

A Multi-Domain Framework for Auditing Examination Quality Assurance in Higher Education: Evidence from Teacher Education in The Gambia

Sheriff Ceesay¹, Lamin L. Kanteh², Omar Ceesay³

¹Directorate of Research and Development, University of Education, The Gambia (UEG), Banjul, The Gambia

²Social and Environmental Studies Department, University of Education, The Gambia (UEG), Banjul, The Gambia

³Commerce Department, University of Education, The Gambia (UEG), Banjul, The Gambia

Email: sheriffceesay@gambiacollege.edu.gm

How to cite this paper: Ceesay, S., Kanteh, L. L., & Ceesay, O. (2026). A Multi-Domain Framework for Auditing Examination Quality Assurance in Higher Education: Evidence from Teacher Education in The Gambia. *Open Journal of Social Sciences*, 14, 646-672.

<https://doi.org/10.4236/jss.2026.149039>

Received: July 28, 2026

Accepted: September 21, 2026

Published: September 24, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This paper evaluates key quality-assurance dimensions: validity, reliability, examination quality control, item quality control, administrative security, and moderation feedback at the University of Education, The Gambia School of Education (UEG-SoE). Using two institutionally developed instruments, the Exam Committee Questionnaire (ECQ; N = 29; $\alpha = 0.87$) and the Staff Assessment Practices Questionnaire (SAPQ; N = 68; $\alpha = 0.85$), the study found strong perceived face and content alignment (composite means = 4.0138 and 4.0483), moderate perceived construct alignment (3.1517), and generally positive perceptions of marking-related consistency (3.8103) anchored in marking-guide practices. Examination quality control was found to be generally robust (3.7356), though difficulty-level consistency across departments was weak (2.8621). Item quality control was positive (3.5733), with strong review for clarity (4.3103) and clear intent (4.1724), but question repetition remains a concern (2.3793). Sound administrative quality control was manifested (3.6802), with strong secure storage and access controls (≈ 4.17), but a critical training gap on confidentiality/security (2.3448). Moderation compliance was high (98.5% submitted), yet feedback was inconsistent, with 53% reporting rarely/never receiving feedback; where provided, feedback was rated useful/very useful by 83.8%, most often via one-on-one consultations (36.7%), although staff preferred email (32.4%) and written comments (19.1%). The study recommends standardised difficulty calibration, formalised and documented feedback protocols, item banking with exposure tracking, cognitive blueprints, and targeted training to close gaps and align with competency-based policy. Beyond the institutional case, the study

proposes a practical multi-domain framework for auditing examination quality assurance in higher education, spanning alignment, marking consistency, examination and item quality control, administrative security, and moderation feedback. The findings highlight implementation tensions likely to extend beyond the case institution, particularly the gap between procedural compliance and developmental feedback, the challenge of achieving cross-department comparability in examination difficulty, and the need to align assessment security systems with staff training and access logging. These insights may inform assessment reform in comparable teacher education and resource-constrained higher education settings.

Keywords

Examination Quality Assurance, Moderation, Assessment Quality, Teacher Education, Higher Education, Institutional Assessment, The Gambia

1. Introduction

Assessment quality in higher education is a cornerstone of academic standards, fairness, and public trust, with particular salience in teacher education, where institutional practices model the assessment literacy that future teachers will enact in schools. Contemporary frameworks emphasise that quality is multidimensional, spanning validity (the extent to which interpretations and uses of scores are supported), reliability and consistency, moderation and standard-setting, item quality assurance, administrative security, and effective feedback loops for continuous improvement (Shaw & Crisp, 2011; AERA, APA, & NCME, 2014; Kane, 2013). Within outcome-based education, constructive alignment positions assessment as one component of a coherent system linking intended learning outcomes (ILOs), teaching/learning activities (TLAs), and assessment tasks, with cognitive demand calibrated against recognised taxonomies such as Bloom's (Biggs & Tang, 2011; Anderson & Krathwohl, 2001). This calibration is not only a psychometric concern but a curricular one: what is assessed shapes what is taught and how students' study.

In The Gambia, recent reforms prioritise competency-based curricula, diagnostic assessment, and institutionalised continuous assessment across school phases (Ministry of Basic and Secondary Education [MoBSE], 2015; UNESCO, 2018; IIEP-UNESCO, 2021). The University of Education, The Gambia (UEG) School of Education (SoE) thus occupies a system-critical position: strengthening the validity, reliability, and moderation of its assessments has consequences for teacher quality throughout the sector. Against this backdrop, the present study evaluates UEG-SoE's assessment quality across six dimensions: face, content, and construct validity; reliability; examination and item quality control; and administrative security, alongside moderation compliance and feedback practices. We triangulate the perspectives of Examination Committee Questionnaire (ECQ; $N = 29$; Cronbach's $\alpha = 0.87$) respondents and Staff Assessment Practices Questionnaire (SAPQ; N

= 68; $\alpha = 0.85$) respondents, offering an institution-wide snapshot of strengths and risks. We combine two internally consistent instruments to capture staff and committee perceptions across six quality-assurance domains, including perception-based proxies for training coverage, access control, and item repetition. In doing so, this work contributes empirical evidence from a low- and middle-income context where comprehensive assessments of higher education quality assurance (QA) systems remain comparatively rare, and it aligns recommendations with international best practice and local policy priorities.

Although the present study is institution-specific, it is intended as an analytically informative case rather than a purely local audit. Teacher education institutions occupy a system-critical position because their assessment practices not only certify student achievement but also model the assessment literacy that graduates may later reproduce in schools. In low- and middle-income and resource-constrained contexts, universities are frequently required to strengthen examination quality assurance without access to extensive psychometric infrastructure, making practical, institution-wide audit approaches especially relevant. The University of Education, The Gambia School of Education, therefore, provides a useful case through which to examine how validity-related alignment, marking consistency, moderation, item control, and examination security are operationalised under reform conditions associated with competency-based policy.

This study contributes in three ways. First, it operationalises examination quality assurance as a multi-domain institutional system encompassing assessment alignment, marking consistency, examination quality control, item review quality, administrative security, and moderation feedback. Second, it provides empirical evidence from a teacher education institution in The Gambia, a context underrepresented in the higher education assessment literature. Third, it identifies implementation tensions that are likely to be relevant beyond the focal institution, including the coexistence of strong moderation compliance with weak feedback loops, fair grading perceptions with inconsistent difficulty calibration across departments, and secure storage practices with inadequate staff training on confidentiality and security.

Accordingly, the purpose of this study is not merely to report the status of examination practices at one institution, but to use the UEG-SoE case to examine how multiple quality-assurance domains interact within a teacher education assessment system and to derive practical lessons that may inform comparable institutional contexts. The study asks how respondents perceive the strengths and vulnerabilities of the current system across alignment, marking consistency, examination and item quality control, administrative security, and moderation feedback, and what these patterns imply for institutional assessment reform.

