

Curriculum, Teaching Methods, and Assessment in Arabic Language Education: A Department-Level Quality Assurance Evaluation

Saproni Muhammad SAMIN

Abstract

This study evaluates a department-level holistic assessment system in Arabic language teacher education using Goal-Free Evaluation within Indonesia's Determination–Implementation–Evaluation–Control–Improvement quality assurance cycle. Drawing on a convergent mixed-methods design, the analysis triangulates three years of institutional indicators (2022–2025): cohort grade point average and on-time graduation, a student satisfaction survey of 179 respondents, graduate tracer outcomes (39 of 42 responses), quality-assurance documentation (department-to-course learning outcome mapping; 40–30–30 cognitive–affective–psychomotor weighting), and standardized micro-teaching parameters. Results show sustained high grade point average (3.69–3.78), improved on-time completion in the most recent year (52%), consistently high very good service ratings (at least 75% across dimensions), auditable micro-teaching operations (at least four practices per semester; at least ten skills; B-minus threshold), and strong graduate outcomes (tracer coverage 92.86%; placement 97.5%; mean time-to-job 4.8 months). The study contributes an audit-ready account of how constructive alignment and culturally responsive assessment support quality education at department scale, while identifying priorities for rubric standardization, assessor moderation, reliability quantification, and outcome-based education analytics pipelines.

Keywords: Quality education, outcome-based education, holistic assessment, goal-free evaluation, Arabic language education, quality assurance

INTRODUCTION

Outcome-Based Education (OBE) fundamentally transforms educational paradigms by shifting focus from teaching activities to demonstrable student capabilities. This approach reframes curricula around what graduates can accomplish, translating Department Learning Outcomes (DLO) into concrete Course Learning Outcomes (CLO), tasks, and precise assessment criteria (Biggs, 2003; Biggs & Collis, 2014; Harden, 2007; Munzenmaier & Rubin, 2013; Spady, 1994). In language teacher education—specifically Arabic Language Education—this means assessment must transcend traditional knowledge evaluation, capturing not only cognitive mastery but also professional dispositions and performative teaching/language abilities that reflect authentic graduate profiles. Despite the promise of OBE, contemporary educational research reveals persistent implementation challenges. State-of-the-art reviews highlight recurring weaknesses: overreliance on self-reported evidence, limited longitudinal tracking, uneven assessment standardization, and immature data governance at the Department level (Constantinou, 2020; Mufanti et al., 2024; Yadav et al., 2024). These systemic gaps compromise the potential of outcome-based approaches to genuinely transform educational experiences and graduate readiness. Digital transformation offers promising remedies to these challenges. Advanced technologies can enable workflow automation, accelerate formative feedback, and generate sophisticated learning analytics—provided that robust policy frameworks, technological infrastructure, and assessment literacy are in place (Anderson et al., 2024; Kim & Su, 2024; Lugay et al., 2020; Qiao & Fu, 2022; Thompson & Asanov, 2024). Simultaneously, pedagogical innovations like experiential learning and reflective practice position repeated authentic tasks—such as micro-teaching—as critical mechanisms for developing higher-order professional skills (Kolb et al., 2014; Kolb et al., 2024; Schön et al., 2024).

Emerging assessment scholarship advocates for holistic design approaches that comprehensively span cognitive, affective, and psychomotor domains. By combining holistic and analytic scoring methods, educators can generate more nuanced diagnostic feedback and enhance decision-making quality (Singh et al., 2019; Zlatkin-Troitschanskaia et al., 2019). Innovative assessment formats like Multiple Mini Viva (MMV) demonstrate how multi-dimensional evaluation can triangulate cognitive, interpersonal, and professional competencies through blended assessment strategies (Hadie et al., 2022).

In culturally diverse contexts, particularly within Indonesian and faith-based educational settings, assessment design must balance local values with international standards. Vygotsky's socio-cultural theory

provides a foundational framework for understanding how cultural tools, social interactions, and zone of proximal development shape learning outcomes (Vygotsky, 1978). This theoretical perspective emphasizes that assessment practices are not culturally neutral but embedded within specific socio-cultural contexts where learning occurs through mediated social interactions. Research indicates that culturally responsive approaches, grounded in socio-cultural principles, can significantly boost student motivation, engagement, and long-term educational sustainability (Conradty & Bogner, 2022; Handrianto et al., 2025; Theodoridou, 2021). Within Arabic Language Education, this means creating ecosystems that integrate Islamic–Malay pedagogical principles—where knowledge transmission occurs through culturally meaningful interactions and authentic community practices—while maintaining rigorous academic standards. The theoretical framework of this study thus integrates constructive alignment (Biggs, 2003) with socio-cultural perspectives (Vygotsky, 1978) to examine how holistic assessment systems function within culturally embedded educational contexts.

My research setting—the Arabic Language Education Department at Islamic University of Riau—embodies these complex educational dynamics. The Department operates an Outcome-Based Education curriculum supported by micro-teaching and language laboratories, comprehensive service documentation, and systematic graduate tracking mechanisms. The core challenge lies in maintaining a 40–30–30 (cognitive–affective–psychomotor) assessment portfolio that remains traceable to institutional indicators within the DIECI quality assurance cycle (Determination–Implementation–Evaluation–Control–Improvement).

This study advances three critical contributions to the field. First, I apply Goal-Free Evaluation (GFE) at the Department level to rigorously assess the effectiveness of a DLO–ELO-aligned holistic assessment system (Biggs, 2003; Biggs & Collis, 2014; Harden, 2007; Spady, 1994; Yadav et al., 2024). Second, I formalize the micro-teaching ecosystem as auditable psychomotor evidence, directly linking it to key institutional performance indicators like Grade Point Average (GPA) and on-time graduation (Al-Mubaid & Bettayeb, 2017; Kolb et al., 2014; Richards & Rodgers, 2014). Third, I propose a comprehensive quality assurance strengthening agenda, including cross-course rubric standardization, assessor moderation, and automated data pipelines that connect DLO–ELO–tasks–scores–outcomes (Lugay et al., 2020; Mitra & Wadegaonkar, 2024; Thompson & Asanov, 2024; Zlatkin-Troitschanskaia et al., 2019).

