Toward a Validated School Head Qualifying Examination:A PRISMA-Guided Review and PrototypeFor Philippine Public Schools

Roy M. Maldecino[1], Las Johansen B. Caluza[2]

[1]Sto. Niño Elementary School, Department of Education, Tacloban City, Philippines

[2]College of Arts and Sciences, Leyte Normal University, Tacloban City, Philippines

[*]Corresponding Authors

DOI: https://doi.org/10.51584/IJRIAS.2026.11110014

Received: 06 November 2026; Accepted: 11 November 2026; Published: 21 November 2026

ABSTRACT

A defensible qualifying examination for school heads must do more than produce reliable scores. It must show that candidates can demonstrate leadership practice, support comparable interpretations across settings, and justify the consequences of classification. This exploratory evidence map, completed by one reviewer, examined fixed Scopus and Google Scholar exports and used the findings to frame a Philippine research prototype. The source files contained 6,250 rows. Exact normalized-title deduplication removed 78 records, leaving 6,172 unique titles for screening under a documented rule set. Ninety-five records were mapped. To strengthen the interpretation, 12 linked scholarly full texts and four authoritative documents were purposefully verified. The historical Google Scholar query, export date, filters, and 500-result retrieval limit were recoverable, but the original Scopus query and database-run date were not preserved. Of the mapped records, only 12 directly addressed principal qualification or readiness. Binary signals in indexed text, used only as reporting indicators rather than estimates of method prevalence, showed factor-analysis terms in 27 records, self-report or survey terms in 23, predictive or criterion terms in two, invariance or differential-item-functioning terms in two, and performance-oriented terms in one. No signal was found for standard setting, longitudinal validation, or test–retest evidence. The literature helps inform construct and method design, but it does not support an integrated Philippine qualification decision. In response, this paper proposes a five-domain, PPSSH-aligned multimethod architecture, a traceability matrix linking claims, tasks, scoring, and validation, and a staged validation program. All weights and item counts remain provisional. The prototype is neither the current National Assessment for School Heads nor a developed or validated examination. It should not support pass/fail, appointment, promotion, or exclusion decisions unless policy authorization is secured and evidence establishes score meaning, fairness, classification consistency, prediction, and consequences.

Keywords: school heads, principal qualification, qualifying examination, Philippine Professional Standards for School Heads, PRISMA;

INTRODUCTION

Selecting school heads is a high-consequence assessment decision. A qualifying score can shape a candidate's career, affect a school's access to leadership, and influence public confidence in the appointment process. Reliability matters, but it cannot carry the decision by itself. The intended interpretation is that a candidate is ready to assume the responsibilities of a Philippine public-school head. Supporting that interpretation requires evidence on content representation, candidate response processes, score structure, relationships with later practice, fairness across relevant groups, and the consequences of score use. This view is consistent with the Standards for Educational and Psychological Testing, which define validity through the evidence and reasoning that support intended score interpretations and uses (American Educational Research Association et al., 2014).

The Philippine Professional Standards for School Heads (PPSSH) offer a policy-based starting point for defining what the assessment should cover. DepEd Order No. 24, s. 2020 sets out five domains: Leading Strategically; Managing School Operations and Resources; Focusing on Teaching and Learning; Developing Self and Others; and Building Connections. The standards guide recruitment, selection, placement, performance appraisal, and professional development (Department of Education, 2020). Terminology in national policy has also changed over time. Earlier memoranda refer to the National Qualifying Examination for School Heads (NQESH), while DepEd Memorandum No. 69, s. 2025 administers the National Assessment for School Heads (NASH). DepEd later described NASH as a PPSSH-based assessment intended to gauge aspiring school heads' readiness, identify competency gaps, and inform targeted professional development (Department of Education, 2025, 2026). This policy context is essential, but it does not validate the prototype presented in this paper.

International research provides many instruments related to school leadership, although their coverage remains fragmented. Recent studies examine digital leadership (Aydin & Ceylan, 2026), moral leadership (Arjanto et al., 2025), teacher-leadership competencies among aspiring principals (Darwis et al., 2025), innovation leadership (Fuad et al., 2022), technological leadership (Durnalı, 2022), principal emotion (Chen, 2021), leadership training (Yuliana et al., 2022), and broader principal competencies (Kadri et al., 2021). Taken together, these studies illuminate parts of the role rather than the whole. None of the mapped records demonstrates an integrated Philippine qualifying examination whose classifications have been tested for fairness, decision consistency, longitudinal prediction, and consequences.

Method choice creates another challenge. Questionnaires can measure perceptions and dispositions efficiently, while a qualifying examination must also sample applied judgment under conditions that resemble the role. Studies from related areas show why the distinction matters. Direct reviews of teacher-generated test items did not always agree with teachers' self-reported competence (Kissi et al., 2023; Quansah et al., 2025). By comparison, a theory-driven situational judgment test can place respondents in professional scenarios and ask them to make decisions rather than simply describe themselves (Aldrup et al., 2020). These findings do not establish one best format for school-head selection. They show that the method used shapes what can reasonably be inferred from a score.

This paper has two related aims. It first provides a transparent, PRISMA-informed synthesis of the supplied Scopus and Google Scholar records, supported by targeted verification of linked full texts. It then turns the observed evidence gaps into an evidence-centered prototype and a prospective validation program. Because development and validation have not yet been conducted, the paper presents no field results. Calibration, standard setting, and predictive validation remain tasks for future research. Being explicit about that boundary is necessary for an ethical and publishable account of a proposed high-stakes assessment.

Objectives

General Objective

To synthesize the evidence on school-head and leadership assessment and use the identified construct, method, fairness, and decision gaps to propose a Philippine PPSSH-aligned prototype qualifying examination and a complete, testable validation framework.

Specific Objectives

1. Map the volume, topical composition, and evidence completeness of the supplied Scopus and Google Scholar exports using a transparent PRISMA-informed record flow.
2. Identify what principal-focused instruments measure, how they were developed, and which validity evidence and limitations are visible in the exported records and verified full texts.
3. Determine whether the mapped literature supports an integrated, high-stakes qualification interpretation for Philippine school heads.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

4. Translate the five PPSSH domains into provisional readiness claims, assessment components, score weights, and performance evidence for a research prototype.
5. Specify the qualitative, psychometric, fairness, standard-setting, predictive, consequential, and governance studies required before operational use.
6. Formulate research questions that can falsify or refine the proposed construct, method mix, score interpretation, and qualifying decision.
7. Identify a future research agenda for localization, multilingual comparability, longitudinal prediction, security, and periodic revalidation.

Research Questions

Against this background, the study addresses the following research questions:

1. Which PPSSH-aligned knowledge, judgment, and evidence-based practices distinguish a minimally qualified school-head candidate from a developing candidate?
2. What content structure and combination of selected-response, situational-judgment, and constructed-response tasks best represent the intended readiness construct?
3. Does the proposed five-domain structure fit the field data, and is a total score defensible beyond domain profiles?
4. Are scores sufficiently precise and comparable across regions, school levels, locale and resource conditions, language versions, and relevant candidate groups?
5. What criterion-referenced performance standards support defensible qualification decisions, and how consistent are classifications across forms and replications?
6. How well do prototype scores predict later leadership practice and school-process outcomes after baseline context and prior experience are accounted for?
7. Does the multimethod composite add reliable and incremental evidence beyond a selected-response test alone?
8. What intended and unintended consequences arise for candidates, subgroups, schools, and professional-development systems?

METHODOLOGY

Review Design and Reporting Boundary

The review used a single-reviewer exploratory evidence map informed by PRISMA 2020 (Page et al., 2021). The term “PRISMA-informed” is used deliberately. The source material consisted of fixed exports rather than a prospectively registered search; the original Scopus query and database-run date were unavailable; Google Scholar retrieval stopped at 500 ranked records; one reviewer handled rule-assisted screening and coding; and full texts could not be obtained for all 95 mapped records. No independent second review or estimate of coding agreement was available. The claims are therefore limited to patterns visible in the supplied indexed records and the purposefully verified full-text core. This work is not presented as a complete, dual-screened systematic review.

Data Sources and Search Provenance

The supplied data consisted of 5,750 Scopus records and 500 Google Scholar records. The Google Scholar export retained a query timestamp of 19 September 2026 at 12:56 (timezone not stored), ranked positions 1–500, a 2020–2026 year restriction, an English interface, and the following captured query: “"test development"|"instrument development"|"test construction"|"assessment development" "validation"|"validity"|"content validity"|"construct validity"|"factor analysis"|"exploratory factor analysis"|"confirmatory factor analysis" "reliability"|"internal consistency"|"cronbach alpha"|"item analysis"|"item response theory"|"rasch model"”. In this syntax, vertical bars represent OR alternatives, while spaces separate the three concept groups. The export did not retain a recency-sort flag, so the records were treated as the first 500 positions under Google Scholar’s default ranking. The Scopus file contained 22

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

bibliographic fields, but its original query and database-run or export timestamp could not be recovered. The supplied workbook also does not allow language, document-type, or subject-area filters to be reconstructed. These details are reported as unavailable rather than inferred.