2. Literature Review

2.1. Validity and Validation Frameworks

Validity concerns the degree to which evidence and theory support the interpre-

tations of test scores for proposed uses (AERA, APA, & NCME, 2014; Kane, 2013). While face and content validity speak to perceived appropriateness and representativeness, construct validity requires that assessments capture the intended competencies and that score-based decisions are defensible given the underlying claims. According to Kane (2013), argument-based approaches emphasise building and evaluating an interpretive/use argument with multiple sources of evidence, content, response processes, internal structure, relations to other variables, and consequences. In programme contexts, constructive alignment operationalises validity by ensuring that ILOs, TLAs, and assessments are coherent and that cognitive demand matches intended outcomes; superficial alignment (high-level verbs with low-level tasks) undermines construct validity (Biggs & Tang, 2011; Mor & Erşen, 2023).

2.2. Reliability, Standardisation, and Marking Quality

Reliability encompasses score consistency across markers, occasions, and parallel forms. In higher education, inter-marker reliability is improved by explicit criteria, calibrated rubrics, exemplars, and structured moderation (Crisp, Johnson, & Constantinou, 2019; Sadler, 2013). Marking guides can reduce subjectivity and support fair differentiation by cognitive level when well specified and used consistently; however, rubric design and marker calibration are prerequisites for realising these benefits (Brookhart, 2018; Prichard et al., 2025). Routine post-assessment checks (e.g., double-marking samples, blind remarking, marker drift monitoring) further bolster reliability and equity across cohorts and departments.

2.3. Moderation and Feedback

Moderation is a suite of pre-, during-, and post-assessment activities that aims to ensure comparability and defensibility of judgments across assessors, cohorts, and sites (University of Birmingham, & University College Birmingham, 2020; University of Edinburgh, 2024). Effective moderation features timely, documented, and actionable feedback to item writers and markers; transparent processes; and calibration exercises that align tacit standards (Prichard et al., 2025; Sadler, 2013). Where feedback is irregular or undocumented, opportunities for item improvement and professional learning are lost (Winstone & Pitt, 2025; Williams, 2024). Emerging work on feedback literacy underscores that both staff and students need capabilities to interpret, use, and seek feedback; formalising feedback channels (e.g., templates, email with tracked actions) supports uptake and organisational learning (Carless & Boud, 2018; Hattie & Timperley, 2007).

2.4. Item Quality Control and Post-Examination Analytics

Item quality is central to validity and reliability, established guidelines address clarity, relevance, avoidance of cueing, and cognitive demand, particularly for selected-response formats (Haladyna, Downing, & Rodriguez, 2002; Downing, 2005;

Haladyna & Rodriguez, 2015). Post-exam analytics quantify performance and discrimination (e.g., facility indices, point-biserials) and test reliability (e.g., Kuder-Richardson Formula 20 (KR-20) and Cronbach's alpha (α)), informing item retention, revision, or retirement (Tavakol & Dennick, 2011). Item banks with metadata (topic, Bloom level, difficulty) and exposure tracking mitigate repetition and leakage risks while enabling intentional cognitive sampling across cycles (Crisp et al., 2019).

2.5. Administrative Quality Control and Assessment Security

Administrative controls, secure storage, role-based access, encryption, access logging, and breach protocols protect the integrity of assessments. Training for staff with access to materials is a critical control; absent or ad hoc training elevates operational risk (Quality Assurance Agency for Higher Education, 2024; Myer & Smith, 2024). As digital workflows expand, policies must address file handling, version control, and traceability, supported by periodic audits and drills to test response capacity.

2.6. Cognitive Alignment and Blueprinting

Blueprinting maps assessment content and cognitive demand to ILOs and ensures planned coverage across Bloom's levels, with higher-order skills weighted appropriately to year/level (Anderson & Krathwohl, 2001; Biggs & Tang, 2011). Without blueprints and calibration, departments can drift towards recall-heavy assessments, reducing comparability and undermining competency-based goals. Stimulus-rich items and authentic tasks are often required to elicit analyse/evaluate/create, supported by rubrics that operationalise complex performances (Pellegrino, Chudowsky, & Glaser, 2001; Crisp et al., 2019).

2.7. A Multi-Domain Institutional Framework for Examination Quality Assurance

Drawing on the literature reviewed above, the present study conceptualises examination quality assurance as a connected institutional system rather than as a set of isolated technical checks. At minimum, such a system includes six interacting domains: (1) alignment-related validity, referring to the extent to which examination tasks appear appropriate, represent intended content, and reflect intended constructs; (2) marking consistency, supported by marking guides, clarity of criteria, and moderation processes; (3) examination quality control, including communication of criteria, fairness, comparability, and appeal mechanisms; (4) item quality control, including clarity, non-duplication, topical range, and error reduction; (5) administrative security, including secure storage, role-based access, logging, breach response, and staff training; and (6) moderation feedback, which links quality assurance to continuous improvement by ensuring that item writers and assessors receive timely, actionable, and documented feedback. Framed this way, examination quality assurance is both a governance function and a learning

system. The value of the present case study lies not only in documenting the status of these domains at UEG-SoE but also in demonstrating a practical structure that other institutions may adapt when reviewing their own assessment systems.

The framework is intended as a practical heuristic rather than a fixed or fully standardised instrument. Its value lies in helping institutions review examination quality assurance as an integrated system that combines alignment, consistency, comparability, item integrity, security, feedback, and continuous improvement. While specific indicators may vary by institutional context, the domains and associated review questions offer a structure that can be adapted for institutional audit, policy development, staff training, or comparative research (**Table 1**).

Table 1. Core domains of a transferable institutional examination quality-assurance framework.

Domain	Purpose	Sample evidence	Priority actions
Assessment alignment	Align exams with intended learning outcomes and cognitive demand	Blueprints, learning outcome maps, Bloom-level distributions	Mandate blueprinting and alignment checks
Marking consistency	Improve the fairness and dependability of scoring	Marking guides, rubrics, exemplars, and double-marking records	Use standard rubrics and calibration
Examination quality control	Improve fairness, transparency, and comparability across units	Approval records, appeal records, and cross-department reviews	Use calibration panels and monitor difficulty patterns
Item quality control	Improve clarity, coverage, and technical quality of questions	Review checklists, duplication scans, and item-bank metadata	Require a second review and item banking
Administrative security	Protect exam materials and maintain traceability	Access logs, encryption records, breach reports	Enforce role-based access and security audits
Moderation compliance and feedback	Ensure moderation is timely, documented, and improvement-oriented	Submission logs, feedback templates, revision archives	Introduce feedback SLAs and archive records

3. Methods

3.1. Design and Setting

The study employed a descriptive, cross-sectional survey design to examine assessment quality assurance practices within the UEG SoE. Two complementary instruments targeted distinct respondent groups: the Examination Committee Questionnaire (ECQ) and the Staff Assessment Practices Questionnaire (SAPQ). The design ensured that respondents' opinions identified their compliance with quality assurance protocols and ascertained the standard required for examination validity and reliability. The study spanned the school's two campuses (Brikama and Basse) and multiple departments, enabling institution-level inferences.