RESEARCH GAP AND NOVELTY

Existing studies in Arabic language education predominantly operate at the course level, relying heavily on self-report methodologies (Al-Mubaid & Bettayeb, 2017; Richards & Rodgers, 2014). Systematic reviews reveal that these studies rarely connect holistic assessment portfolios to auditable Department-level indicators within formal quality assurance cycles (Mufanti et al., 2024; Yadav et al., 2024). A comprehensive analysis of international teacher education literature demonstrates that few studies employ Goal-Free Evaluation to minimize goal bias or report standardized micro-teaching parameters as psychomotor evidence with rigorous moderation and digital workflows (Constantinou, 2020; Lugay et al., 2020). Specifically, GFE practices are rare at the departmental level in the context of international teacher and language education, and existing evidence is mostly limited to classroom practices or institutional policy narratives (David, 2025; Qiao & Fu, 2022). This gap is particularly evident in Islamic higher education contexts, where Department-level evaluations connecting assessment design to institutional indicators remain limited and often goal-biased (Handrianto et al., 2025). My systematic review of recent literature (2019–2024) confirms that comparable Department-level GFE applications in teacher and language education remain scarce, with most evidence confined to classroom-level interventions or institutional policy narrative levels rather than comprehensive Department-level quality assurance cycles.

By applying GFE at the Department scale, triangulating institutional aggregates across three years, and presenting traceable tables that link assessment domains to indicators within the DIECI cycle, my study addresses these critical gaps. Table 1 provides a comprehensive map of indicators and sources, while Tables 2–5 in the Discussion section report detailed Department performance and outcomes.

Constructive alignment at the Department scale demands robust evidence chains that connect intended outcomes to specific tasks and assessment scores. Prior research demonstrates that explicit criteria and shared rubrics can effectively shape learning behaviors and yield more stable attainment across student cohorts (Biggs & Collis, 2014; Harden, 2007). I extend this logic by explicitly tying aligned tasks to institutional indicators—specifically GPA and on-time completion—thereby rendering alignment auditable beyond individual course boundaries.

My holistic assessment model integrates cognitive mastery with professional attitudes and observable performance. Synthesizing existing scholarship, I operationalize this integration through a carefully designed 40–30–30 portfolio and systematic micro-teaching evidence collection. This approach aligns with contemporary recommendations that combine holistic judgments with analytic sub-scores to improve diagnostic usefulness and scoring dependability (Singh et al., 2019; Zlatkin-Troitschanskaia et al., 2019).

LITERATURE REVIEW

CONSTRUCTIVE ALIGNMENT AND OBE

Classic and contemporary works argue that clarity of intended outcomes and assessment criteria drives learning behaviors and attainment (Biggs, 2003; Biggs & Collis, 2014; Harden, 2007; Munzenmaier & Rubin, 2013; Spady, 1994). Reviews in OBE emphasize mapping outcomes to tasks and evidence at scale and caution against reliance on self-report metrics (Yadav et al., 2024).

At Department level, constructive alignment requires auditable chains linking PLO–CLO–tasks–rubrics–scores–indicators so that claims of attainment can be externally verified. Practical enablers include CO-PO matrices in course outlines, cross-course rubric libraries, and moderation routines that stabilize criteria interpretation across assessors and semesters (Biggs & Collis, 2014; Harden, 2007; Yadav et al., 2024). I adopt this view by reading alignment not only inside courses but also in institutional indicators (GPA, on-time graduation), which situates OBE evidence within quality assurance rather than isolated classroom practice.

A recurrent gap is translation fidelity from curriculum design to enacted assessment. Studies note that without shared exemplars and performance-level descriptors, alignment degrades at scale and weakens the validity of inferences drawn from Department metrics (Biggs & Collis, 2014). This motivates my emphasis on standardized rubrics and moderation as infrastructure for credible Department-level evaluation.

HOLISTIC ASSESSMENT AND SCORING

Evidence supports integrating cognitive–affective–psychomotor domains and combining holistic with analytic sub-scores to improve diagnostic power and reliability (Singh et al., 2019; Zlatkin-Troitschanskaia et al., 2019). Station-based oral formats like MMV triangulate indicators across domains (Hadie et al., 2022).

For Department-level teacher education, holistic ratings (overall lesson quality) provide efficient judgments of authentic performance, while analytic sub-skills (e.g., questioning, reinforcement, classroom management) localize improvement needs for feed-forward. Reliability is strengthened through common rubric descriptors, assessor calibration, double-rating where feasible, and routine moderation sessions; where item-/rater-level files exist, alpha and ICC quantify internal consistency and inter-rater agreement to support equitable decisions (Zlatkin-Troitschanskaia et al., 2019).

Recent syntheses further caution that holistic scores alone can mask differential growth across sub-skills, while purely analytic scoring risks fragmentation; hence, blended scoring architectures with linked remediation plans are preferred in practice (Singh et al., 2019). I therefore interpret holistic outcomes alongside analytic indicators of sub-capabilities when available.

EXPERIENTIAL LEARNING AND MICRO-TEACHING

Experiential and reflective traditions justify repeated, authentic practice with structured feedback cycles, which micro-teaching operationalizes in teacher education (Kolb et al., 2014; Kolb et al., 2024).

Cycled rehearsal (plan–teach–record–review–reteach) cultivates automatization of core skills, metacognitive monitoring, and fluent professional communication. When aligned with course outcomes and assessed using shared rubrics, these cycles generate evidence that transfers to school placements and early career indicators, linking classroom performance to employability signals observed in tracer studies (Kolb et al., 2014; Richards & Rodgers, 2014).

In language teacher education, micro-teaching also strengthens disciplinary language use (listening/speaking) and classroom discourse management. Beyond skill rehearsal, it strengthens disciplinary language use and classroom discourse. Capturing videos, annotated rubrics, and short reflections in an e-portfolio builds a longitudinal record of growth that supports Department reviews and accreditation cycles.

DIGITAL TRANSFORMATION IN ASSESSMENT

Technology-enabled workflows, e-portfolios, rubric banks, and learning analytics can enhance timeliness, traceability, and feedback—subject to governance and assessor literacy (Anderson et al., 2024; Lugay et al., 2020; Qiao & Fu, 2022; Thompson & Asanov, 2024).