Table 1. Search Sources, Export Characteristics, and Retrieval Limitations

Source	Rows	Search/export record	Fields and filters recoverable	Retrieval rule / limitation
Scopus	5,750	Exact historical query and database-run/export date unavailable.	Authors; author IDs; title; year; source; volume/issue/pages; citations; DOI/link; abstract; keywords; document/publication/open-access fields; EID.	All supplied rows analyzed. Any platform export cap or upstream filtering is unknown.
Google Scholar	500	19 Sep 2026, 12:56; timezone not stored. Exact query reproduced in text and Supplement S1.	Default indexed search; years 2020–2026; English interface. Export includes title, authors, source, abstract/snippet, URLs, DOI, ranks, and citation fields.	First 500 ranked results only; no preserved date-sort parameter. Ranking and truncation bias remain.

Eligibility and Screening

Table 2. Eligibility Criteria and Evidence Boundaries

Dimension	Included	Excluded or Bounded
Population/setting	School heads, principals, aspiring leaders, teachers or educators where evidence transferred to leadership assessment, and school-governance contexts	Learner-only assessment without a transferable instrument-development contribution; health/clinical populations unrelated to educational leadership
Phenomenon	Qualification, readiness, leadership competencies, governance, assessment development, validation, or decision use	Topics outside education leadership or measurement
Evidence	Empirical development/validation studies, methodological papers, or directly relevant policy/standards	Records without sufficient topical relevance after title and indexed-text review
Language/source	Records represented in the supplied exports; full texts used when accessible through supplied links	No exclusion solely because a full text was inaccessible; such evidence was labeled partial

Records were deduplicated with a deterministic key derived from each title. Titles were normalized for whitespace, transformed with Unicode NFKD, lowercased, stripped of diacritics, converted to ASCII, and reduced to alphanumeric characters. Records with the same resulting key were grouped. A preferred row was then selected in this order: Scopus presence, longer abstract, and DOI presence. Database provenance, citation counts, DOI, and any available full-text URL were merged into that record. Rows without titles received no

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

key. The process removed 78 rows, comprising 76 cross-database overlaps and two duplicates within a source. Because the procedure did not use fuzzy title or DOI matching, some near duplicates may have been missed and rare false merges remain possible.

Eligibility decisions were rule-assisted and based mainly on the normalized title. Authors, year, source, and any indexed abstract or snippet were used to clarify borderline terminology. Categories were applied in a hierarchy: principal qualification/readiness; school or instructional leadership; teacher readiness/competency; school organization/governance; and psychometric/validation methodology. Health or clinical records, learner-only studies, and other material outside the scope were excluded. When a title was ambiguous, the complete exported metadata was checked and the record was resolved conservatively against the category definitions below. No claim was inferred from inaccessible full text. Supplementary File S1 records the decision and exclusion reason for every unique record.

Table 3. Evidence-Mapping Category Codebook and Classification Boundaries

Code	Category	Inclusion rule	Positive example	Exclusion boundary
PQR	Principal qualification/readiness	School heads/principals plus assessment, readiness, competency, selection, or qualifying construct.	Principal digital-leadership scale.	Principal-component analysis or clinical "principal investigator."
SIL	School/instructional leadership	Educational leadership measure or leadership–practice/outcome study in a school setting.	Instructional-leadership scale.	Business, nursing, or learner leadership without school-head relevance.
TRC	Teacher readiness/competency	Teacher professional readiness, competency, or performance measure with transferable assessment design.	Teacher SJT or test-construction competency.	Student readiness or health-related teacher screening.
SOG	School organization/governance	School governance, budgeting, resources, stakeholder participation, or organizational process with assessment relevance.	Teacher involvement in school budgeting.	Non-education governance.
PVM	Psychometric/validation methods	Instrument development, validation, translation, factor analysis, Rasch/IRT, or score-use methodology.	Cross-cultural scale-validation framework.	Purely substantive study without measurement contribution.

Targeted Full-Text Verification

Full-text review covered twelve linked scholarly works and examined their design, sample, item-development procedures, analyses, findings, and limitations. The set comprised four principal-focused instrument studies (Durnalı, 2022; Fuad et al., 2022; Kadri et al., 2021; Yuliana et al., 2022), two direct studies of assessment competence and work products (Kissi et al., 2023; Quansah et al., 2025), one situational judgment study (Aldrup et al., 2020), one school-budgeting instrument (Hadi et al., 2022), and four papers on measurement development or cross-cultural guidance (Fenn et al., 2020; Kalkbrenner, 2021; Stefana et al., 2025; Zhao et al., 2024). Claims in this manuscript that depend on full text are confined to this verified set. Other studies are discussed only to the extent supported by their exported abstract or publisher abstract page.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

Synthesis and Prototype Development

Because the studies varied widely in constructs, populations, methods, and outcomes, and because many records lacked complete numerical results, the synthesis was narrative and configurative rather than meta-analytic. Descriptive counts summarized the composition of the corpus and the validation signals stated in the indexed text. A framework synthesis then compared the mapped evidence with six requirements for interpreting a high-stakes score: content, response process, internal structure and reliability, relations to other variables, fairness and comparability, and consequences or decision quality. The proposed blueprint draws from the five PPSSH domains rather than from frequency counts in the international literature. Formats were chosen for the evidence they can elicit: selected-response tasks for applied knowledge, situational-judgment tasks for contextual prioritization, and constructed-response cases for integrated reasoning and communication.

Quality Assurance and Limitations

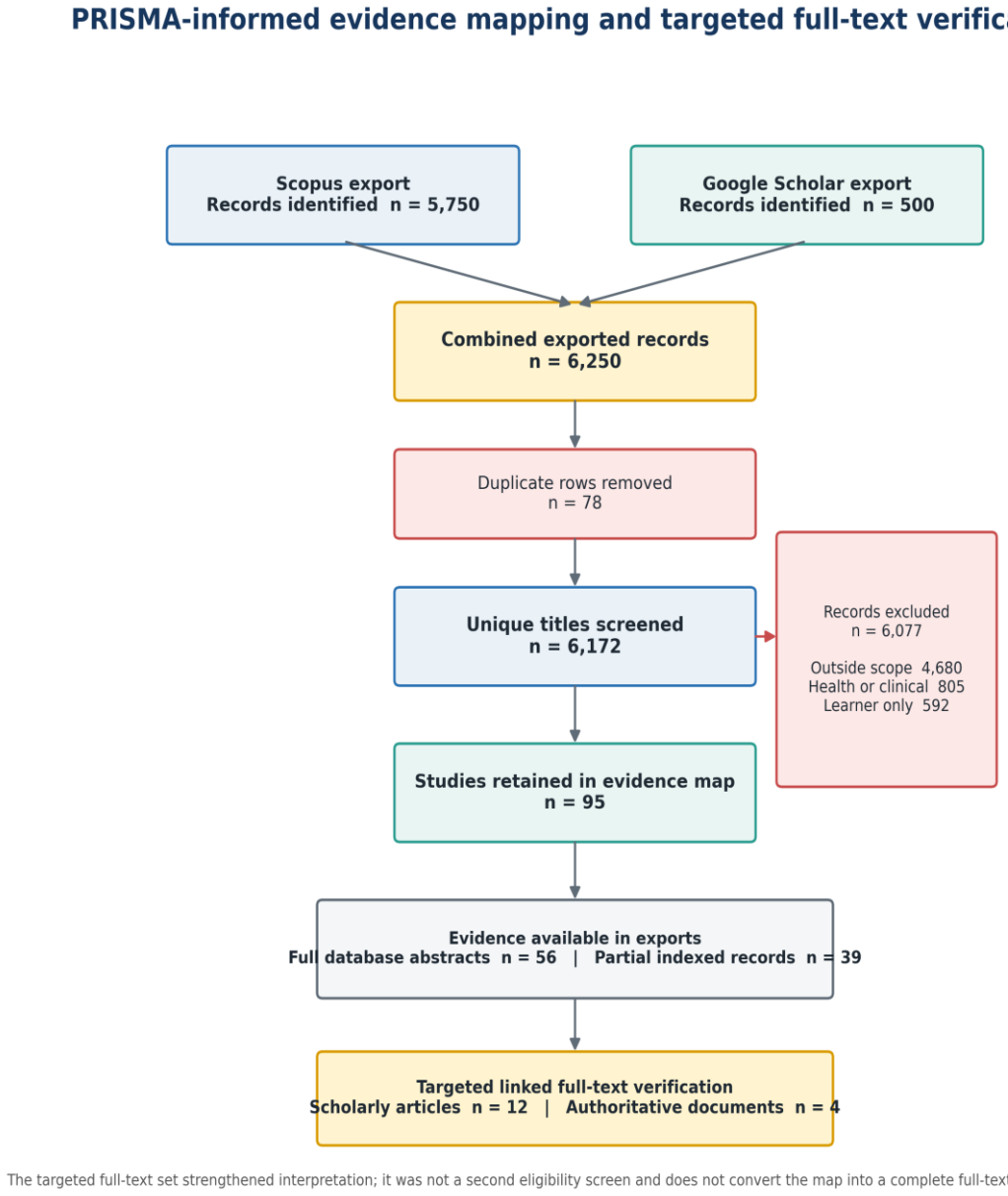
Each record was classified as linked-full-text verified, full database abstract, partial indexed record, policy-supported, or proposed. When an effect size, sample statistic, or limitation was not reported, none was inferred. One reviewer completed the coding with deterministic rules and maintained an auditable screening log. An independent second review was not feasible for this revision, so no inter-rater agreement coefficient is reported. Ambiguous records were checked against the complete exported metadata and retained only when they met a category rule. The resulting design supports an exploratory evidence map; it does not support prevalence estimates, a formal risk-of-bias synthesis, or causal inference. Supplementary File S1 provides the reproducibility materials, including the 95 mapped records, the 6,172-record screening log, query provenance, the codebook, evidence-completeness codes, method signals, and source basis.