The study was designed as an institutional case study intended to generate both local diagnostic insights and broader lessons about examination quality assurance in teacher education. A case-study approach was appropriate because the research

questions concerned how multiple interrelated quality-assurance processes operate within a real institutional setting. Rather than seeking statistical generalisation across all universities, the study aimed for analytical transferability by documenting a structured quality-assurance examination system that may be informative for comparable institutions operating under similar governance, resource, and reform conditions.

3.2. Participants and Sampling

The eligible population comprised all academic staff involved in the examination setting at UEG-SoE (approximately $N \approx 110$ across the Brikama and Basse campuses) and all members of the Examination and Moderation Committee (approximately $N \approx 35$) used to select academic staff. Purposive sampling was used to select the committee, of which 29 returned questionnaires. For academic staff, a sampling frame was drawn from departmental staff lists, and 68 SAPQ responses were completed. Participation was voluntary and anonymous, and non-response was not followed up to preserve anonymity.

Purposive sampling and simple random sampling were used to select the examination committee and staff, respectively. Examination and Moderation Committee members ($N = 29$) completed the ECQ; academic staff ($N = 68$) completed the SAPQ. Departments represented included Education & Professional Studies, Physical & Health Education, Mathematics, Science, Agriculture, Commerce, English, Home Science, Information and Communication Technology (ICT), Arts, Gender Studies, Social & Environmental Studies, and others. The staff sample was predominantly academic (over 98%), with qualifications ranging from Bachelor's to Doctor of Philosophy (PhD) degrees. Campus participation was approximately 70% Brikama and 30% Basse for the ECQ, consistent with staff distribution. Participation was voluntary and anonymous.

The respondent groups were selected because they occupy distinct but complementary positions within the examination system. Examination and Moderation Committee members are directly involved in question review, moderation, and security processes, making them appropriate informants on institutional quality controls. Academic staff are the primary setters and submitters of examination questions and direct recipients of moderation feedback, making them especially well placed to report on workflow compliance and feedback practices. The sample was therefore aligned with the study's purpose of examining operational strengths and vulnerabilities within the institutional assessment system.

3.3. Instruments

The ECQ (Cronbach's alpha (α) = 0.87) assessed perceived face, content, and construct alignment; marking-related consistency; examination and item quality control; and administrative security. The SAPQ ($\alpha = 0.85$) captured moderation submission practices, frequency and modalities of feedback, perceived usefulness of

feedback, and preferred feedback channels. Both instruments were developed to reflect the study's six quality-assurance domains derived from the literature reviewed in Section 2. Initial item pools were generated from recurring themes in the literature on validity, marking reliability, moderation, item writing, assessment security, and feedback processes. The draft questionnaires were reviewed by subject specialists with experience in higher education assessment and institutional quality assurance to assess wording clarity, domain relevance, and coverage, and revisions were made before administration. Items were rated on a five-point Likert scale (1 = Strongly disagree/Never; 5 = Strongly agree/Always). A priori, a threshold mean of 3.0 denoted acceptable agreement (Boone & Boone, 2012). The instruments were designed primarily for institutional diagnostic use; although internal consistency estimates were acceptable, further validation would be needed before wider cross-institutional deployment.

3.4. Procedures and Data Management

Questionnaires were administered during routine assessment cycles. Responses were recorded digitally and checked for completeness. Descriptive statistics (means, medians, standard deviations, and percentages) were computed by domain and item. Composite/grand means summarised domain-level perceptions. Where appropriate, medians were reported to respect the ordinal nature of Likert data, and item-level standard deviations were used to characterise response dispersion (Boone & Boone, 2012). Free-text comments (where provided) were reviewed to contextualise quantitative patterns. No inferential comparisons between campuses or departments were conducted due to sample sizes. Negatively worded items (e.g., "Examination questions lack clarity"; "Questions were repeated in the examination") were [reverse-coded prior to analysis so that higher scores consistently indicate more favourable practice / or retained in original polarity].

3.5. Ethical Considerations

The study adhered to institutional guidelines for research with human participants. Participation was voluntary, no incentives were offered, and no personally identifying information was collected. Data were stored on secure, access-controlled drives, with access limited to the research team.

3.6. Reporting

This study did not analyse independent operational records (e.g., access logs, training registers, or item-exposure databases). "Operational indicators" such as training coverage, access logging, and question repetition are therefore derived solely from respondents' perceptions on the ECQ/SAPQ. Findings are interpreted against contemporary validity/reliability frameworks and sector guidance on moderation and assessment security (AERA, APA, & NCME, 2014; University of Birmingham, & University College Birmingham, 2020; University of Edinburgh,

2024).

Because the instruments capture respondent perceptions rather than psychometric evidence from scored artefacts, results are reported as perceived alignment and perceived marking consistency; they do not establish score validity or reliability in the psychometric sense. This distinction is especially important for construct validity, which would require additional evidence from assessment artefacts, blueprint reviews, item statistics, and scoring studies.

4. Results

4.1. Demographic Information of Respondents

As shown in **Table 2**, the Examination and Moderation Committee Questionnaire (ECQ) drew responses from 29 participants. In terms of gender, the sample was male-dominated: males accounted for 23 respondents (79%), while females made up only 6 respondents (21%). This imbalance indicates a clear gender skew in the composition of the committee sample.

Table 2. Examination and moderation committee.

Section	Category	Frequency	Percent
Gender	Female	6	21%
	Male	23	79%
	Total	29	100%
Current status	Administrative Staff	1	3%
	Assistant Lecturer	1	3%
	Lecturer	18	62%
	Senior Lecturer	9	31%
	Total	29	100%
Department	Agriculture	2	7%
	Arts	1	3%
	Commerce	2	7%
	Education & Professional Studies	4	14%
	English	2	7%
	Gender Studies	1	3%
	Home Science	2	7%
	Information & Communication Technology	1	3%
	Mathematics	3	10%
	Others	3	10%
	Physical & Health Education	4	14%
	Science	3	10%
	Social & Environmental Studies	1	3%
	Total	29	100%

Continued

Highest academic qualification	B.Ed./B.Sc./B. A	7	24%
	H.T.C/Ad. Dip	1	3%
	M.Ed./M.Sc./M. A	17	59%
	PGD	1	3%
	Ph.D.	3	10%
	Total	29	100%

With regard to the current status, the respondents were spread across four staff categories. Lecturers formed the largest group with 18 respondents (62%), followed by Senior Lecturers with 9 respondents (31%). Assistant Lecturers and Administrative staff were each represented by a single respondent (3% each). This distribution shows that the ECQ sample is centred on the Lecturer grade, supported by a smaller senior cadre, with minimal representation from junior and administrative roles. Taken together with the gender profile, these patterns suggest that gender and rank composition should be considered when interpreting committee feedback and planning moderation practices.

The respondent pool also spanned a wide range of academic departments. The largest contributions came from Education & Professional Studies and Physical & Health Education, with 4 respondents each (14%). These were followed by mid-sized groups in Mathematics, Science, and the “Others” category, each contributing 3 respondents (10%). Smaller clusters were drawn from Agriculture, Commerce, English, and Home Science, with 2 respondents each (7%), while single respondents (3% each) represented Arts, Gender Studies, Information & Communication Technology, and Social & Environmental Studies. Overall, representation was diffused across disciplines, with no single department dominating, although several areas were represented by only one respondent.