For Arabic teacher education, priority capabilities include rubric versioning and sharing across courses; video-annotated micro-teaching with timestamped comments; automated calculation of criterion scores and CO-PO dashboards; and safeguards for privacy/ethics (role-based access, consent, retention). These affordances shorten feedback loops and create an auditable trail from performance artifacts to Department-level indicators (Anderson et al., 2024; Qiao & Fu, 2022).

Adoption, however, is uneven: institutions report gaps in assessor readiness, policy clarity, and system integration (Thompson & Asanov, 2024). This underlines the need for targeted capacity-building and minimal-viable pipelines that demonstrate value without imposing excessive administrative burden.

CONTEXTUALIZATION IN INDONESIAN HIGHER EDUCATION

Integrating local Islamic–Malay values without compromising standards is linked to student engagement and sustainability (Conradty & Bogner, 2022; Handrianto et al., 2025; Theodoridou, 2021). Department-level evaluations that connect assessment design to institutional indicators remain limited and often goal-biased (David, 2025; Qiao & Fu, 2022).

Within Indonesia’s DIECI cycle, much documentation centers on compliance artifacts rather than outcome-anchored evaluation designs. By adopting GFE, this study minimizes bias toward formal targets and recenters empirical appraisal on what the indicators show—GPA, on-time completion, satisfaction, and tracer—thereby contributing a department-level, audit-ready account of how holistic assessment links to institutional signals.

Additionally, the Islamic–Malay context foregrounds professional ethics and community engagement. Aligning affective indicators (assurance, empathy, service climate) with culturally grounded practices may explain sustained satisfaction scores; documenting these linkages provides externally legible evidence for international audiences.

This study evaluates a department-level holistic assessment system and its alignment with cycle Determination, Implementation, Evaluation, Control, and Improvement (DIECI) using GFE, and it examines how institutional indicators and graduate outcomes evolved across three academic years.

METHOD

Design and approach. I used Goal-Free Evaluation (GFE) to appraise a department-level holistic assessment system while deliberately bracketing formal targets (Scriven, 1972; Youker, 2013). GFE was selected to minimize goal bias and enable objective evaluation based on actual empirical evidence rather than predetermined program objectives. The study followed a convergent mixed-methods design (Creswell & Plano Clark, 2016), where quantitative and qualitative data were collected simultaneously, analyzed separately, and then integrated to provide comprehensive understanding of the assessment system's effectiveness. This design choice was justified by the need to triangulate multiple data sources (academic records, QA documents, surveys, tracer data) while maintaining methodological rigor through independent quantitative and qualitative analyses before integration (Tashakkori & Teddlie, 2010). The quantitative strand employed descriptive statistics to summarize indicator trends across three academic years, while the qualitative strand utilized document analysis to interpret contextual factors and institutional processes supporting the assessment system.

CONTEXT AND UNIT OF ANALYSIS

The setting was the Arabic Language Education Department at Islamic University of Riau. The unit of analysis comprised Department-level indicators covering three assessment domains (cognitive, affective, psychomotor) and graduate outcomes across the three most recent academic years (Y-2 to Y; 2022–2023 to 2024–2025). Participants Department cohorts from Y-2 through Y; student respondents in the institutional satisfaction survey (N=179); and graduates included in the tracer study (target 42; responses 39).

DATA COLLECTION AND ANALYSIS

Data sources; (1) Academic records: cohort mean GPA, on-time completion rate, and the 144-credit requirement. (2) Quality-assurance documentation: course outlines; DLO–CLO–sub-CLO mapping; the 40–30–30 assessment distribution; semesterly DLO/CLO monitoring; internal audits and management reviews, DIECI implementation files. (3) Institutional student satisfaction survey (N=179) across five service dimensions: reliability, responsiveness, assurance, empathy, facilities. (4) University-level tracer of graduates (coverage 42; responses 39) tracking placement, post-graduation activity, and time-to-job. (5) Micro-teaching artefacts: observation forms and rubrics for ≥10 basic skills, minimum pass threshold B-, records of ≥4 practices/semester, and laboratory support.

INSTRUMENTS AND PROCEDURES

I consolidated data from institutional academic systems and Department QA archives following established protocols. Cognitive-domain artefacts included quizzes, assignments, essays, mid-/final-term exams, and portfolios aligned with PLO/CLO.

For the affective domain, I employed multiple validated instruments. The institutional student satisfaction survey utilized SERVQUAL-based dimensions (reliability, responsiveness, assurance, empathy, facilities) with established psychometric properties (Parasuraman et al., 1988; Petrušić, 2023). The survey employed a five-point Likert scale (1=very poor to 5=very good), with "very good" ratings (score ≥4) serving as the primary affective indicator. Additionally, affective evidence was triangulated through documented activities from the Sexual-Violence Prevention and Handling (SVPH) Taskforce, Occupational Safety and Health (OSH) programs, and counselling services, which reflect institutional commitment to student welfare and professional development (Handrianto et al., 2025; Theodoridou, 2021). These complementary sources provided convergent evidence of affective domain outcomes beyond self-reported satisfaction.

Psychomotor-domain evidence included standardized micro-teaching rubrics aligned with course outcomes, practice recordings, and facilities logs. All data collection followed QA Standards Operational Procedures (SOPs) and internal audit routines to ensure consistency and traceability.

RESEARCH PROCESS FLOW

The study followed a multi-stage sequential process: (1) Preparation and instrument validation—reviewing QA documentation, validating survey instruments, and establishing data access protocols; (2) Quantitative data collection—extracting academic records, compiling satisfaction survey responses, and gathering tracer data; (3) Qualitative data collection—analyzing QA documents, micro-teaching artifacts, and institutional reports; (4) Independent analysis—computing descriptive statistics for quantitative data and conducting thematic analysis for qualitative documents; (5) Integration and triangulation—comparing findings across sources to identify convergent patterns and discrepancies; (6) Interpretation and reporting—synthesizing integrated findings within the GFE framework and DIECI cycle context.

INDICATORS AND OPERATIONAL DEFINITIONS

Cognitive indicators: cohort GPA and on-time graduation (Y-2, Y-1, Y). Affective indicator: proportion rated "very good" in each service dimension. Psychomotor indicators: operational parameters of micro-teaching (practice frequency, pass threshold, skills coverage, facilities). Graduate-outcome indicators: tracer coverage, total placement (employment/entrepreneurship/further study/Teacher-Certification Department [TCP]), time-to-job, and distribution of post-graduation activities. The 40–30–30 distribution underpinned Table 1.