RESULTS

Study Selection and Evidence Base

Together, the two exports contained 6,250 rows: 5,750 from Scopus and 500 from Google Scholar. Title normalization removed 78 duplicate rows and left 6,172 unique titles for screening. Of these, 6,077 were excluded (98.5%) and 95 were retained in the evidence map. The retained set included 56 full database abstracts and 39 partial indexed records. Figure 1 distinguishes the formal mapping flow from the later, targeted verification of full texts.

Figure 1. PRISMA-Informed Record Flow and Targeted Full-Text Verification Boundary



Source: Authors' analysis of the supplied Scopus and Google Scholar exports. Counts describe the fixed source files.

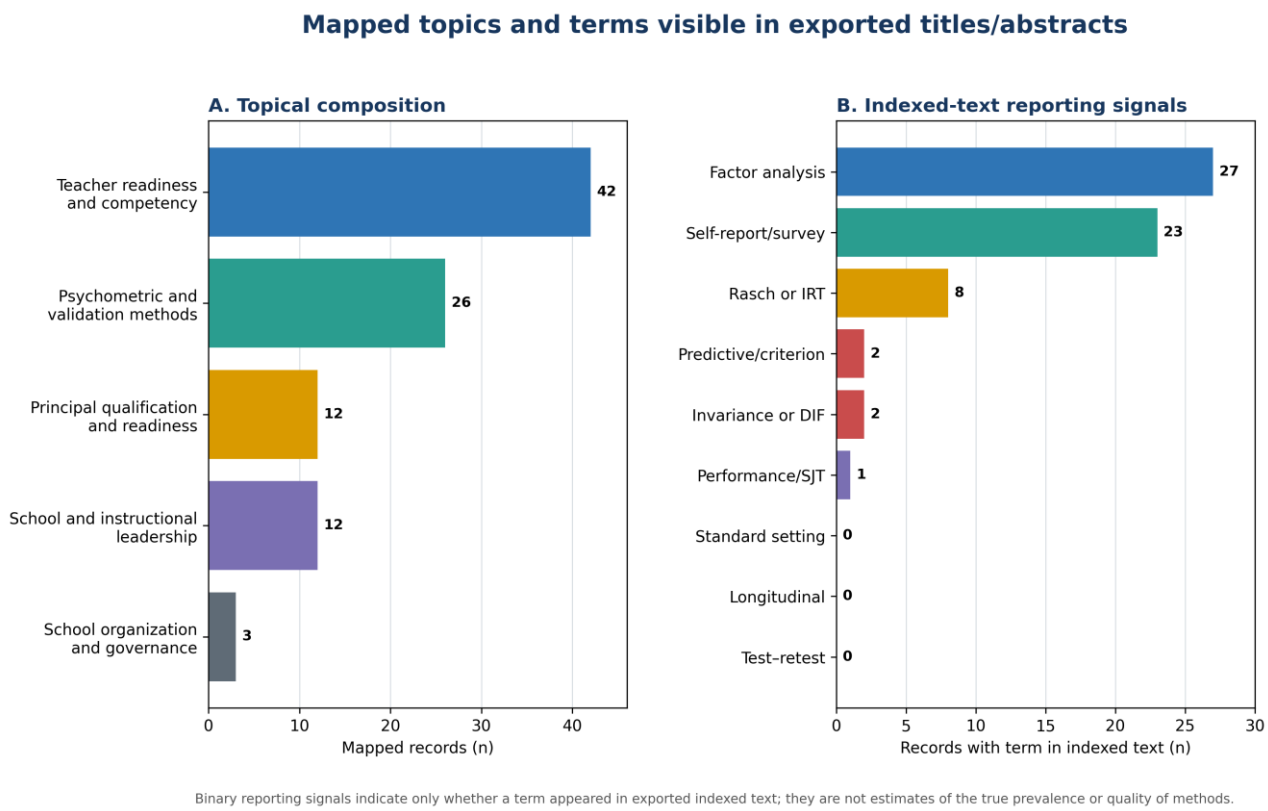
The large number of exclusions reflects the breadth of the source exports. Records outside the topic accounted for 4,680 exclusions, health or clinical records for 805, and learner-only records for 592. The resulting map of 95 studies is therefore a relevant subset of the supplied files, not a census of principal-assessment research worldwide. Study-level conclusions are further limited by the 41.1% share of partial records. The map is best used to identify patterns and missing evidence, which can then be checked against the smaller verified full-text core.

Topical Distribution and Validation-Evidence Imbalance

Teacher readiness and competency was the largest category, with 42 studies, followed by psychometric and validation methodology with 26. Principal qualification/readiness and school/instructional leadership contributed 12 studies each, while school organization/governance contributed three. Only 12.6% of the mapped corpus, therefore, dealt directly with principal qualification or readiness. Studies in the broader categories can inform method choices, but they cannot replace evidence that an assessment reflects the work and decision context of Philippine school heads.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

Figure 2. Topical Composition and Binary Indexed-Text Reporting Signals



Source: Authors' coding of the supplied exports (n = 95). Each count is a binary indexed-text reporting signal, presence of a specified term in the exported title or abstract/snippet, not a verified method, quality rating, or estimate of the true prevalence of methods in the literature.

These counts are useful as signals of what the indexed text reports, but they should not be interpreted as estimates of method prevalence. Factor-analysis terms appeared in 27 records, and self-report, questionnaire, or survey terms appeared in 23. Rasch or IRT terms appeared in eight records; predictive or criterion terms in two; invariance or DIF terms in two; and performance, observation, portfolio, multi-source, or SJT terms in one under the prespecified pattern. No standard-setting, longitudinal, or test–retest term appeared in the exported indexed text. A zero means only that the term was not named in the indexed title or abstract field, not that the method was never used. Full-text methods may be undercounted, and the presence of a term does not show that the method was carried out well. Supplementary File S1 contains the codebook, regular expressions, examples, and record-level binary flags.

Construct and Transfer Boundaries

For this map, principal qualification/readiness referred to evidence directly relevant to whether an aspiring or incumbent school head meets an entry or role-readiness standard. Studies of school or instructional leadership were kept separate because they describe leadership constructs, practices, or outcomes without necessarily supporting a qualifying classification. This distinction prevents a reliable leadership scale from being treated as proof that a candidate is minimally qualified for a high-stakes public role.

Teacher-readiness studies were included only as sources of transferable methods. They show how professional competencies may be operationalized, how self-report can differ from produced work, how situational-judgment or performance evidence can be elicited, and how items and scores may be validated. Their teacher-specific constructs and score interpretations do not transfer directly to school heads. Any claim about school-head readiness must be established anew through PPSSH-linked job analysis, studies of candidate responses, local calibration, fairness analysis, standard setting, prediction, and monitoring of consequences.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

What Principal-Focused Studies Contribute and Do Not Establish

Table 4. Principal-Focused Assessment Studies and Qualification-Use Boundaries

Study	Construct	Evidence Basis	Reported Contribution	Boundary for Qualification Use
Durnalı (2022)	Technological leadership behavior	Full text	EFA with 308 teachers; CFA with an independent 240; 18 items/four factors	One Ankara district; teacher ratings; no cut score, invariance, or prediction
Fuad et al. (2022)	Principal innovation leadership	Full text	Six experts; CVI and modified kappa; 58-item pool reduced to 54 across eight dimensions	Face/content evidence only; authors position the instrument for further validation
Yuliana et al. (2022)	Leadership-training indicators	Full text	255 teachers, 57 principals, 315 students; factor extraction and reliability evidence	Reported "CFA" relies on PCA/Varimax procedures; no decision or outcome evidence
Kadri et al. (2021)	Principal/teacher leadership competencies	Full text	52 items; EFA and internal consistency; principal dimensions include personal/social, managing, leading, transformative	Urban ordinary secondary schools in one Malaysian state; restricted generalizability
Chen (2021)	Principal emotions	Publisher abstract	Mixed-method program; 25-item, five-factor inventory reported	Full-text appraisal unavailable; emotion is one component, not a qualifying construct
Arjanto et al. (2025)	Moral leadership	Database abstract	Two experts; readability n=100; validation n=362; 65 to 52 items; hierarchical CFA	Sample-dependent percentile norms do not define minimum competence
Darwis et al. (2025)	Teacher-leadership competencies of aspiring principals	Database abstract	Expert review and CFA with 174 teachers; three dimensions; strong reported fit and reliability	Authors note contextual variability and long-term applicability; no Philippine evidence
Aydin & Ceylan (2026)	Digital leadership in an AI-driven era	Database abstract	EFA n=302 principals and independent CFA n=227; three-factor structure; high internal consistency	Does not establish readiness across the full school-head role or qualification decisions

Taken together, the studies cover separate pieces of school-head readiness. Technological and digital leadership instruments make transformation and support behaviors visible (Aydin & Ceylan, 2026; Durnalı, 2022). Innovation instruments examine how principals may enable new practices (Fuad et al., 2022), while moral leadership instruments address ethical conduct (Arjanto et al., 2025). Measures of emotion and well-being consider the sustainability of leadership (Chen, 2021), and tools for aspiring principals emphasize learning vision, culture, and environment (Darwis et al., 2025). Each area contributes useful evidence, but none represents the full role. Combining all of the scales would not solve the problem: it would give too much

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

weight to self-report, introduce uncontrolled overlap among constructs, and still leave the meaning of a qualifying cut score unsettled.