Finally, in relation to the highest academic qualification, the results revealed that the minimum qualification among respondents was an advanced diploma in education (H.T.C/Ad. Dip). This suggests that most participating staff had attained at least a first degree, except for one holding a diploma, with many others holding higher qualifications. The greatest proportion of respondents were master’s degree holders (M.Ed./M.Sc./M. A) at 59% (17 respondents), followed by first-degree holders (B.Ed./B.Sc./B. A) at 24% (7 respondents). Ph.D. holders formed the third-largest group at 10% (3 respondents), while single respondents each held a PGD and an advanced diploma (3% each).

4.2. Evaluation by Examination and Moderation Committee Members

Table 3 reports respondents’ perceptions of key examination quality-assurance domains. Because these results are based on staff and committee perceptions rather than direct psychometric evidence, they should be interpreted as institution-level indicators of perceived alignment, consistency, control, and security rather

than as definitive validation evidence. Given a 5-point Likert scale, according to the literature (Boone & Boone, 2012), the reference point for success or agreement is 3.0 or higher. With all five questions assessing face validity achieving a mean score of at least 3.6, which is well above 3.0, this study therefore concludes that respondents perceived strong face alignment of the items. This outcome is further substantiated by the higher composite mean value for face validity, 4.0138, from a possible maximum mean score of 5.0 and a median value of 4.0 for all items.

Table 3. Perceived face alignment.

Statements on Face Validity	Mean	Median	Std. Deviation
Examination questions lack clarity, and is not easy to comprehend	4.2069	4	0.90156
Examination questions are often confusing and misleading	4.2759	4	0.92182
Examination questions measure accurately what they are intended to measure.	3.6552	4	1.07822
Examination questions are suitable for the intended population.	3.931	4	0.75266
Examination questions are relevant to the topics being measured.	4	4	0.70711
Grand/Composite Mean	4.0138		

In the construct validity review, we are concerned with how well a test or assessment actually measures the theoretical concept, or construct, it claims to measure. Citing the agreed threshold of 3.0 as a benchmark for a 5-point Likert scale, it is indicated that all items assessing the extent of construct validity have passed the minimum threshold as indicated in Table 4. This suggests that the University of Education, The Gambia, School of Education respondents perceived moderate construct alignment. This result is further justified by the composite mean score of 3.15172, which is also more than the acceptable minimum benchmark, and a median value of 3.0 or above.

Table 4. Perceived construct alignment.

Statements on Construct Validity	Mean	Median	Std. Deviation
Examination questions comprehensively cover all topics intended to be assessed	3.0345	3	1.0171
Examination questions focus on the subject matter and do not assess unrelated skills	3.3793	4	1.11528
The difficulty level of the examination questions is appropriate for the intended Examinees	3.2069	4	1.0481
The examination questions reflect real-world applications of the concepts being tested	3.1379	3	1.02554
The examination questions do not assess unrelated skills or knowledge areas	3	3	1.28174
Grand/Composite Mean	3.15172		

The questionnaire responses on the content validity status of items were also examined to determine the suitability of the items shown in **Table 5**. This was done to determine whether the instruments adequately and accurately cover the full scope of the subject matter or constructs they are intended to measure. The indication of content validity guarantees that the instruments are representative, relevant, and aligned with learning objectives or curriculum standards. The responses revealed that all items assessing content validity obtained a mean value higher than the threshold of 3.0, suggesting that respondents perceived strong content alignment of the instruments. This assertion is further strengthened by the higher grand mean and median values way higher than the 3.0 threshold.

Table 5. Perceived content alignment.

Statements on Content Validity	Mean	Median	Std. Deviation
The most important concepts in the course are adequately represented in the examination	3.6552	4	1.04457
The examination questions are relevant to the learning objectives of the course	4.2414	4	0.57664
Examination instructions are clear and understandable	4.3103	4	0.54139
Examination questions have a good balance of easy, moderate, and difficult questions	3.7241	4	0.88223
The time allotted for the examination questions is enough to complete all the questions, to avoid rushing	4.3103	4	0.60376
Grand Mean	4.0483		

A further review of the responses on the reliability of the assessment instruments ensued in **Table 6**. This is meant to indicate the consistency, stability, and dependability of our assessment results, knowing that a reliable assessment yields similar outcomes under consistent conditions. With the same statistical threshold as the previous, the mean values for all items testing reliability have surpassed the minimum acceptable benchmark of 3.0. This indicates positive perceptions of marking consistency. Further to that, the composite mean for the test of reliability has also surpassed the 3.0 threshold, reaffirming the perceived marking consistency of our examination items from the viewpoint of the respondents. With the median values all being way higher than 3.0, this further strengthens the respondent's assertion that the instruments are reliable. Marking guides are central to reliability. Clarity and use are good, with strong perceptions that guide, reduce subjectivity, and enable cognitive differentiation.

The conduct of assessments at the University of Education, The Gambia (School of Education) was examined to determine the level of effectiveness of the existing examination quality control mechanisms and their alignment with recognised standards. This section reviews the responses of staff members directly involved in the examination process, with particular attention to how well current

Table 6. Perceived marking consistency.

Statements on Reliability	Mean	Median	Std. Deviation
Marking guides are made available to the committee for evaluation	3.3448	4	1.34366
The criteria in the marking guide are clear	3.4483	4	1.15221
Examiners often refer to the marking guide while marking	3.5172	4	1.02193
The marking guide reduces subjective judgment in scoring	4.1034	4	0.97632
The benchmark provided in the marking guide is helpful	4.2759	4	0.5914
The marking guide allows for fair differentiation between cognitive levels	4.069	4	0.59348
Scoring of questions is consistent with similar responses	3.7241	4	0.95978
The marking guide helps assign scores consistently across different responses	4	4	0.84515
Grand Mean	3.8103		

practices meet established quality assurance benchmarks. The findings are presented in **Table 7**. Relating the aggregate responses on a 5-point Likert scale to the grand mean and the threshold of 3.0, the results indicate an overall positive perception of examination quality control across the sample. For a review of the specific parameters designed to assess examination quality control effectiveness levels, 89% (8 out of 9) agree that the required quality control measures are in place. This is highly recognizable in the following cases: The examination process was fair and consistent across all departments and programmes (Mean = 4.3103, SD = 0.4708). The grading standards were applied uniformly to all students regardless of department or programme (Mean = 4.2414, SD = 1.0231). Examination criteria were clearly communicated to all students before the examination (Mean = 4.1724, SD = 0.7106), with many other favourable results affirming high pedigree of examination quality control mechanisms. These results are further verified by the higher median values of at least 3.0 across the parameters. However, there still exists a point of disagreement on the aspect of consistency on the difficulty level of examinations across departments and programmes, with many believing that this is not sustained. The results detail that the difficulty level of the examination was consistent across different departments and Programmes (Mean = 2.8621, SD = 1.0255). With the mean value considerably falling short of the threshold of 3.0, this suggests that the respondents have disputed the existence of a consistent difficulty level of examinations across departments and programmes. It shows that the standard and difficulty level of School of Education examinations vary across departments and programmes, asserting that while some design tough exam materials, others develop much easier ones to assess learning outcomes.