DATA ANALYSIS

Quantitative analyses were conducted using SPSS 26 (IBM Corporation, 2019). I computed descriptive statistics including means (cohort GPA), proportions (satisfaction ratings, tracer coverage, placement rates, activity distribution), and frequency distributions for micro-teaching parameters. Year-over-year trend analyses (Y-2–Y-1–Y) were performed for GPA and on-time graduation rates to identify temporal patterns. Cross-tabulation analyses examined relationships between assessment domains and institutional indicators.

Qualitative document analysis followed systematic procedures (Bowen, 2009): (1) familiarization with QA documents and institutional reports; (2) identification of key themes related to assessment implementation, DIECI cycle processes, and quality assurance mechanisms; (3) coding of relevant sections

using a priori categories (DLO–ELO alignment, assessment standardization, moderation practices) and emergent themes; (4) triangulation of qualitative findings with quantitative results to identify convergent and divergent patterns.

Following recent guidance on holistic learning assessment and complementary holistic–analytic scoring, I prioritized Department-level holistic indicators while using analytic sub-skill evidence diagnostically where available (Singh et al., 2019; Zlatkin-Troitschanskaia et al., 2019). Findings appear as Tables 1–5 in the Conclusion and Discussion section to maintain traceability between indicators, definitions, and results.

VALIDITY, RELIABILITY, AND TRIANGULATION

Triangulation was systematically implemented across multiple dimensions to enhance validity and credibility (Denzin, 2017; Flick, 2018). Source triangulation integrated evidence from five independent data sources: (1) academic records (GPA, graduation rates), (2) QA documentation (DLO–ELO mapping, monitoring reports), (3) student satisfaction surveys, (4) graduate tracer data, and (5) micro-teaching artifacts. Method triangulation combined quantitative descriptive analyses with qualitative document analysis to examine the same phenomena through different methodological lenses (Creswell & Plano Clark, 2016). Data triangulation involved collecting data across three academic years (Y-2, Y-1, Y) to identify consistent patterns over time. Investigator triangulation was achieved through independent review of findings by multiple researchers and cross-validation with Department QA personnel.

Internal validity was strengthened through systematic cross-source verification: quantitative findings from academic records were compared with qualitative evidence from QA documents, and survey responses were triangulated with documented service activities. Process reliability was supported by standardized procedures: common rubrics across courses, systematic DLO/CLO monitoring protocols, and consistent data collection procedures following QA SOPs.

Reliability coefficients were not computed due to the absence of item-/rater-level data files. Planned reliability analyses include: (1) Cronbach's alpha for internal consistency of satisfaction survey scales (Parasuraman et al., 1988) (2) two-way random-effects, absolute-agreement Intraclass Correlation Coefficients—ICC(2,1) for single-rater reliability and ICC(2,k) for mean-rater reliability—for micro-teaching rubric scores (Zlatkin-Troitschanskaia et al., 2019). These will be calculated once item-level survey responses and rater-level rubric scores become available for comprehensive psychometric evaluation.

SOFTWARE

Quantitative descriptive analyses were conducted using IBM SPSS Statistics version 26 (IBM Corporation, 2019) to ensure replicability and maintain an auditable analytic trail. SPSS 26 was selected for its comprehensive statistical capabilities, established reliability in educational research, and compatibility with institutional data formats. All analyses followed documented procedures to enable replication and verification.

STATEMENT ON COMPANION MANUSCRIPT (DATA OVERLAP)

This article draws on the same institutional aggregates (Y-2–Y) as a related published study of in-course holistic assessment (40–30–30) using Goal-Free Evaluation (Samin et al., 2026). That Ijaz Arabi article evaluates classroom/course-level assessment design and its association with academic attainment and graduate outcomes. The present manuscript does not re-analyze that assessment architecture; instead, it examines Department-level indicators within the Determination, Implementation, Evaluation, Control, and Improvement (DIECI) quality-assurance cycle. Shared descriptive statistics (e.g., GPA bands, on-time graduation, tracer coverage, placement, and time-to-job) may overlap because both papers draw on the same institutional LED records; analytical focus, tables, and contributions remain distinct. The research design is described above and is appropriate for the purpose of the study.

CONCLUSION and DISCUSSION

PERFORMANCE OVERVIEW

Guided by GFE, I prioritized what the data show rather than stated targets. The Department demonstrates robust performance across cognitive, affective, and psychomotor domains with clear links to

institutional indicators. Cohort GPA stayed high (Y-2: 3.73; Y-1: 3.69; Y: 3.78); on-time graduation reached 52% in Y; at least three-quarters of students rated services “very good” across dimensions; the micro-teaching ecosystem operated to specification (≥ 4 practices/semester; ≥ 10 core skills; B- threshold); and graduate outcomes were strong (tracer coverage 92.86%; placement 97.5%; mean time-to-job 4.8 months; activity distribution: education 76.9%, entrepreneurship 10.3%, further study 10.3%, teacher-certification 0%). Tables 2–5 summarize these results; Table 1 outlines the framework and data sources.

FINDINGS

Table 1. Assessment framework and Department-level indicators (mapping and sources)

Domain	Primary indicators	Primary sources
Cognitive	Cohort mean GPA; on-time graduation (%)	Academic records; QA monitoring
Affective	Proportion “very good” by service dimension	Institutional satisfaction survey (N=179)
Psychomotor	Micro-teaching frequency; skills coverage; pass threshold	Micro-teaching rubrics and labs
Graduate outcomes	Tracer coverage; placement; time-to-job; activity distribution	University tracer (coverage 42; responses 39)

Note. The Department’s 40–30–30 weighting (cognitive–affective–psychomotor) anchors course-level assessment and feeds institutional indicators.

Table 2. Cohort GPA and on-time graduation by year

Year	Mean GPA	On-time graduation (%)
Y-2	3.73	39.13
Y-1	3.69	39.53
Y	3.78	52.00

Note. Y-2 and Y-1 on-time values are from the main institutional report; Y is from the Department’s verified records.

Table 3. Student satisfaction (N=179): proportion “very good” by dimension

Dimension	Very good (%)
Reliability	76.20
Responsiveness	75.88
Assurance	76.86
Empathy	76.28
Facilities	77.44

Note. Institutional survey across five service dimensions; categories follow institutional standards.