Some studies also offer development practices worth carrying into the prototype. Durnalı (2022) reduced a large initial item set through theory, expert review, exploratory analysis, and confirmation with an independent sample. Darwis et al. (2025) followed a theory-based model of instrument development and reported evidence of model fit and reliability. Arjanto et al. (2025) tested readability before moving to a larger validation sample. Even so, most end products are rating scales completed either by teachers describing principals or by respondents describing themselves. Such scales can indicate a construct, but they do not directly show whether a candidate can interpret policy, weigh competing interests, diagnose instruction, allocate scarce resources, or justify an action that is ethically and legally defensible.

Lessons from the Verified Full-Text Core

Table 5. Verified Full-Text Sources and Design Implications for the Prototype

Source	Design Contribution	Implication for the Prototype	Transfer Boundary
Aldrup et al. (2020)	Theory-driven SJT	Professional scenarios can sample knowledge about emotion regulation and relationship management.	Additional validity evidence is still needed; transport to principal selection requires new construct work.
Kalkbrenner (2021)	MEASURE development sequence	Begin with purpose, empirical/theoretical frameworks, blueprint, item pool, expert review, sampling, and validity/reliability evaluation.	Guidance is generic; high-stakes fairness, standards, and consequences must be added explicitly.
Fenn et al. (2020)	Test-development workflow	Pilot items; study reliability including temporal stability; examine content, construct, and criterion evidence; document use in a manual.	Psychological-test guidance requires adaptation to mixed-format occupational assessment.
Kissi et al. (2023)	Direct work-product assessment	Observed item-construction quality can diverge from self-reported competence.	Teacher assessment literacy is not principal readiness, but the method discrepancy is transferable.
Quansah et al. (2025)	Survey + document analysis + item analysis	Triangulation exposed weak actual items despite reported competence; CFA, content analysis, IRT, and rank association were combined.	One national and occupational setting; not a leadership selection study.
Hadi et al. (2022)	Budgeting-involvement scale	Resource stewardship can be decomposed and evaluated through content evidence and internal structure.	Perceptual scale evidence does not demonstrate performance in a budget simulation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

Source	Design Contribution	Implication for the Prototype	Transfer Boundary
Stefana et al. (2025)	Best-practice synthesis	Independent samples, test–retest evidence, criterion/predictive validation, and measurement invariance strengthen score interpretation.	Guidelines do not determine the local construct or acceptable decision risk.
Zhao et al. (2024)	Cross-cultural 10-step framework	Target-population interviews, collaborative translation, cognitive interviewing, independent EFA/CFA, invariance and DIF are integral to adaptation.	The scoping review found uneven uptake; techniques still require context-specific implementation.

The full-text studies suggest three practical lessons for development. First, the work begins before anyone writes an item. Kalkbrenner (2021) and Fenn et al. (2020) place purpose, theory, blueprinting, expert review, piloting, and several forms of validity evidence in a deliberate sequence. Following that sequence helps prevent the examination from becoming a set of plausible questions with no defensible construct behind them. Second, evidence should be triangulated because self-report and demonstrated work can diverge (Kissi et al., 2023; Quansah et al., 2025). Third, localization must be studied rather than assumed. Zhao et al. (2024) show that translation, cognitive interviewing, and cross-group measurement analyses address different risks; a single forward translation or a high coefficient alpha cannot replace them.

The verified full texts also caution against several common interpretations. High internal consistency does not demonstrate one-dimensionality, fairness, predictive validity, or suitability for pass/fail decisions. Likewise, a satisfactory factor model does not identify the level of competence that should count as minimally qualified. Percentile norms show where candidates stand relative to a sample, but they do not define what a school head must be able to do. Teacher ratings of principals may provide useful evidence of leadership behavior, although they are shaped by rater perspective, school climate, and opportunity to observe. For a high-stakes prototype, each score must therefore be tied to a specific claim, with task, rater, form, and contextual sources of error modeled explicitly.

Philippine Construct Translation

Because the PPSSH defines the national role in policy, it provides a stronger starting point than an imported overseas scale. The proposed examination retains all five domains but treats them as provisional, correlated claims about readiness rather than assuming that policy headings automatically function as psychometric dimensions. The prototype is conceptually related to the historical NQESH and the current NASH because each concerns aspiring school heads and PPSSH-based readiness. It is not a replacement form, revision, or validation of NASH/NQESH. It has no DepEd authorization and should not be presented as an operational examination.

Intended Use, Candidate Population, and Decision Boundary

The research population consists of aspiring Philippine public-school heads who meet the eligibility rules in effect when a study is conducted. Sampling should represent variation in region, school level, locale, resources, language background, prior leadership opportunity, and accessibility need. The first intended use is developmental: to examine PPSSH-aligned readiness claims and provide feedback at the domain level. Prototype v0.1 cannot justify pass/fail, eligibility, appointment, promotion, reclassification, or exclusion decisions. Any future qualifying use would need formal DepEd authorization, independently reviewed

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

technical evidence, a defensible performance standard, secure administration, due process, an appeals mechanism, and continuing monitoring of impact.

Table 6. Intended Use, Candidate Population, Decision Boundaries, and Safeguards

Use boundary	Candidate population	Intended decision	Consequences / safeguards
Research/developmental use	Aspiring school-head candidates meeting contemporaneous DepEd eligibility rules.	Estimate domain evidence; study tasks, scoring, fairness, and feedback.	Development plan or research participation only; no employment consequence.
Possible future qualifying use	Only the population and conditions covered by a validated interpretation and policy authorization.	Support a criterion-referenced readiness decision alongside specified safeguards/evidence.	Potential access to eligibility/appointment pathways; requires error, appeal, retest, and impact controls.
Prohibited at prototype stage	Any applicant or incumbent.	Pass/fail, ranking, appointment, promotion, exclusion, or replacement of NASH/NQESH.	Unacceptable risk of misclassification, inequitable opportunity, and policy overreach.

Table 7. Provisional PPSSH Domain Weights, Readiness Evidence, and Assessment Methods

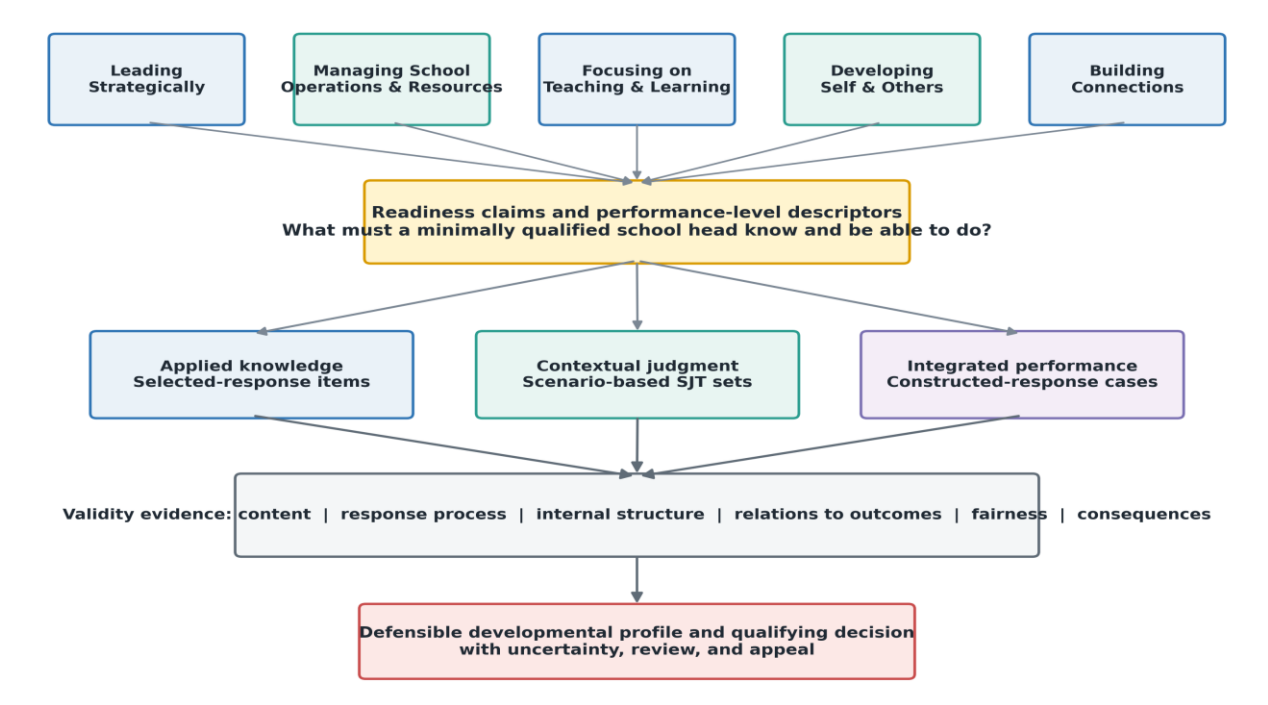
PPSSH Domain	Proposed Score Weight	Illustrative Readiness Evidence	Methods
Leading Strategically	20%	Policy interpretation; improvement priorities; ethical and evidence-informed direction	Selected response, SJT, constructed case
Managing School Operations and Resources	20%	Budget, personnel, facilities, safety, records, risk, and transparent stewardship	Selected response, SJT, constructed case
Focusing on Teaching and Learning	25%	Curriculum, assessment, instructional supervision, inclusion, and use of learning evidence	Selected response, SJT, constructed case
Developing Self and Others	20%	Professional learning, coaching, distributed leadership, well-being, and reflective practice	Selected response, SJT, constructed case
Building Connections	15%	Family/community engagement, partnerships, conflict, participation, and public accountability	Selected response, SJT, constructed case
Total	100%	Weights are provisional and must be tested through job analysis, information modeling, fairness review, and standard setting.	No operational use at prototype stage