Table 7. Examination quality control.

Statements on Examination Quality Control	Mean	Median	Std. Deviation
Examination criteria were clearly communicated to all students before the examination	4.1724	4	0.7106
Examination questions were relevant to the course content taught	4.0345	4	0.6805
The difficulty level of the examination was consistent across different departments and Programmes	2.8621	3	1.0255
The grading standards were applied uniformly to all students, regardless of department or programme	4.2414	5	1.0231
The feedback on examination performance was provided in a timely and transparent manner	3.069	3	1.1317
The collation of examination questions was well organized across departments and programmes	3.5172	4	1.1219
There are opportunities given to students to appeal examination results	3.1034	3	1.1755
The examiners demonstrated impartiality when grading the examinations across all departments and programmes	4.3103	4	0.6038
The examination process was fair and consistent across all departments and programmes.	4.3103	4	0.4708
Grand Mean	3.7356		

The item quality control process is concerned with ensuring that assessment instruments meet the required standards of validity, reliability, fairness, and clarity before they are administered in an assessment. To determine the effectiveness of this process, a series of statements was presented to staff responsible for examination processes, with the aim of gauging the current quality status of the instruments in use. The responses are summarised in **Table 8**. Relating to a 5-point Likert scale and a threshold of 3.0, the grand mean of 3.5733 revealed that the mechanisms for item quality control are highly effective. These findings can be reaffirmed by the results generated for specific parameters, such as; The examination questions were reviewed thoroughly to ensure clarity and accuracy (Mean = 4.3103, SD = 0.6603), The examination avoided unnecessary repetition of content across different questions (Mean = 3.9310, SD = 0.7527), The examination covered an appropriate range of topics without overemphasizing any single area (Mean = 3.7241, SD = 0.7972) and some others asserting the same position. Conversely, there is an agreement that questions were often repeated, which has the tendency of leakage, thereby undermining the level of item quality control. This perspective is asserted by the result; Questions were repeated in the examination (Mean = 2.3793, SD = 0.9029). The result highlights that questions are often repeated, which raises a concern that stringent measures need to be put in place to maintain the high credibility and standards of our assessments.

Table 8. Item quality control.

Statements on Item Quality Control	Mean	Median	Std. Deviation
The examination avoided unnecessary repetition of content across different questions	3.931	4	0.7527
Questions were repeated in the examination	2.3793	2	0.9029
The examination contained multiple questions that tested the same concept in a similar way	3.2414	4	1.0575
The intent of each examination is clear	4.1724	4	0.5391
Certain questions are open to multiple interpretations	3.5517	4	1.0885
The examination questions were reviewed thoroughly to ensure clarity and accuracy	4.3103	4	0.6603
There were no typographical or factual errors in the examination questions	3.2759	4	1.0986
The examination covered an appropriate range of topics without overemphasizing any single area	3.7241	4	0.7972
Grand Mean	3.5733		

To further strengthen quality assurance and enforce compliance with standard practices, the review was extended to the administrative quality control measures currently in place. This aspect examines the procedures and controls that govern the administration of examinations, ensuring that established protocols are consistently followed throughout the assessment process. The responses of staff involved in these procedures are summarised in **Table 9**. These measures point to the assessment processes, systems, and checks that ensure administrative work in an organization is accurate, efficient, consistent, and aligned with standards, policies, and goals. Checking the rate of compliance from our results on a 5-point Likert scale data, we rely on the minimum benchmark of 3.0 to gauge agreement on the parameters with respect to their descriptive statistics and the composite mean. Overall, the composite mean score of 3.6802, which is way above the benchmark of 3.0, suggests a strong level of agreement on the diligent implementation of administrative quality control measures. On the specific parameters put forward to assess compliance with the administrative quality control measures, the mean values are mainly above the threshold. This is substantiated by some of the following points: Examination rooms are securely locked to protect examination materials from unauthorized access (Mean = 4.1724, SD = 1.0375), Only authorized personnel have access to examination materials before, during and after examination period (Mean = 4.0690, SD = 1.2516), Examination materials are stored in secured, access-controlled locations at all times (Mean = 4.1724, SD = 0.9285), There are effective measures in place to prevent unauthorized copying or sharing of examination content (Mean = 3.8621, SD = 1.1252), which all upheld the steadfastness of the administrative quality control measures.

Table 9. Administrative quality control.

Statements on Administrative Quality Control	Mean	Median	Std. Deviation
Examination materials are stored in secure, access-controlled locations at all times	4.1724	4	0.9285
Only authorized personnel have access to examination materials before, during, and after the examination period	4.069	4	1.2516
Procedures for handling examination materials are clearly documented and consistently followed	3.7586	4	1.0575
There are effective measures in place to prevent unauthorized copying or sharing of examination content	3.8621	4	1.1252
Examination rooms are securely locked to protect examination materials from unauthorized access	4.1724	4	1.0375
Digital examination files are protected with strong passwords and encryption	3.7241	4	1.0315
There is a clear process of tracking and logging all access to the examination materials	3.2759	3	1.0656
Any breach or suspected breach of examination materials security is promptly reported and investigated	3.7586	4	0.8724
The examination and moderation committee regularly reviews and updates its security protocols for examination materials	3.5862	4	1.0183
The examination and moderation committee receives regular training on examination confidentiality and security protocols	2.3448	2	0.8567
There is designated staff responsible for overseeing examination materials security	3.7586	4	1.2721
Grand Mean	3.6802		

On the contrary, there exists a single point of agreement against the administrative quality control measures put in place, specifically on the training of the moderation committee. This is evident by the point: The examination and moderation committee receives regular training on examination confidentiality and security protocols (Mean = 2.3448, SD = 0.8567). With the mean value here being considerably lower than the threshold of 3.0, this indicates that the examination and moderation committee members have not benefited from any form of training on examination confidentiality and security protocols. This highlights a capacity gap that needs to be strengthened to enhance their effectiveness and efficiency.

Taken together, the domain-level findings suggest an examination system with relatively strong procedural foundations but uneven developmental depth. Respondents reported mostly positive perceptions of question clarity, content alignment, marking-guide use, fairness, and secure storage. At the same time, three vulnerabilities emerged repeatedly across domains: limited consistency in exami-

nation difficulty across departments and programmes, continued risk of item repetition, and insufficient routine capacity-building for staff involved in confidentiality and examination security. These patterns suggest that the institution's quality-assurance system is stronger in compliance-oriented controls than in mechanisms that promote comparability, learning, and continuous improvement.