Table 4. Graduate tracer outcomes (coverage 42; responses 39)

Indicator	Value
Coverage (% of target cohort)	92.86
Total placement (work/entrepreneurship/further study/TCP) (%)	97.50
Mean time-to-job (months)	4.8
Post-graduation activity distribution (%)	Education 76.9; Entrepreneurship 10.3; Further study 10.3; TCP 0.0

Note. University-coordinated tracer aligned with institutional KPIs; distribution sums to ~100%.

Table 5. Micro-teaching operational parameters

Parameter	Department standard
Practice frequency per semester	≥4 sessions
Skills coverage (basic teaching skills)	≥10 skills
Passing threshold	Minimum B-
Facilities and support	Dedicated micro-teaching and language labs

Note. Parameters are standardized at Department level to support comparability and auditability.

COGNITIVE DOMAIN: ACADEMIC PERFORMANCE AND EFFICIENCY

High and steady cohort GPA (Table 2) signals durable academic quality. Achieving 52% on-time completion in Y under a 144-credit plan reflects throughput efficiency and consistent academic governance. These patterns are coherent with OBE's constructive alignment linking DPO–CLO–tasks–assessment (Biggs, 2003; Biggs & Collis, 2014; Harden, 2007; Spady, 1994; Yadav et al., 2024).

AFFECTIVE DOMAIN: SERVICE EXPERIENCE AND ACADEMIC CLIMATE

Roughly 76–77% of respondents rated reliability, responsiveness, assurance, empathy, and facilities as “very good” (Table 3). Protection mechanisms (SVPH), counselling, comprehensive OSH, and an Islamic–Malay culture likely sustain comfort, trust, and persistence. Conceptually, service quality and procedural clarity align with student engagement and retention (Petrušić, 2023; Theodoridou, 2021).

PSYCHOMOTOR DOMAIN: PRACTICE-BASED READINESS TO TEACH

A codified micro-teaching regime—≥4 practices per semester, ≥10 foundational skills, B- threshold, dedicated labs—supports automatization of teaching skills and fluent oral communication (Table 5). Transfer is evident in external outputs (25 awards in five years; 18 copyrights; 23 ISBN books; 7 publications), consistent with experiential learning and structured, repeated practice, complemented by self-regulated and collaborative strategies (Kolb et al., 2014).

GRADUATE OUTCOMES AND TRACEABILITY

The tracer covered 39/42 graduates (92.86%), with 97.5% placed in employment/entrepreneurship/further study and a mean time-to-job of 4.8 months (Table 4). Predominant placement in education (~77%) indicates labor-market alignment; integrating knowledge, dispositions, and demonstrable performance appears to smooth school-to-work transitions (David, 2025).

RESOURCES SUPPORTING QUALITY

DIECI governance is operationalized via internal academic audits and management reviews with semesterly monitoring. Typical annual allocations approximate IDR 2.884B (education), IDR 300.96M (research), and IDR 141.27M (community service). Facilities comprise 13 classrooms (capacity 520), micro-teaching and language labs (37 PCs/38 headsets), and integrated IT (78 systems, SSO, 2 Gbps, 403 access points). These enable stable academic performance and evidence-based evaluation.

THEORETICAL DISCUSSION: HOLISTIC ASSESSMENT DESIGN

A 40–30–30 portfolio linked to DLO–CLO appears to enhance the relevance and utility of learning evidence (Biggs, 2003; Biggs & Collis, 2014). My findings support this theoretical proposition, demonstrating that balanced domain weighting (cognitive–affective–psychomotor) within a department-level quality assurance cycle produces sustained academic performance and favorable graduate outcomes. This extends previous research that primarily examined holistic assessment at the course level (Singh et al., 2019) by demonstrating scalability to Department-level indicators.

Syntheses emphasize balanced domains and standardized processes so that results remain meaningful and auditable (Singh et al., 2019; Yadav et al., 2024). My study contributes empirical evidence supporting this emphasis: the standardized micro-teaching parameters (≥4 practices/semester, ≥10 skills, B- threshold) and systematic DLO–CLO monitoring creates auditable evidence chains that previous studies have recommended but rarely implemented at Department scale (Lugay et al., 2020; Mitra & Wadegaonkar, 2024). Combining

holistic and analytic scores strengthens diagnostic feedback and decision reliability (Zlatkin-Troitschanskaia et al., 2019). While my current study prioritizes Department-level holistic indicators, the micro-teaching rubrics provide analytic sub-skill evidence that supports targeted improvement—a blended approach that aligns with contemporary assessment best practices but is uniquely operationalized within a formal QA cycle in my context.

DIGITAL TRANSFORMATION OF ASSESSMENT

Technology-enabled integration can accelerate formative feedback and process analytics, contingent on robust data governance (Anderson et al., 2024; Qiao & Fu, 2022). For psychomotor–affective evidence, multimodal analytics and structured station-based/oral exams such as Multiple Mini Viva (MMV) add objectivity and triangulation (Hadie et al., 2022; Zhang & Wang, 2024). HAF’s participatory-inclusive principles translate well to language disciplines focused on authentic evidence (Mitra & Wadegaonkar, 2024).

COMPANION-PAPER NOTE

A classroom-/course-level analysis of the 40–30–30 holistic assessment design is reported in Samin et al. (2026) in *Ijaz Arabi: Journal of Arabic Learning*. Here I advance the Department-level QA (DIECI) perspective and policy implications not addressed there.

CULTURAL CONTEXTUALIZATION

Integrating Islamic–Malay values into assessment and services—without compromising international standards—improves contextual fit and acceptability. This finding aligns with Vygotsky’s socio-cultural theory, where learning and assessment are embedded within cultural practices and social interactions (Vygotsky, 1978). My study demonstrates how culturally responsive assessment design, grounded in Islamic–Malay pedagogical principles, contributes to sustained satisfaction ($\geq 75\%$ “very good” ratings) and positive academic outcomes—supporting the theoretical proposition that assessment effectiveness is enhanced when aligned with learners’ cultural contexts.