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

Tasks that cross domains require careful handling. An emergency school-closure scenario, for example, may involve strategy, operations, communication, and continuity of teaching. If the same task were coded under every domain, dependencies would be inflated and the resulting scores would be harder to interpret. The blueprint should identify one primary claim and only a small number of secondary claims, then score no more than the response format can support. Ethical and equity considerations should run across the domains rather than appear in a small, separate section.

Proposed Examination Architecture

Figure 3. Evidence-Centered Architecture for the Proposed PPSSH-Aligned Examination



Source: Proposed synthesis. Domain weights, formats, scoring, and decision rules remain hypotheses until validated.

PPSSH Domain-to-Evidence Traceability

Table 8 shows the chain of reasoning behind the prototype. Its tasks and descriptors are examples of claims to be tested, not finished operational specifications. Before item production is finalized, the meaning of "minimally qualified" must be refined through job analysis and standard-setting panels.

Table 8. PPSSH Domain-to-Evidence Traceability Matrix

PPSSH domain	Observable tasks	Provisional minimally qualified descriptor	Formats / coverage	Scoring criteria	Proposed validation evidence
Leading Strategically	Interpret school data and policy; select and justify an improvement priority.	Identifies the material problem, weighs evidence and risks, and proposes a feasible, equitable, policy-aligned sequence.	12 knowledge items; 3 SJT sets; evidence in both cases; 20%.	Accuracy; evidence use; prioritization; feasibility; ethical/policy alignment.	Job analysis/content validity; cognitive interviews; item information/fit; DIF; relation to school-improvement-plan quality.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

PPSSH domain	Observable tasks	Provisional minimally qualified descriptor	Formats / coverage	Scoring criteria	Proposed validation evidence
Managing School Operations and Resources	Resolve budget, staffing, facility, safety, and continuity constraints.	Allocates transparently and equitably, recognizes compliance/risk, and explains resource trade-offs.	12 knowledge items; 3 SJT sets; Case 1 primary; 20%.	Compliance; prioritization; resource trade-off; risk control; transparency.	Content validity; rater/generalizability studies; DIF by resource context; relation to stewardship indicators.
Focusing on Teaching and Learning	Diagnose learning evidence; plan supervision, coaching, and monitoring.	Identifies the instructional problem, chooses evidence-based support, and specifies inclusive monitoring.	15 knowledge items; 4 SJT sets; both cases; 25%.	Diagnostic accuracy; instructional validity; monitoring; inclusion; use of evidence.	Response-process evidence; CFA/IRT; school-level invariance; relation to supervision and learning-process outcomes.
Developing Self and Others	Give feedback; design coaching/PD; manage professional learning and well-being.	Provides constructive, ethical feedback and builds distributed, sustained professional learning.	12 knowledge items; 3 SJT sets; Case 2 primary; 20%.	Feedback quality; development logic; relational judgment; ethics; follow-through.	Rater agreement; SJT keying; DIF by experience; relation to teacher-development processes.
Building Connections	Address stakeholder conflict; design participation, partnership, and accountability.	Communicates inclusively, enables legitimate participation, resolves conflict, and maintains public accountability.	9 knowledge items; 2 SJT sets; Case 2 secondary; 15%.	Stakeholder mapping; communication; conflict resolution; cultural responsiveness; accountability.	Language/cultural review; stakeholder content validity; regional/language DIF; relation to participation indicators.

Table 9. Provisional Examination Components, Weights, Claims, and Required Analyses

Component	Prototype v0.1 Field Form	Score Weight	Primary Claim	Required Analysis
A. Applied knowledge	60 selected-response items in the initial field form	50%	Policy application, data interpretation, resource decisions, instructional leadership principles	CTT plus IRT/Rasch; distractor functioning; local dependence; form information
B. Contextual judgment	15 scenario sets with calibrated response options and rationale-based scoring where feasible	30%	Prioritization, ethical judgment, stakeholder balance, escalation, and context-sensitive action	Expert-derived keys tested empirically; IRT or nominal-response models; response-process interviews

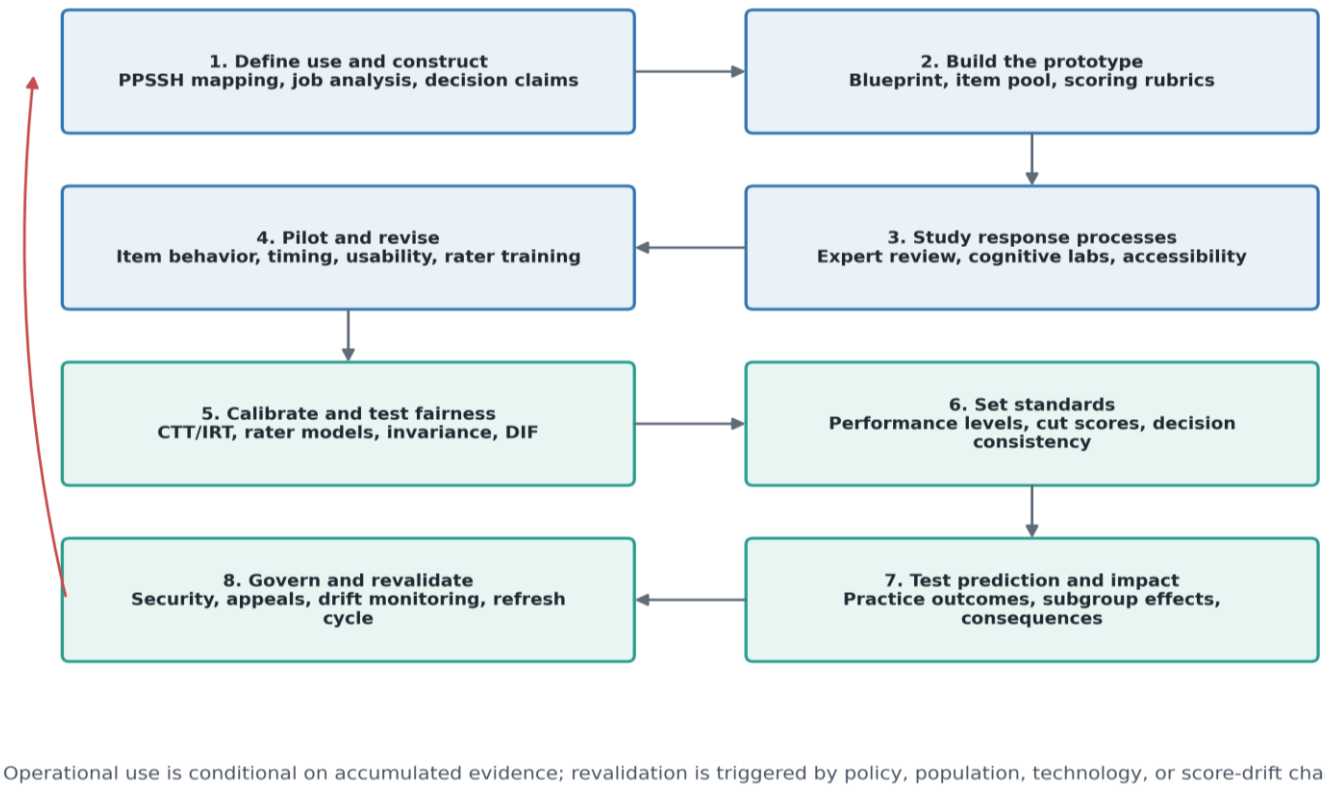
Component	Prototype v0.1 Field Form	Score Weight	Primary Claim	Required Analysis
C. Integrated cases	Two constructed-response cases scored with analytic rubrics by trained raters	20%	Diagnosis, option appraisal, implementation planning, communication, and justification	Rater training; double scoring; generalizability or many-facet analysis; moderation and audit

The proposed counts and weights are working assumptions rather than policy specifications: 60 selected-response items, 15 SJT sets, two constructed-response cases, domain weights of 20%, 20%, 25%, 20%, and 15%, and component weights of 50%, 30%, and 20%. A national job analysis could change the domain weights if task frequency, criticality, entry-level necessity, or consequences differ from the current blueprint. It could also add, combine, or remove claims and redistribute evidence across formats. Pilot findings on item information, local dependence, completion time, fatigue, security exposure, rater reliability, subgroup precision, and classification may lead to further changes in item counts and scoring. Any revision should remain traceable to evidence about the job and the intended decision, rather than to statistical fit alone.