4.3. Examination Question Submission and Moderation Feedback Practices and Preferences

Table 10 presents the findings on the academic staff's compliance with submitting their drafted examination questions for moderation. The results revealed that 98.5% of academic staff (67 out of 68 sample) report submitting their examination questions to the Examination and Moderation Committee for review, and only 1 staff member (1.5%) indicated non-submission. This shows a near-universal submission rate, reflecting a high level of adherence to internal quality assurance protocols. It is also an indication that the moderation of examination questions is an embedded and accepted practice within the School of Education.

Table 10. Submission of examination questions to the examination and moderation committee.

Response	Frequency	Percent
No	1	1.5
Yes	67	98.5
Total	68	100

Table 11 presents the findings on the reception of feedback on the moderated examination questions. The results reveal that feedback inconsistency is evident. While 23.5% of staff consistently receive feedback, a combined 53% (rarely and never) report receiving little to no feedback after moderation. The 32.4% reporting "Never" receiving feedback is particularly concerning, given that moderation is intended to enhance assessment quality and alignment. The absence or irregularity of feedback undermines the purpose of moderation and limits opportunities for professional growth and improvement of defective examination items for improved quality assessment. This irregularity in feedback provision further highlights the inactivity of some moderators.

Table 11. Receipt of feedback after moderation of submitted examination questions.

Response	Frequency	Percent
Always	16	23.5
Sometimes	16	23.5
Rarely	14	20.6
Never	22	32.4
Total	68	100

On the feedback after the moderation and the frequency of its receipt, respondents reported the kind of feedback they usually receive from the moderators, as presented in **Table 12**. The results reveal that one-on-one consultation (36.7%) is the dominant feedback method, suggesting a preference or institutional norm for personalized, informal dialogue. Despite this process's ability to foster deeper understanding, there is concern that it may lack documentation or consistency. Further to that, nearly one-third (30.9%) of staff report receiving no feedback at all, reinforcing earlier findings that moderation lacks a reliable feedback loop. The results further reveal that written feedback and email communication are severely underutilized, despite their professional status and the potential to provide clear, traceable, and actionable input. These unveil the dominance of informal feedback-sharing mechanisms and the absence of feedback from the moderators, potentially defeating the ultimate purpose of moderation.

Table 12. Types of moderation feedback on examination questions.

Responses	Frequency	Percent
No response	1	1.50%
Email communication	3	4.50%
Feedback during departmental/committee meetings	14	20.60%
No feedback provided	21	30.90%
One-on-one consultation	25	36.70%
Written comments on the exam script	4	5.90%
Total	68	100.00%

On the usefulness of the feedback received to improve the quality of examination questions, only respondents who reported receiving feedback ($n = 46$) were asked to rate its usefulness, ensuring that the estimate does not conflict with reports of no feedback received. As presented in **Table 13**, a combined 67% of these respondents (Useful + Very useful) perceived the feedback as beneficial for improving the quality of their examination questions, reflecting a generally positive orientation towards assessment refinement where feedback is provided. However, a substantial one-third (33%) rated the feedback as slightly useful or not useful (11% and 22%, respectively), indicating that the quality, relevance, or specificity of feedback remains a concern for a considerable proportion of staff. This more tempered picture, where feedback is valued by most recipients but judged inadequate by a sizeable minority, suggests that improving the consistency and substance of feedback is as important as increasing its frequency.

Table 14, shows that the staff's responses on their preferred way of receiving moderation feedback. The results reveal that email communication is the top preference (32.4%), indicating a desire for formal, asynchronous, and documented feedback. This is followed by one-on-one consultation (27.9%), reflecting the

value placed on interactive, personalized feedback, which may foster deeper understanding and professional growth. Written comments on exam papers (19.1%) are also well-regarded, likely due to their direct linkage to specific content and actionable nature. However, other feedback means, such as Group-based feedback (10.3%) and informal channels like WhatsApp (4.4%), are less favoured, possibly due to their generality or lack of documentation.

Table 13. Perceived usefulness of moderation feedback for improving examination question quality.

Responses	Frequency	Percent
Not useful	10	22%
Slightly useful	5	11%
Useful	11	24%
Very useful	20	43%
Total	46	100%

Table 14. Preferred mode of receiving moderation feedback on examination questions.

Responses	Frequency	Percent
No response	1	1.50%
Email communication	22	32.40%
One-on-one consultation	19	27.90%
Others	3	4.40%
Verbal feedback during departmental meetings	7	10.30%
WhatsApp message	3	4.40%
Written comments on the submitted exam paper	13	19.10%
Total	68	100.00%

The moderation findings also indicate that compliance and developmental value should be treated as separate dimensions of quality assurance. In this case, submission compliance is near-universal, yet the feedback loop is inconsistent and often undocumented. This distinction is likely to be relevant in other institutions where moderation exists formally but does not reliably generate actionable learning for item writers or markers. The pattern suggests the need for a moderation model that combines mandatory submission with documented, timely, and traceable feedback processes.

5. Discussion

5.1. Summary of Principal Findings

The findings portray examination quality assurance at UEG-SoE as a partially mature institutional system: many procedural controls appear to be in place, yet sev-

eral components required for comparability, traceability, and continuous improvement remain underdeveloped. Respondents reported strong perceived face and content alignment and generally positive marking-related consistency, suggesting that examinations are largely seen as relevant, clear, and supportable through marking guides. Examination quality control was also viewed favourably overall, especially in relation to fairness, communication of criteria, and grading consistency. However, difficulty comparability across departments and programmes emerged as a notable weakness, indicating that fairness in grading does not necessarily imply equivalence in assessment demand. Item quality control was generally positive, but concerns about repeated questions signal a meaningful risk to assessment integrity. Administrative security mechanisms were seen as reasonably strong in relation to storage and authorised access, yet limited training on confidentiality and security represents a significant human-capacity vulnerability. Finally, moderation appears to function strongly as a compliance mechanism but much less consistently as a developmental feedback process.

5.2. Interpretation in Light of Recent Evidence

The pattern of strong face/content validity and rubric-enabled reliability alongside middling construct validity is consistent with multi-institution evidence that assessments often under-sample complex, authentic performances even where outcomes-based language is used (Crisp, Johnson, & Constantinou, 2019; AERA, APA, & NCME, 2014; Kane, 2013). Studies in health and teacher education similarly report a dominance of lower-order cognitive demand in written examinations relative to staff intentions, with the gap attributed to design constraints, workload, and confidence in marking higher-order work (Tavakol & Dennick, 2011; Mor & Erşen, 2023). The finding of difficulty variability across departments echoes work on comparability, which argues that fairness requires not only common criteria but also shared standards enacted through calibration practices that surface tacit expectations (Sadler, 2013; University of Birmingham, & University College Birmingham (UCB), 2020). Recent synthesis indicates that moderation is most effective when it embeds calibration conversations supported by exemplars and annotated rubrics, rather than relying solely on post hoc paper checks (Prichard et al., 2025).

