistry*, *19*(3), 720-736. https://doi.org/10.22146/ijc.39218

Hair, J., & Alamer, A. (2022). Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. *Research Methods in Applied Linguistics*, *1*(3), 1–16. https://doi.org/10.1016/j.rmal.2022.100027

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM) (3rd ed.)*. Sage. https://uk.sagepub.com/en-gb/eur/a-primer-on-partial-least-squares-structural-equation-modeling-pls-sem/book270548

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. In *European Business Review* (Vol. 31, Issue 1, pp. 2–24). Emerald Group Publishing Ltd. https://doi.org/10.1108/EBR-11-2018-0203

Hull, M. M., Jansky, A., & Hopf, M. (2022). Does Confidence in a Wrong Answer Imply a Misconception? *Physical Review Physics Education Research*, *18*(2), 20108.

https://doi.org/10.1103/PhysRevPhysEducRes.18.020108

Imaduddin, M., Shofa, M. M., Riza, M. F., & Fikri, A. A. (2023). Exploring the Pre-Service Basic Science Teachers' Misconceptions Using the Six-Tier Diagnostic Test. *International Journal of Evaluation and Research in Education*, *12*(3), 1626–1636. https://doi.org/10.11591/ijere.v12i3.24603

Isra, R. A., & Mufit, F. (2023). Students' Conceptual Understanding and Causes of Misconceptions on {Newton}'s Law. *International Journal of Evaluation and Research in Education*, *12*(4), 1914–1924. https://doi.org/10.11591/ijere.v12i4.25568

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The development of a four-tier diagnostic test based on modern test theory in physics education. *European Journal of Educational Research*, *12*(1), 371-385. https://doi.org/10.12973/eu-jer.12.1.371

Jadwiszczak, M., Wawrzyniak, S., & Pezdek, K. (2025). More than movement: A systematic review of moral and social development in adolescents' physical education. *BMC Public Health*, *25*(1), 2076. https://doi.org/10.1186/s12889-025-23169-2

Jani, E., Yono, T., & Nurhasyim, M. (2025). Sports learning methods based on Islamic values. *Gladi : Jurnal Ilmu Keolahragaan*, *16*(5), 269–274. https://doi.org/10.21009/GJIK.162.11

Judijanto, L., Siminto, S., & Rahman, R. (2024). The Influence of Religious Beliefs and Religious Practices on Social Cohesion in Modern Society in Indonesia. *The Eastasouth Journal of Social Science and Humanities*, *1*(6), 139 – 150. https://doi.org/10.58812/esssh.v1i03.276

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education*, *35*(2), 238-260. https://doi.org/10.1080/02635143.2017.1310094

Manninen, M., & Campbell, S. (2022). The effect of the Sport Education Model on basic needs, intrinsic motivation and prosocial attitudes: A systematic review and multilevel meta-analysis. *European Physical Education Review*, *28*(1), 78-99. https://doi.org/10.1177/1356336X211017938

Putra, A. T. A., & Anhusadar, L. (2025). Survey of assessment techniques for religious and moral values in early childhood in Baruga Sub-district, Kendari City. *EduBasic Journal: Jurnal Pendidikan Dasar*, *7*(1), 30-45. https://doi.org/10.17509/ebj.v7i1.82857

Rachmawati, D., Anjania, D., Lubis, M., Ali, A., Oktavianti, C., Wati, Y., Sabrina, L., Noviyanti, E., & Nihayah, A. (2024). The Role of Islamic Religious Education in Shaping the Character of Children in Banyuurip Village in the Digital Era. *Prosiding Seminar Nasional Pendidikan Dan Agama*, *5*(5), 48–59. https://doi.org/10.55606/semnaspa.v5i1.2035

Rahayu, N. Y., Maulida, T., Lestari, A., & Surawan, S. (2026). Fostering Empathy and Positive Character in Elementary School Students Through the School Environment. *Indonesian Journal of Pedagogy and Teacher Education*, *4*(2), 85–91. https://doi.org/10.58723/ijopate.v4i2.661

Sarstedt, M., Hair, J. F., Pick, M., Liengaard, B. D., Radomir, L., & Ringle, C. M. (2022). Progress in partial least squares structural equation modeling use in marketing research in the last decade. *Psychology & Marketing*, *39*(5), 1035-1064. https://doi.org/10.1002/mar.21640

Shen, Y., Martinek, T., & Dyson, B. P. (2022). Navigating the Processes and Products of The Teaching Personal and Social Responsibility Model: A Systematic Literature Review. *Quest*, *74*(1), 91–107. https://doi.org/10.1080/00336297.2021.2017988

Soeharto, S., & Csapó, B. (2022). Exploring Indonesian student misconceptions in science concepts. *Heliyon*, *8*(9), e10720. https://doi.org/10.1016/j.heliyon.2022.e10720

Taber, K. (2019). Constructivism in Education: Interpretations and Criticisms from Science Education. In *In Early childhood development: Concepts, methodologies, tools, and applications* (pp. 312–342). IGI Global. https://doi.org/10.4018/978-1-4666-9634-1.ch006

Teh, W. L., Abdin, E., Asharani, P. V, Kumar, F. D., Roystonn, K., Wang, P., Shafie, S., Chang, S., Jeyagurunathan, A., Vaingankar, J. A., Sum, C. F., Lee, E. S., Dam, R. M., & Subramaniam, M. (2023). Measuring social desirability bias in a multi-ethnic cohort sample: Its relationship with self-reported physical activity, dietary habits, and factor structure. *BMC Public Health*, *23*(1), 415. https://doi.org/10.1186/s12889-023-15309-3

Yan, Y. K., & Subramaniam, R. (2018). Using a Multi-Tier Diagnostic Test to Explore the Nature of Students' Alternative Conceptions on Reaction Kinetics. *Chemistry Education Research and Practice*, *19*(1), 213–226. https://doi.org/10.1039/C7RP00143F

Zhang, J., Soh, K. G., Bai, X., Anuar, M. A., & Xiao, W.

(2024). Optimizing learning outcomes in physical education: A comprehensive systematic review of hybrid pedagogical models integrated with the Sport Education Model. *PLOS ONE*, *19*(12), e0311957. https://doi.org/10.1371/journal.pone.0311957

Zhang, J., Xiao, W., Soh, K. G., Yao, G., Anuar, M. A. B., Bai, X., & Bao, L. (2024). The effect of the Sport Education Model in physical education on student learning attitude: A systematic review. *BMC Public Health*, *24*(1), 949. https://doi.org/10.1186/s12889-024-18243-0