For version 0.1, self-report, portfolios, and multi-rater observations are not included in the proposed examination score. They still have a place in a broader readiness system. Self-reflection may guide professional development, portfolios can document sustained work, and structured ratings may offer behavioral evidence after appointment. Keeping these sources outside the initial examination score avoids treating secure standardized tasks as equivalent to context-dependent evidence, while leaving room for later studies of incremental validity.

Proposed Validation Program

Figure 4. Staged Development, Validation, and Revalidation Program



Source: Proposed program informed by AERA et al. (2014), Fenn et al. (2020), Kalkbrenner (2021), Stefana et al. (2025), and Zhao et al. (2024).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

Table 10. Proposed Development, Validation, and Governance Program

Phase	Evidence-Producing Activities	Decision Product
1. Intended use and construct	Job analysis across varied Philippine school contexts; policy map; candidate and stakeholder interviews	Domain definitions, exclusions, claims, performance-level descriptors, prohibited interpretations
2. Content development	Blueprint; writer training; item/task review; expert panels; content-validity evidence	Traceable item-to-claim map and revision log
3. Response processes	Cognitive interviews/think-alouds; language and accessibility review; rater cognition study	Evidence that candidates and raters use the intended knowledge and criteria
4. Pilot and usability	Small diverse pilot; timing; missingness; item statistics; rubric trials; security/usability review	Revised items, tasks, rubrics, administration rules, and test specifications
5. Structure, precision, and fairness	Independent calibration/validation samples; IRT/Rasch; CFA/ESEM as justified; rater models; invariance and DIF	Domain/composite score model, conditional error, subgroup comparability, flagged content
6. Standards and decisions	Modified Angoff, Bookmark, or empirically appropriate process; panel replication; decision studies	Performance levels, provisional cut score(s), classification accuracy/consistency, review rules
7. External and consequential evidence	Prospective follow-up after appointment/training; multilevel models; adverse-impact and response studies	Predictive/incremental evidence, consequence profile, use restrictions, development utility
8. Governance and revalidation	Technical manual, security, privacy, appeals, accommodations, drift monitoring, periodic review	Versioned evidence record, release decision, refresh triggers, public technical reporting

A national sample should be stratified by DepEd region and should represent elementary, junior high, senior high, integrated, multigrade, geographically isolated, urban, rural, small, large, and differently resourced schools. It should also reflect relevant language backgrounds, prior opportunities for leadership, disability and accessibility needs, and differences in candidate experience. Translation and adaptation should include forward translation, reconciliation, back review, bilingual cognitive interviews, and empirical checks of comparability rather than relying on literal equivalence. Planned accommodations, accessible digital delivery, device and connectivity contingencies, and alternative response channels should all be tested for construct equivalence before use.

Standard setting should begin only after the construct, tasks, and scoring rules are sufficiently stable. Panelists need training to distinguish “minimally qualified” performance from what is merely “typical” or “desirable.” Their judgments should consider empirical difficulty, examples of candidate work, subgroup evidence, and likely consequences, without allowing a norm-based pass rate to replace the criterion definition. The technical report should describe uncertainty around the cut score and estimate the consistency of classification. Candidates whose scores fall near the standard should have a clear route for review or retesting, since a binary label can overstate differences within the score’s conditional error band.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

DISCUSSION

The Main Problem Is an Inference Gap, Not an Item Shortage

Although the mapped literature contains many item pools and reliable scales, it does not establish the inference needed for a school-head qualifying decision. Most existing instruments answer narrower questions. They ask, for example, how often a principal is perceived to show technological leadership, how strongly a respondent endorses innovation-leadership behaviors, or which emotional states are present. These questions are useful, but they do not determine whether an aspiring school head meets a criterion-referenced minimum for a complex public role. Beginning item writing before the decision claim is defined would reproduce the literature's central weakness in a new local form.

Evidence-centered design begins with the decision the assessment is meant to support. Performance-level descriptors then define minimally qualified practice, domain claims identify the required knowledge and judgment, and tasks are chosen because they can elicit relevant evidence. Scoring rules are subsequently tested against actual responses. This sequence is consistent with the staged approaches described by Kalkbrenner (2021) and Fenn et al. (2020) and with the emphasis on validity and use in the testing standards (American Educational Research Association et al., 2014). It also creates clear conditions under which the prototype must change. If tasks fail to elicit the intended reasoning, scores are not comparable, or classifications do not predict relevant practice, the instrument must be revised or its intended use narrowed.

Why a Multimethod Examination Is Justified

No single format captures readiness in the same way. Selected-response items can cover broad content and allow secure, efficient scoring, yet they may reward recognition without showing how a candidate frames an ambiguous problem. Situational-judgment tasks reveal priorities, ethical reasoning, and stakeholder trade-offs, although their scoring keys can depend on culture and context. Constructed cases allow candidates to diagnose a problem and justify a response, but they sample fewer situations and introduce variation from tasks and raters. The proposed 50:30:20 weighting is therefore a testable design choice, not a claim of psychometric optimality. It seeks a workable balance among breadth, contextual judgment, and integrated performance.

Method-discrepancy research provides the most direct rationale for a multimethod design. Kissi et al. (2023) and Quansah et al. (2025) found that reported competence in test construction was not interchangeable with the quality of the items teachers actually produced. The domain is different, but the implication carries over: a belief about one's capability and evidence from completed work are not the same. Aldrup et al. (2020) also demonstrate how a theory-driven situational judgment test can represent professional knowledge about relationship management through scenarios. For school-head qualification, the important empirical question is whether each method contributes information that is reliable, nonredundant, fair, and incrementally predictive, not simply whether a multimethod assessment appears more comprehensive.

PPSSH Alignment Is Necessary but Not Sufficient

Alignment with policy establishes relevance, but not validity. Because the PPSSH is the national professional standard, its five domains should guide construct definition and content coverage (Department of Education, 2020). The labels themselves, however, are too broad to function directly as score scales. "Building Connections," for example, has to be translated into observable judgments, behaviors, boundaries, and evidence. Without that work, item writers may substitute personal preferences for the intended construct. Job analysis must further distinguish competencies required at entry from those developed after appointment and from responsibilities that depend heavily on context and should not be treated as stable personal attributes.

Resource management makes this boundary especially clear. Hadi et al. (2022) show that participation in school budgeting can be measured as an organized construct, but a qualifying claim demands more than agreement with statements about good practice. A candidate should be able to interpret a constrained budget, identify compliance and equity risks, compare alternatives, and explain a defensible allocation. At the same

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

time, the task should not penalize candidates who have not previously worked in a resource-rich assignment. Context belongs within the problem to be reasoned through; it should not operate as an unmeasured advantage.

Fairness Requires More Than Equal Administration

A common national administration procedure does not guarantee that scores carry the same meaning or that all candidates have an equal opportunity to demonstrate readiness. Candidates differ in language background, disability and access needs, school level, region, connectivity, school size, access to technology and budgets, exposure to formal leadership work, mentoring, and familiarity with dominant administrative cases. A scenario built around resource-rich operations may measure prior opportunity rather than judgment. Cross-cultural development therefore calls for input from the target population, careful translation and adaptation, cognitive interviews, accessibility studies, separate structural confirmation, invariance and DIF analyses, subgroup precision, and impact analysis (Zhao et al., 2024). Each source answers a different question; no single source can establish fairness on its own.

A fairness review needs both statistical evidence and substantive judgment. A flagged item may reflect relevant experience, an irrelevant difference in opportunity, ambiguous language, an accessibility barrier, regional or school-level specificity, rater effects, or model misfit. Review panels should therefore include expertise in regional and Indigenous contexts, language, disability and accessibility, school levels, and resource conditions. Depending on the evidence, the response may be to revise an item, use contextual branching or alternate forms, provide accommodations, gather additional evidence, or narrow the interpretation. Statistical concealment is not an acceptable response. Ongoing monitoring should examine access to preparation, device and connectivity burden, accommodation use, retesting, appeals, subgroup classifications, and whether resource-related opportunities continue to predict scores after relevant readiness has been controlled.

Standard Setting Changes the Question

Reliability and standard setting answer different questions. Reliability concerns whether scores are consistent enough for an intended use; standard setting identifies the evidence that should count as minimally qualified performance. The absence of explicit standard-setting signals in the mapped corpus is therefore a serious gap. Percentile categories reported in some scale studies show where a candidate stands relative to a sample, but they do not define the competence needed to lead safely and effectively. A qualifying examination requires performance-level descriptors, trained panel judgment, empirical information about tasks, replication, and analysis of classification consistency. Its cut score should be treated as a policy judgment supported by evidence, not as a psychometric constant produced by software.