The observed weakness in the feedback loop aligns with sector-wide concerns that moderation is often compliance-focused, with limited actionable feedback recorded for item writers or markers (Winstone & Pitt, 2025). Contemporary feedback research emphasises feedback processes and literacies, timely, specific, and documented feedback that recipients are positioned and supported to use, over one-off comments (Carless & Boud, 2018; Winstone & Carless, 2020). Where institutions have adopted structured templates and service levels for staff-to-staff feedback in moderation, reported gains include clearer actionable points, better traceability, and improved subsequent item quality (Williams, 2024). Finally, item repetition and limited access logging are known test-security vulnerabilities. In

the absence of systematic exposure tracking and audits, item recall and circulation can compromise validity and inflate scores, particularly in tight subject communities (Haladyna & Rodriguez, 2015; Wollack & Fremer, 2013). As assessment increasingly relies on digital workflows, recent work underscores the need to couple policy with regular, documented security training and scenario-based drills (Myer & Smith, 2024; Dawson, 2020).

5.3. Transferability beyond the Case Institution

Although the findings arise from a single institution, several features of the case support analytical transferability. First, the institution operates within a broader reform context in which competency-based expectations, increased quality assurance demands, and constrained assessment infrastructure coexist, a combination common across many emerging higher education systems. Second, the tensions identified in this study are not unique to UEG-SoE: the literature similarly documents gaps between rubric availability and actual calibration, between formal moderation structures and effective feedback, and between security policy and staff training. Third, the six-domain framework used here can be adapted by other institutions as a practical audit structure for reviewing examination systems. What is most transferable is therefore not the exact numeric profile of UEG-SoE, but the pattern of strengths and vulnerabilities that the case reveals and the structured approach used to diagnose them.

The findings can be interpreted through the transferable framework shown in **Table 1**, which highlights that examination quality assurance depends not only on compliance structures but also on calibration, documentation, staff capability, and feedback loops.

5.4. A Practical Framework for Institutional Examination Quality Assurance

Based on the findings and literature, a practical institutional examination quality-assurance framework can be distilled into six linked functions: Alignment: ensure that examination tasks map explicitly to intended learning outcomes and intended cognitive demand; Consistency: support reliable scoring through compulsory marking guides, exemplars, and calibration; Comparability: review difficulty and standards across departments and programmes using blueprinting and cross-unit moderation; Item integrity: reduce ambiguity, duplication, and overexposure through structured item review and item banking; Security: protect assessment materials through secure storage, access control, logging, and regular staff training; Feedback for improvement: require documented moderation feedback, traceable revision processes, and routine review of recurring issues. This framework is intended as a practical heuristic rather than a fixed model. Its value lies in helping institutions move beyond isolated quality checks toward an integrated system that treats examination quality assurance as both an accountability process and a continuous-improvement cycle.

5.5. Implications for Policy and Practice

The implications of the present findings extend beyond the focal institution because they reflect recurring implementation challenges in higher education assessment systems that seek to balance fairness, validity, security, and feasibility. Strengthening construct validity requires intentional design choices. Mandating cognitive blueprints that map items to intended learning outcomes (ILOs) and Bloom levels, with explicit targets for Analyse/Evaluate/Create appropriate to year/level, is a feasible first step (Anderson & Krathwohl, 2001; Biggs & Tang, 2011). Stimulus-rich selected-response items (e.g., data displays, case vignettes, extended-matching formats) can elicit multi-step reasoning at scale when paired with rigorous item-writing guidelines and peer review; constructed-response tasks should be supported with criterion-referenced rubrics and exemplars to stabilise marking while preserving higher-order demand (Haladyna, Downing, & Rodriguez, 2002; Downing, 2005; Brookhart, 2018). Difficulty consistency is best addressed through cross-department calibration panels that agree on anchor items, marking expectations, and target facility distributions by course level; post-examination dashboards can flag outliers for targeted support and re-standardisation (Sadler, 2013; University of Edinburgh, 2024). To mitigate item repetition and leakage risks, a secure item bank with metadata (topic, Bloom level, difficulty), version control, and exposure tracking should be implemented, alongside periodic audits and duplication scans. Where resources allow, automatic item generation (AIG) and templating can diversify item pools while maintaining cognitive alignment (Gierl, Lai, & Turner, 2012; Gierl, Bulut, Guo, & Zhang, 2017). Administrative security should be reinforced through annual training for all personnel handling examination materials, role-based access logging with routine reviews, and breach-response drills (Myer & Smith, 2024; Dawson, 2020). Moderation should be reoriented from compliance to learning by instituting feedback service level agreements (SLAs), for example, documented written/email feedback within 10 working days using a standard template, paired with scheduled calibration sessions and an archive of feedback to support organisational learning (Prichard et al., 2025; Winstone & Carless, 2020).

5.6. Limitations

Several limitations should be considered when interpreting the findings. First, the study is based on a single institutional case, which limits statistical generalisation even though the case may still support analytical transferability. Second, the evidence relies mainly on self-reported perceptions from staff and committee members rather than direct analysis of examination papers, moderation records, scoring data, or security logs. Third, although the instruments showed acceptable internal consistency, further work is needed to establish their factor structure and broader validity for use across other institutions. Fourth, some domains, particularly construct validity and reliability, were assessed indirectly through perceptions and operational indicators rather than through direct psychometric evidence. Finally, the descriptive design does not allow causal inference about which

quality-assurance interventions would produce the strongest improvements.

6. Conclusion

This study used a single teacher education institution in The Gambia as a case through which to examine examination quality assurance as a multi-domain institutional system. The findings suggest that while important procedural foundations are in place, including marking-guide use, secure storage, and widespread moderation submission, stronger mechanisms are still needed to support comparability across departments, reduce item repetition, strengthen security-related staff capacity, and make moderation feedback more systematic and developmental. Viewed more broadly, the study illustrates how institutions may appear procedurally compliant while still lacking the feedback, calibration, and learning mechanisms necessary for a fully mature assessment quality-assurance system. Addressing these gaps is central to fairness, integrity, and the constructive alignment of intended learning outcomes, teaching/learning activities, and assessment tasks. We recommend five mutually reinforcing priorities: mandate cognitive blueprints with explicit targets for higher-order thinking; institute cross-department calibration to stabilise difficulty and standards; deploy a secure, metadata-rich item bank with exposure tracking and duplication safeguards; formalise moderation as a feedback-rich process with service level agreements, templates, and archived records; and embed routine post-examination analytics and annual security training with access-log audits. These measures align with international standards for validity and reliability, are feasible within existing governance structures, and directly support The Gambia's competency-based policy by modelling robust assessment practice for future teachers.

Implementation should be staged and data-informed. A pragmatic 12-month roadmap, encompassing policy updates, staff development, pilot blueprinting and calibration, item-bank rollout, and dashboard monitoring, can build momentum while enabling iterative refinement through plan-do-study-act cycles. Clear ownership, resourcing, and routine reporting to programme boards will be essential to sustain gains and ensure accountability. As processes mature, incorporating elements of programmatic assessment, staged authentic tasks with iterative feedback, and structured aggregation of evidence can further enhance validity and learning value.