Appropriate cultural adaptation is associated with stronger motivation, engagement, and sustained outcomes (Conradty & Bogner, 2022; Handrianto et al., 2025). My findings extend this literature by providing Department-level evidence of how cultural integration operates within formal quality assurance cycles, whereas previous studies typically examined cultural adaptation at the classroom or individual level (Handrianto et al., 2025; Theodoridou, 2021). The observed pattern—where service dimensions reflecting Islamic–Malay values (assurance, empathy) show consistently high ratings (76–77%)—suggests that cultural alignment enhances not only student satisfaction but also institutional climate, potentially contributing to the observed academic performance and graduation outcomes.

CHALLENGES AND LIMITATIONS

(1) Reliability: Cronbach’s alpha (surveys) and ICC (micro-teaching rubrics) were not computed due to unavailable item/rater files; planned models are two-way random-effects, absolute-agreement—ICC(2,1) and ICC(2,k). (2) Standardization and moderation: cross-course assessor calibration and an integrated rubric bank are required. (3) Process automation: a PLO–CLO–tasks–scores–outcomes analytics dashboard is needed for early anomaly detection. (4) Longitudinal: periodic tracking is required to estimate domain contributions to institutional indicators (GPA, on-time graduation, tracer). (5) Cross-report consistency: Y-2/Y-1 on-time rates (39.13%; 39.53%) exist in the main report; integration into Department QA should be strengthened.

PRACTICAL IMPLICATIONS

Priorities include: (i) standardizing rubrics across practice-based courses and scheduling assessor moderation; (ii) building integrated data pipelines and OBE analytics dashboards; (iii) expanding school/community practice partnerships to enrich authentic evidence; (iv) adopting digital workflows (e-portfolio, rubric banks, video review) to speed feedback; (v) conducting reliability (alpha/ICC) and longitudinal studies on domain contributions; and (vi) strengthening assessor assessment literacy via continuous professional development.

Viewed through the GFE lens, a DLO–CLO-aligned holistic portfolio implemented within the DIECI cycle and supported by adequate resources corresponds with sustained high GPA, improved on-time completion in Y (52%), $\geq 75\%$ “very good” service ratings, a standardized micro-teaching regime, and strong

graduate outcomes (tracer coverage 92.86%; placement 97.5%; time-to-job 4.8 months). These patterns are consistent with constructive alignment and contemporary assessment syntheses, while indicating clear needs for quantified reliability, systematic moderation, and longitudinal analytics.

Interpreting Tables 2–5 jointly, the consistency of academic attainment, affective climate, and graduate outcomes suggests a coherent pattern rather than isolated gains.

These findings converge with OBE syntheses reporting efficiency–quality gains when outcomes, tasks, and rubrics are coherently aligned (Biggs & Collis, 2014; Yadav et al., 2024). However, my study extends this literature by demonstrating these gains at the Department level through a formal quality assurance cycle (DIECI), whereas previous studies primarily focused on course-level alignment (Constantinou, 2020; Mufanti et al., 2024). My findings align with service-climate evidence linking assurance and facilities to engagement and persistence (Petrušić, 2023; Theodoridou, 2021), but uniquely document how culturally embedded service dimensions (Islamic–Malay values) contribute to sustained satisfaction in faith-based higher education contexts—a dimension less explored in previous research.

The observed employability signals align with reports where authentic performance evidence is embedded across courses and documented through shared rubrics (Kolb et al., 2014; Richards & Rodgers, 2014). However, my study differs from previous research in three key ways: (1) I employ GFE to minimize goal bias, whereas most employability studies rely on goal-oriented evaluation designs; (2) I connect micro-teaching evidence directly to Department-level indicators (GPA, graduation rates) and graduate outcomes (tracer data), creating an auditable evidence chain that previous studies rarely document; (3) I operationalize standardized micro-teaching parameters (≥ 4 practices/semester, ≥ 10 skills, B- threshold) as psychomotor evidence within a formal QA cycle, whereas previous studies typically report micro-teaching as isolated classroom practice without systematic Department-level integration.

A plausible mechanism is the cumulative effect of repeated, rubric-guided micro-teaching that accelerates automatization and metacognitive control, increasing transfer to school practice (Kolb et al., 2014). This mechanism aligns with experiential learning theory, where structured practice cycles (plan–teach–review–reteach) develop professional competencies through reflective engagement (Kolb et al., 2024). My findings extend this theoretical framework by demonstrating how standardized micro-teaching parameters, embedded within Department-level quality assurance, create systematic pathways from skill development to employability outcomes—a connection that previous studies have suggested but rarely documented with institutional evidence (Richards & Rodgers, 2014).

Alternative explanations include concurrent institutional initiatives and labor-market shifts; however, the stability across Y-2–Y and triangulation with QA documents reduce the likelihood that the results are policy-only artifacts. The consistency of high GPA (3.69–3.78), improved graduation rates (39% to 52%), and sustained satisfaction ($\geq 75\%$ "very good" across dimensions) across multiple years suggests systematic rather than episodic effects. This pattern differs from studies reporting short-term intervention effects (Hadie et al., 2022) and aligns more closely with comprehensive Department-level reform initiatives (Yadav et al., 2024), though my GFE approach provides more objective evidence of effectiveness than goal-oriented evaluations typically yield.

Internal validity remains bounded by the absence of item/rater files; planned Cronbach's alpha for surveys and ICC(2,1)/(2,k) for rubric scores will quantify measurement consistency. External validity and scalability rest on prerequisites—shared rubric banks, assessor moderation, laboratory access, and video-data governance. Equity requires parallel submission modes and targeted support for students with limited digital access.

Department dashboards should present indicator uncertainty (e.g., 95% CIs in Tables 2–5) alongside action thresholds to enable earlier anomaly detection and targeted remediation. Aligning DIECI review moments with course-level analytics can institutionalize these feedback loops and reduce cycle time from evidence to action.

GENERALIZABILITY AND TRANSFERABILITY

Although the setting is Arabic language education, the underlying mechanisms—Department-level constructive alignment, standardized rubrics with moderation, and auditable evidence flows—are relevant to cognate teacher-education and other performance-based Departments. Key enablers include a stable CO-PO map, shared rubric banks, assessor preparation, authentic practice contexts (or laboratories), and baseline data

governance. Where resources are limited, a minimal-viable pathway—limited rubric set, double-rating a subsample, simple dashboards—can stage adoption while maintaining credibility.