Any decision rule also has to account for weakness within a domain. Under a fully compensatory composite, very high strategic knowledge could offset poor ethical judgment about resources. Requiring a hard minimum in every domain creates a different problem, since each subscore contains less information and may produce unstable classifications. A hybrid rule, combining an overall standard with safeguards for one or more critical domains, may prove defensible. It should be selected only after simulations examine information, error, and consequences. The rule must be public enough to permit review while still protecting item security.

Prediction, Consequences, and Governance Complete the Validity Argument

An assessment of readiness should predict relevant practice beyond the test itself. Prospective validation should relate scores to structured observations of leadership, the quality of school-improvement planning, teacher-development processes, stakeholder participation, resource stewardship, and other outcomes close to the construct. Student or school performance may also be examined, but attribution requires caution because those indicators reflect prior attainment, staffing, poverty, location, policy, and unexpected shocks. Multilevel models with baseline adjustment are preferable to simple correlations. Primary predictive outcomes should also remain close enough to the readiness construct to support a meaningful interpretation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

The consequences of assessment extend well beyond pass rates. A system may encourage preparation that is narrowly coached to the test, create unequal access to review materials, deter candidates in remote settings, or lead to disclosure of secure scenarios. A well-designed developmental report, however, may direct preparation toward authentic demands of the role. Monitoring should cover subgroup opportunity, access to preparation, use of accommodations, appeal patterns, security incidents, candidate response, appointment outcomes, and changes in professional learning after feedback. Moving the prototype into high-stakes use should be a governance decision based on the accumulated evidence, not an automatic step once alpha and model-fit statistics reach acceptable values.

Gap Analysis

Table 11. Evidence Gaps, Risks, and Required Research Responses

Gap	Mapped Evidence	Risk	Required Research Response
Construct integration	Only 12 of 95 mapped studies directly concerned principal qualification/readiness; principal instruments addressed segmented constructs.	No mapped integrated Philippine qualification construct.	PPSSH job analysis and evidence-centered construct map; test correlated-domain versus higher-order structures.
Method representation	23 self-report/survey signals; one strict performance/SJT signal in indexed text.	Perceived readiness may be mistaken for demonstrated judgment.	Compare selected response, SJT, constructed response, and external evidence using multitrait–multimethod and incremental-validity analyses.
Decision standards	No explicit formal standard-setting signal.	Reliable ranking cannot justify pass/fail classification.	Develop performance levels; conduct and replicate criterion-referenced standard setting; estimate classification consistency.
Fairness/comparability	Two invariance/DIF signals.	National scores may conflate readiness with language, locale, school opportunity, or item context.	Cognitive labs, accessibility review, invariance/DIF, subgroup precision, impact and qualitative follow-up.
Temporal evidence	No longitudinal or test–retest signal.	Readiness stability, responsiveness, and score drift are unknown.	Test temporal stability, training sensitivity, appointment follow-up, and scheduled revalidation.
Outcome linkage	Two predictive/criterion signals.	A passing score may not forecast relevant practice.	Prospective multilevel validation and incremental prediction beyond prior credentials and experience.
Reporting completeness	39 of 95 records were partial indexed records; only a targeted core was full-text verified.	Study limitations and methods may be under-observed.	Rerun a reproducible multi-database search, retrieve all reports, dual-screen, and conduct formal appraisal.
Philippine evidence	One mapped Philippine-context record concerned pre-service teachers, not school heads.	No local empirical basis for a principal qualifying interpretation was identified.	National multiphase study across varied Philippine school and candidate contexts.

These gaps cannot be addressed as if they were independent. A larger sample cannot repair a weak construct map, and a well-fitting model cannot correct construct underrepresentation. Likewise, a fair item bank does not

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue XI November 2026

justify an arbitrary cut score, and a stable cut score does not justify use when an assessment fails to predict relevant practice or causes unacceptable harm. The proposed validation program is therefore cumulative and conditional. When evidence fails at an early stage, the prototype should return to redesign rather than allowing later statistical complexity to conceal the problem.

Future Research

Immediate Development Studies

- Conduct a national PPSSH-linked job analysis using policy review, critical-incident interviews, task inventories, and diverse expert panels to distinguish entry-level requirements from competencies developed after appointment.
- Develop performance-level descriptors before final item writing so "minimally qualified" has an observable meaning within each domain.
- Generate substantially more items and scenarios than the field form requires, document every revision, and maintain a traceable claim–evidence–task matrix.
- Run cognitive interviews in relevant language versions and contexts, including candidates from geographically isolated, resource-constrained, and accessibility-relevant settings.
- Pilot constructed-response rubrics with multiple raters and investigate which dimensions can be scored reliably without collapsing complex reasoning into superficial checklists.

Field Validation and Decision Research

- Use independent calibration and confirmation samples; compare plausible domain structures and retain a composite only when its interpretation is supported.
- Model item and task information across the readiness continuum; investigate local dependence, speediness, form equivalence, and rater severity.
- Test measurement invariance and DIF across policy-relevant groups, then conduct substantive reviews of flagged items rather than relying on mechanical deletion.
- Compare compensatory, conjunctive, and hybrid qualification rules through standard-setting replication and decision-accuracy/consistency simulations.
- Estimate the incremental contribution of SJT and constructed-response components beyond applied-knowledge scores, prior experience, training, and existing credentials.

Longitudinal, Consequential, and Governance Research

- Follow candidates through preparation, appointment, and early tenure to examine stability, responsiveness, score drift, and prediction of proximal leadership practice.
- Pre-register primary outcomes and analysis plans; separate confirmatory prediction from exploratory school-performance associations.
- Monitor subgroup access, accommodation, preparation opportunity, retest behavior, appeals, security incidents, and unintended coaching effects.
- Study whether developmental score reports improve targeted learning and whether feedback benefits candidates near and far from the standard.
- Establish independent technical review, version control, public documentation, and explicit triggers for suspension, revision, or revalidation.

A future systematic review should be planned prospectively. Its protocol should be registered, exact database strings and run dates published, additional education and psychology databases searched, and Google Scholar ranking and retrieval limits documented. It should retrieve all eligible reports, use independent dual screening and extraction, report agreement and adjudication, apply critical appraisal suited to each design, and update the search before submission. Because the historical Scopus query and timestamp cannot be recovered from the supplied export, a reproducible Scopus search should be developed for that future review. The current

evidence map remains a useful, auditable exploratory baseline with a record-level supplement, but it should not be relabeled as a complete full-text systematic review.

Limitations

Limitations of the Evidence Map

Several features of the evidence map limit what can be concluded. The analysis began with fixed source exports; the historical Scopus query and search date could not be recovered; Google Scholar was limited to its first 500 ranked results; screening was title-led and rule-assisted; one reviewer completed the coding; no agreement estimate was available; and full-text access was incomplete. Exact normalized-title deduplication may also miss near duplicates or, in rare cases, merge title variants. Because indexed abstracts and snippets often omit methods and limitations, the method-signal counts describe terms reported in the supplied fields rather than the true prevalence or quality of methods. The 12-article full-text core was selected purposefully for its interpretive value. It is neither a second eligibility stage nor a representative sample. Together, these conditions limit the map's completeness, the reproducibility of the historical Scopus search, and the strength of the inferences that can be drawn.

Limitations of the Proposed Prototype

The prototype has not yet been subjected to a national job analysis, operational item writing, cognitive testing, pilot administration, calibration, rater validation, invariance or DIF analysis, standard setting, predictive validation, security review, or evaluation of consequences. Every major design feature remains provisional, including domain and component weights, item counts, descriptors, formats, scoring criteria, and decision rules. Alignment with the PPSSH and support from the literature establish plausibility; they do not establish validity. Unless DepEd authorizes the use and the complete evidence program supports the intended population, testing conditions, language and accessibility arrangements, classifications, and safeguards, the model must not replace NASH/NQESH or be used for employment-related consequences.

CONCLUSION

The literature points to a straightforward conclusion. School-head assessment has many useful component instruments, but the evidence is not yet sufficient for an integrated, high-stakes qualification decision in the Philippines. Most mapped studies focus on construct structure and self-report, while fairness, standard setting, longitudinal prediction, classification consistency, and consequences are seldom visible. The appropriate next step is to build a prototype whose assumptions can be tested systematically, not to claim that a validated examination already exists.

The proposed model connects the five PPSSH domains with applied-knowledge items, situational-judgment scenarios, and constructed-response cases. Validation would begin with job analysis and performance-level descriptors, then proceed through studies of content and response processes, psychometric calibration, rater modeling, fairness across groups, criterion-referenced standard setting, predictive follow-up, monitoring of consequences, and continuing governance. Until this evidence has been assembled, the prototype should be used only for research and developmental feedback. If the program supports defensible score meaning, comparable measurement, stable classification, incremental prediction, and acceptable consequences, the model could offer a transparent basis for qualification. If it does not, the intended use should be revised or withheld.

Supplementary material.

Supplementary File S1 provides the 95-record evidence map, evidence-completeness and source-basis codes, record-level indexed-text signals, category and signal codebooks with examples, search and export provenance, the normalization and deduplication algorithm, and the complete screening log for 6,172 records.