LIMITATIONS AND THREATS TO INFERENCE

My main limitations are the absence of item/rater-level files to compute reliability (Cronbach's alpha; ICC) and a short time-series for estimating domain-level contributions. Even so, triangulation across academic records, QA documents, surveys, and tracer data adds confidence to the inferences.

Implications and future work include: (1) cross-course rubric standardization with scheduled assessor moderation; (2) computation of reliability (alpha; ICC[2,1]; ICC[2,k]) and publication of QA procedures; (3) development of an OBE analytics dashboard linking DLO–CLO–tasks–scores–outcomes; (4) expansion of practice partnerships and e-portfolios to deepen authentic evidence; and (5) longitudinal or quasi-experimental studies testing interventions (e.g., MMV, multimodal analytics) on academic performance and employability.

Educational assessment has long been trapped in rigid, compliance-driven frameworks that reduce learning to mere numerical indicators. My research with the Arabic Language Education Department at Islamic University of Riau offers a compelling alternative: Goal-Free Evaluation (GFE), a transformative approach that breathes life into Department-level quality assurance within Outcome-Based Education. By systematically examining educational performance through a comprehensive 40–30–30 holistic assessment framework, I discovered something profound—that true educational understanding emerges when I look beyond predefined targets and listen to the nuanced stories my data tells.

What emerged was not just a methodology, but a window into educational complexity. My findings reveal that holistic assessment models, anchored within the DIECI quality assurance cycle, can capture the multidimensional nature of student development. The data spoke volumes: consistent high GPAs, improved graduation rates, and robust student satisfaction were not isolated achievements, but interconnected indicators of a learning ecosystem working in harmony. Critically, I demonstrated how technological infrastructure and culturally responsive design could transform assessment from a mechanical process to a meaningful dialogue—integrating Islamic–Malay pedagogical principles with international academic standards to create learning environments that are both rigorous and deeply human.

Yet, my research is not an endpoint, but an invitation. It challenges educational leaders to reimagine assessment not as a compliance mechanism, but as a dynamic process of understanding and nurturing student potential. Future investigations must delve deeper—developing sophisticated reliability methods, conducting longitudinal studies, and exploring how my approach might be scaled and adapted. The DIECI-anchored holistic assessment model I propose is more than an academic construct; it is a pathway towards more empathetic, responsive, and genuinely student-centered educational practices. In an era of rapid technological and cultural change, my study illuminates how thoughtful, technology-enabled, and culturally sensitive assessment can be a powerful mechanism for educational transformation.

LIMITATIONS

While my study provides significant insights into holistic assessment within Outcome-Based Education, several methodological constraints warrant careful consideration: the absence of item-level and rater-level data limits comprehensive reliability metrics computation, preventing precise Cronbach's alpha and Intraclass Correlation Coefficient calculations; the relatively short three-year time series restricts definitive conclusions about long-term domain-level contributions; the context-specific nature of my research—focused on an Arabic Language Education Department at a single Islamic University of Riau—raises generalizability questions; and potential inherent biases in self-reported satisfaction surveys and tracer studies may subtly influence my findings. Despite these limitations, my rigorous triangulation across academic records, QA documents, surveys, and tracer data provides substantive confidence in my inferences, while simultaneously inviting further scholarly investigation, methodological refinement, and nuanced understanding of holistic assessment practices.

Acknowledgements: I thank the Quality Assurance unit and the Arabic Language Education Department at Islamic University of Riau.

Declarations

Conflict of Interest

No potential conflicts of interest were disclosed by the author with respect to the research, authorship, or publication of this article.

Ethics Approval

The formal ethics approval was granted by the Research Ethics Committee of Islamic University of Riau. I conducted the study in accordance with the Helsinki Declaration in 1975.

Funding

No specific grant was given to this research by funding organizations in the public, commercial, or not-for-profit sectors.

Research and Publication Ethics Statement

The study was approved by the author's university ethics committee of the Islamic University of Riau Number 887/KONTRAK/P-NK-PD/DPPM-UIR/11-2024. Hereby, I as the author consciously assure that for the manuscript "Curriculum, Teaching Methods, and Assessment in Arabic Language Education: A Department-Level Quality Assurance Evaluation"

the following is fulfilled:

- This material is the author's own original work, which has not been previously published elsewhere.
- The paper reflects the author's own research and analysis in a truthful and complete manner.
- The results are appropriately placed in the context of prior and existing research.
- All sources used are properly disclosed.

Contribution Rates of Authors to the Article

The author provides sole contribution to this work.

REFERENCES

- Al-Mubaid, H., & Bettayeb, S. (2017). Intervention-based method for improving learning and critical thinking. In *2017 Computing Conference* (pp. 1232–1237). IEEE. <https://doi.org/10.1109/SAI.2017.8252247>
- Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (pp. 413–425). Association for Computing Machinery. <https://doi.org/10.1145/3635636.3656204>
- Biggs, J. (2003). Aligning teaching and assessing to course objectives. *Teaching and Learning in Higher Education: New Trends and Innovations*, 1–9.
- Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.
- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27–40. <https://doi.org/10.3316/QRJ0902027>
- Conradty, C., & Bogner, F. X. (2022). Education for sustainable development: How seminar design and time structure of teacher professional development affect students' motivation and creativity. *Education Sciences*, 12(5), Article 296. <https://doi.org/10.3390/educsci12050296>
- Constantinou, C. S. (2020). A reflexive GOAL framework for achieving student-centered learning in European higher education: From class learning to community engagement. *Societies*, 10(4), Article 75. <https://doi.org/10.3390/soc10040075>
- Creswell, J. W., & Plano Clark, V. L. (2016). *Designing and conducting mixed methods research*. SAGE Publications.
- David, J. O. (2025). Balancing growth, inclusivity, and technology: Rethinking student-centered learning in South African higher education. *International Journal of Learning, Teaching and Educational Research*, 24(4), 127–153. <https://doi.org/10.26803/ijlter.24.4.7>