REFERENCES

1. Aldrup, K., Carstensen, B., Koller, M. M., & Klusmann, U. (2020). Measuring teachers' social-emotional competence: Development and validation of a situational judgment test. Frontiers in Psychology, 11, 892. https://doi.org/10.3389/fpsyg.2020.00892
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
3. Arjanto, P., Senduk, F. F. W., Melati, I. S., Maduretno, T. W., & Kawulur, H. R. (2025). Moral leadership in primary education: Constructing and validating an instrument for principal assessment. Educational Process: International Journal, 14, e2025007. https://doi.org/10.22521/edupij.2025.14.7
4. Aydin, K., & Ceylan, M. (2026). Development and validation of a digital leadership scale for school principals in the AI-driven era. NALANS: Journal of Narrative and Language Studies. https://doi.org/10.59045/nalans.2026.90
5. Chen, J. (2021). Development and validation of the Principal Emotion Inventory: A mixed-methods approach. Educational Management Administration & Leadership. https://doi.org/10.1177/1741143220919764
6. Darwis, A., Bafadal, I., Wiyono, B. B., Sultoni, & Malik, A. R. (2025). Designing an assessment tool for teacher leadership competencies in aspiring elementary school principals in Indonesia. Sustainable Futures, 100803. https://doi.org/10.1016/j.sftr.2025.100803
7. Department of Education. (2020). National adoption and implementation of the Philippine Professional Standards for School Heads (DepEd Order No. 24, s. 2020). Republic of the Philippines. https://www.deped.gov.ph/wp-content/uploads/2020/09/DO_s2020_024.pdf
8. Department of Education. (2024). Administration of the Fiscal Year 2023 National Qualifying Examination for School Heads (DepEd Memorandum No. 13, s. 2024). Republic of the Philippines. https://www.depedncr.com.ph/wp-content/uploads/2024/03/DM_s2024_013.pdf
9. Department of Education. (2025). Administration of the National Assessment for School Heads for Fiscal Year 2025 (DepEd Memorandum No. 69, s. 2025). Republic of the Philippines. https://www.deped.gov.ph/2025/08/?cat=6
10. Department of Education. (2026, May 19). Halos 24,000 na nagnanais maging school principal, sumalang sa nationwide assessment ng DepEd. https://www.deped.gov.ph/2026/05/19/halos-24000-na-nagnanais-maging-school-principal-sumalang-sa-nationwide-assessment-ng-deped/
11. Durnalı, M. (2022). The development and validation of technological leadership behavior instrument for school principal. Journal of Learning and Teaching in Digital Age, 7(2), 210–221. https://doi.org/10.53850/joltida.1035591
12. Fenn, J., Tan, C. S., & George, S. (2020). Development, validation and translation of psychological tests. BJPsych Advances, 26, 306–315. https://doi.org/10.1192/bja.2020.33
13. Fuad, D. R. S. M., Musa, K., Yusof, H., & Hashim, Z. (2022). The development and validation of the principal innovation leadership scale in Malaysian secondary schools. International Journal of Evaluation and Research in Education, 11(1). https://doi.org/10.11591/ijere.v11i1.22038
14. Hadi, S., Maisaroh, S., Hidayat, A., & Andrian, D. (2022). An instrument development to evaluate teachers' involvement in planning the school budgeting at elementary schools of Yogyakarta Province. International Journal of Instruction, 15(2), 1087–1100. https://doi.org/10.29333/iji.2022.15260a
15. Kadri, K., Mansor, A. N., & Mohd Nor, M. Y. (2021). Principal and teacher leadership competencies and 21st century teacher learning and facilitating practices: Instrument development and demographic analysis. Creative Education, 12, 2196–2215. https://doi.org/10.4236/ce.2021.129168
16. Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The MEASURE approach. Practical Assessment, Research & Evaluation, 26, Article 1. https://doi.org/10.7275/svg4-e671
17. Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in examination-oriented educational systems. Frontiers in Education, 8, 1154592. https://doi.org/10.3389/feduc.2023.1154592

18. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
19. Quansah, N., Quansah, F., & Dzakadzie, Y. (2025). Multiple-choice test development competencies of junior high school mathematics teachers in Ghana: A triangulation methodology. Frontiers in Education, 10, 1658971. https://doi.org/10.3389/feduc.2025.1658971
20. Stefana, A., Youngstrom, E. A., Chen, J., Hinshaw, S., Maxwell, V., Michalak, E., et al. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: Best practice guidelines for their development and validation. Frontiers in Psychology, 15, 1494261. https://doi.org/10.3389/fpsyg.2024.1494261
21. Yuliana, L., Prasojo, L. D., & Akalili, A. (2022). Analysis of confirmatory factors of principals' leadership training of vocational high school. Cakrawala Pendidikan, 41(3), 599–618. https://doi.org/10.21831/cp.v41i3.50496
22. Zhao, Y., Summers, R., Gathara, D., & English, M. (2024). Conducting cross-cultural multi-lingual or multi-country scale development and validation: A 10-step framework based on a scoping review. Journal of Global Health, 14, 04151. https://doi.org/10.7189/jogh.14.04151

Appendix A. Evidence-Use Boundary

Table 12. Evidence-Use Boundaries for the Review and Prototype

Evidence Layer	Material	Permitted Interpretation
Mapped evidence	95 records retained after title/indexed-text screening	Supports thematic and gap patterns; not a full-text effect synthesis
Full-text verified evidence	12 linked scholarly articles	Supports the specific method, sample, finding, and limitation statements attributed to those papers
Abstract-supported evidence	Principal studies with detailed database or publisher abstracts	Supports only claims visible in those abstracts; unreported limitations are not inferred
Authoritative context	PPSSH policy; historical NQESH and current NASH documents; testing standards; PRISMA guidance	Supports policy, reporting, and test-use requirements; not counted as empirical studies
Prototype statements	Blueprint, descriptors, tasks, weights, item counts, scoring criteria, validation phases, and research questions	Research proposals requiring empirical confirmation; not presented as validated results

Appendix B. Minimum Evidence Before High-Stakes Use

- A nationally grounded construct map and blueprint endorsed through documented job analysis and stakeholder review.
- Stable scoring rules and administration procedures with secure, accessible forms and trained personnel.
- Adequate reliability and conditional precision at the proposed decision point for each reported score.
- Replicated structural evidence and defensible aggregation of domain/component scores.
- Substantive and statistical fairness evidence across relevant groups and language versions.

- Criterion-referenced performance standards with documented panel process, uncertainty, and classification consistency.
- Prospective evidence linking scores to relevant leadership practice and showing incremental value over existing information.
- A consequence, appeals, privacy, security, monitoring, and revalidation system with independent technical oversight.

Appendix C. Indexed-Text Signal Codebook

A signal was coded as 1 only when a prespecified term appeared in the exported title or abstract or snippet; otherwise, it was coded as 0. The presence of a signal does not show that the method was used, used well, or absent from the full report. Matching was case-insensitive and included common spelling and hyphenation variants.

Table 13. Indexed-Text Signal Codebook and Reporting Counts

Signal	Indexed-text terms	n	Positive example	Interpretation
Factor analysis	EFA; CFA; exploratory/confirmatory factor; factor analysis	27	"confirmatory factor analysis"	Construct-structure term present
Rasch/IRT	Rasch; item response theory; IRT	8	"Rasch model"	Item-model term present
Self-report/survey	self-report; questionnaire; survey	23	"survey questionnaire"	Response-mode term present
Performance/SJT	performance-based; observation; portfolio; situational judgment/SJT; multi-source	1	"situational judgment test"	Applied evidence term present
Longitudinal	longitudinal	0	—	No term visible in indexed text
Test–retest	test-retest; temporal stability	0	—	No term visible in indexed text
Invariance/DIF	measurement invariance; differential item functioning; DIF	2	"measurement invariance"	Comparability term present
Predictive/criterion	predictive; criterion validity; prediction	2	"predictive validity"	Outcome-relation term present
Standard setting	cut score; standard setting; classification accuracy/consistency	0	—	Decision-standard term absent from indexed text

Appendix D. Supplementary File S1 - Evidence Map and Reproducibility Audit Trail

Metric	Value		Interpretation boundary
Scopus export rows	5750		This workbook supports a single-reviewer exploratory evidence map. Method columns are binary indexed-text reporting signals: 1 means the prespecified term appeared in the exported title or abstract/snippet; 0 means the term was not visible there. These values are not verified method-use indicators, quality ratings, or prevalence
Google Scholar export rows	500		
Combined exported rows	6250		

Duplicate rows removed	78		estimates. Twelve purposefully selected linked scholarly full texts strengthened interpretation but were not a representative second eligibility stage.
Cross-database overlaps	76		
Unique titles screened	6172		
Records mapped	95		**Historical-search caveat**
Records excluded	6077		The Google Scholar export preserves its historical query, timestamp, 2020–2026 filter, and first-500 ranked-record limit. The supplied Scopus workbook does not preserve the exact historical query or database-run/export timestamp. Those details are marked unavailable; they have not been reconstructed or invented.
Full database abstracts	56		
Partial indexed records	39		

Access the spreadsheet using this link: https://drive.google.com/drive/folders/1YRAAPnDqlXu-Ry2do5XmmyYI25K4kgvp?usp=sharing