- Denzin, N. K. (2017). *The research act: A theoretical introduction to sociological methods*. Routledge. <https://doi.org/10.4324/9781315134543>
- Flick, U. (2018). Triangulation in data collection. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 527–544). SAGE Publications.
- Hadie, S. N. H., Kasim, F., Mohd Ismail, Z. I., Asari, M. A., Mohd Yusof, N. A., Simok, A. A., Shamsuddin, S. A., & Mohd Amin, M. S. I. (2022). Multiple mini viva as an online assessment tool for postgraduate anatomy programme: The USM's experience amidst COVID-19 pandemic. *Education in Medicine Journal*, 14(4), 99–112. <https://doi.org/10.21315/eimj2022.14.4.9>
- Handrianto, C., Solfema, S., & Jusoh, A. J. (2025). Cultural adaptation and academic success: Exploring the experiences of Indonesian students in Malaysian universities. *Journal of Ethnic and Cultural Studies*, 12(3), 251–280. <https://doi.org/10.29333/ejecs/2496>
- Harden, R. M. (2007). Outcome-based education: The future is today. *Medical Teacher*, 29(7), 625–629. <https://doi.org/10.1080/01421590701729930>
- Kim, A., & Su, Y. (2024). How implementing an AI chatbot impacts Korean as a foreign language learners' willingness to communicate in Korean. *System*, 122, Article 103256. <https://doi.org/10.1016/j.system.2024.103256>
- Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2014). Experiential learning theory: Previous research and new directions. In R. J. Sternberg & L.-F. Zhang (Eds.), *Perspectives on thinking, learning, and cognitive styles* (pp. 227–247). Routledge. <https://doi.org/10.4324/9781410605986-9>
- Kolb, L., Panzer, M., & Gronau, N. (2024). Assessing generalizability in deep reinforcement learning based assembly: A comprehensive review. *Journal of Intelligent Manufacturing*. Advance online publication. <https://doi.org/10.1007/s10845-024-02546-5>
- Lugay, C. I. P., Agustin, B. E., Codico, M. A. E., Tapis, C.-A. P., & Tolentino, E. (2020). A design of an automated assessment process for the outcomes-based education in the industrial engineering department of the faculty of engineering at the University of Santo Tomas. *Journal of Physics: Conference Series*, 1529(3), Article 032046. <https://doi.org/10.1088/1742-6596/1529/3/032046>
- Mitra, G., & Wadegaonkar, A. (2024). Multidimensional holistic assessment framework (HAF): A case study of exploring the discourses from elementary school teachers. *International Electronic Journal of Elementary Education*, 17(1), 51–69. <https://doi.org/10.26822/iejee.2024.363>
- Mufanti, R., Carter, D., & England, N. (2024). Outcomes-based education in Indonesian higher education: Reporting on the understanding, challenges, and support available to teachers. *Social Sciences & Humanities Open*, 9, Article 100873. <https://doi.org/10.1016/j.ssaho.2024.100873>
- Munzenmaier, C., & Rubin, W. N. (2013). *Perspectives Bloom's taxonomy: What's old is new again*. The eLearning Guild. https://publicservicesalliance.org/wp-content/uploads/2013/04/guildresearch_blooms2013.pdf
- Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1), 12–40.

- Petrušić, I. (2023). Student-centred learning in the context of external quality assurance. *Medicina Fluminensis*, 59(4), 427–436. https://doi.org/10.21860/medflum2023_309615
- Qiao, Y., & Fu, H. (2022). Study on OBE teaching concept in the context of deep learning for the construction of university mathematics microcourses. *Computational Intelligence and Neuroscience*, 2022, Article 6860842. <https://doi.org/10.1155/2022/6860842>
- Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge University Press.
- Samin, S. M., Jaafar, A., Supriady, H., Pebrian, R., Zulkifli, A., & Yunita, Y. (2026). Measuring what matters: Goal-free evaluation of holistic assessment in Arabic language education. *Ijaz Arabi: Journal of Arabic Learning*, 9(1), 278–290. <https://doi.org/10.18860/ijazarabi.v9i1.36779>
- Schön, S., Gasplmayr, K., Ebner, M., Alcober, J., Hoppe, C., Koschutnig-Ebner, M., Da Silva, F. M., & Taraghi, B. (2024). The development of “teaching management patterns” from the perspective of IT infrastructure as a tool for consulting and further development in a European university alliance. In T. Bastiaens (Ed.), *Proceedings of EdMedia + Innovate Learning* (pp. 35–47). Association for the Advancement of Computing in Education. <https://www.learntechlib.org/primary/p/224515/>
- Scriven, M. (1972). Objectivity and subjectivity in educational research. *Teachers College Record*, 73(5), 94–142. <https://doi.org/10.1177/016146817207300505>
- Singh, C. K. S., Ong, E. T., & Singh, T. S. M. (2019). Redesigning assessment for holistic learning. *Journal of Social Sciences Research*, 5(3), 620–625. <https://doi.org/10.32861/jssr.53.620.625>
- Spady, W. G. (1994). *Outcome-based education: Critical issues and answers*. American Association of School Administrators. <https://files.eric.ed.gov/fulltext/ED380910.pdf>
- Tashakkori, A., & Teddlie, C. (Eds.). (2010). *SAGE handbook of mixed methods in social & behavioral research* (2nd ed.). SAGE Publications.
- Theodoridou, S. (2021). The importance of student-centered learning in the development of critical thinking (full text in Greek). *Academia*, 23–24, 226–245.
- Thompson, A. S., & Asanov, E. (2024). “Nonnative? Next!” Native-speakerism in world language job advertisements. *Studies in Second Language Learning and Teaching*, 14(1), 49–74. <https://doi.org/10.14746/ssllt.42371>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Yadav, P., Tomar, A., Soni, R., Dubey, G., Bhargava, C. P., & Kumari, J. (2024). Metrics with performance indicators used for calculating the mapping strength of course outcome with program outcome. *Journal of Engineering Education Transformations*, 38(2), 85–95. <https://doi.org/10.16920/jeet/2024/v38i2/24192>
- Youker, B. W. (2013). Goal-free evaluation: A potential model for the evaluation of social work programs. *Social Work Research*, 37(4), 432–438. <https://doi.org/10.1093/swr/svt032>
- Zhang, Z., & Wang, W. (2024). Enhancing dance education through convolutional neural networks and blended learning. *PeerJ Computer Science*, 10, Article e2342. <https://doi.org/10.7717/peerj-cs.2342>

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology*, 89(3), 468–484. <https://doi.org/10.1111/bjep.12286>