This evaluation is based on perceptions and composite indicators; future work should triangulate with independent audits of exam artefacts, item statistics, moderation records, and security logs, report inter-rater reliability for any cognitive classifications, and disaggregate findings by discipline and level. Quasi-experimental evaluations of blueprinting, calibration, and feedback interventions, as well as cautious pilots of decision-support tools (e.g., NLP/ML aids for item review), will help establish causal impacts on item quality, reliability, and student outcomes.

As shown in **Table 1**, institutions may strengthen examination quality assur-

ance most effectively when they treat alignment, consistency, comparability, item integrity, security, and feedback as linked components of one system.

The broader contribution of the study lies in demonstrating a practical framework through which higher education institutions, especially those operating in teacher education and resource-constrained contexts, can review examination quality assurance across linked domains rather than in isolation. Future research should test this framework across multiple institutions, incorporate audit evidence from examination artefacts and moderation records, and examine whether interventions such as blueprinting, calibration panels, item banking, and feedback service-level agreements produce measurable gains in item quality, comparability, and assessment integrity.

7. Recommendations

The recommendations below are framed not only as institution-specific actions for UEG-SoE but also as potentially adaptable implementation strategies for comparable higher education settings seeking to strengthen examination quality assurance.

1. Adopt a School-wide assessment quality assurance policy, mandate cognitive blueprints for all exams, set clear roles/timelines, and appoint leads to oversee implementation.
2. Run regular workshops on item writing for higher-order thinking, rubric design and calibration, moderation facilitation, post-exam analytics, and assessment security.
3. Provide structured induction for new staff on SoE assessment policies, blueprinting, moderation, and security; pair novices with experienced setters/moderators.
4. Deploy a secure, metadata-rich item bank with version control and exposure tracking; mandate duplication scans before approval; set rotation/retirement rules for reused items.
5. Strengthen pre-moderation checklist with explicit clarity/ambiguity and error checks; require an “item intent” note for each question and a second reviewer sign-off; use small-scale cognitive walk-throughs for complex items.
6. Require cognitive blueprints for every paper, mapping items to intended learning outcomes and Bloom levels; set minimum targets for application/analysis/evaluation appropriate to year/level; include authentic, stimulus-rich tasks; align rubrics to intended cognition.
7. Make the marking guide a compulsory component of every submission pack; adopt standard rubric templates with annotated exemplars; double-mark a sample per course and discuss discrepancies.
8. Introduce a moderation feedback service level agreement requiring written/email feedback via a standard template within 10 working days; archive all feedback; schedule brief follow-ups only where needed.
9. Use a standard template that addresses clarity, content/construct alignment,

difficulty, cognitive balance, and security notes; archive all feedback for audit and professional learning.

10. Implement mandatory, annual security/confidentiality training (recorded) for all authorised staff; enforce role-based access with comprehensive logging and quarterly audits; run an annual breach-response drill.

11. Provide cohort-level post-exam feedback briefs (common strengths, misconceptions, next steps) within 10 working days without disclosing secure content.

12. Pilot natural language processing and machine learning (NLP/ML) tools to flag potential ambiguity, duplication, cognitive imbalance, and reading-load issues in draft items, complementing—not replacing—expert review.

Funding

The authors declare that no financial support was received for the research and/or publication of this article.

Ethic

Ethical approval was not required for this study; however, informed consent and institutional approval were obtained in line with national regulations and institutional policies.

Consent to Participate

All the research participants gave voluntary informed consent to participate in the study.

Consent to Publish

All the research participants provided informed consent for participation and publication of the research findings.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing*.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson.
- Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University* (4th ed.). Open University Press.
- Boone, H. N., & Boone, D. A. (2012). Analyzing Likert Data. *Journal of Extension*, 50, Ar-

- title 48. <https://doi.org/10.34068/joe.50.02.48>
- Brookhart, S. M. (2018). *How to Create and Use Rubrics for Formative Assessment and Grading* (2nd ed.). ASCD.
- Carless, D., & Boud, D. (2018). The Development of Student Feedback Literacy: Enabling Uptake of Feedback. *Assessment & Evaluation in Higher Education*, 43, 1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Crisp, V., Johnson, M., & Constantinou, F. (2019). A Question of Quality: Conceptualisations of Quality in the Context of Educational Test Questions. *Research in Education*, 105, 18-41. <https://doi.org/10.1177/0034523717752203>
- Dawson, P. (2020). *Defending Assessment Security in a Digital World*. Routledge.
- Downing, S. M. (2005). The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. *Advances in Health Sciences Education*, 10, 133-143. <https://doi.org/10.1007/s10459-004-4019-5>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87, 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using Automatic Item Generation to Create Multiple-Choice Test Items. *Medical Education*, 46, 757-765. <https://doi.org/10.1111/j.1365-2923.2012.04289.x>
- Haladyna, T. M., & Rodriguez, M. C. (2015). *Developing and Validating Test Items*. Routledge.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15, 309-333. https://doi.org/10.1207/s15324818ame1503_5
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77, 81-112. <https://doi.org/10.3102/003465430298487>
- IIEP-UNESCO (2021). *IIEP in Action 2020-2021*. <https://www.iiep.unesco.org/en/iiep-action>
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50, 1-73. <https://doi.org/10.1111/jedm.12000>
- Ministry of Basic and Secondary Education (MoBSE) (2015). *National Policy Documents on Competency-Based Curricula and Assessment*. Government of The Gambia. <http://www.edugambia.gm/data-area/publications/year-book-2016/251-year-book2016/download.html>
- Mor, E., & Karatoprak Erşen, R. (2023). Implications of Current Validity Frameworks for Classroom Assessment. *International Journal of Assessment Tools in Education*, 10, 164-173. <https://doi.org/10.21449/ijate.1368458>
- Myer, L., & Smith, A. W. (2024). Academic Integrity in the United Kingdom: The Quality Assurance Agency's National Approach. In S. E. Eaton (Ed.), *Springer International Handbooks of Education* (pp. 841-858). Springer. https://doi.org/10.1007/978-3-031-54144-5_173
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press.
- Prichard, R., Peet, J., El Haddad, M., Chen, Y., & Lin, F. (2025). Assessment Moderation in Higher Education: Guiding Practice with Evidence—An Integrative Review. *Nurse Education Today*, 146, Article 106512.

- Quality Assurance Agency for Higher Education (2024). *The UK Quality Code for Higher Education*. <https://www.qaa.ac.uk/the-quality-code>
- Sadler, D. R. (2013). Assuring Academic Achievement Standards: From Moderation to Calibration. *Assessment in Education: Principles, Policy & Practice*, 20, 5-19. <https://doi.org/10.1080/0969594x.2012.714742>
- Shaw, S., & Crisp, V. (2011). Tracing the Evolution of Validity in Educational Measurement: Past Issues and Contemporary Challenges. *Research Matters*, No. 11, 14-17.
- Tavakol, M., & Dennick, R. (2011). Post-Examination Analysis of Objective Tests. *Medical Teacher*, 33, 447-458. <https://doi.org/10.3109/0142159x.2011.564682>
- UNESCO (2018). *Global Education Monitoring Report 2017/18: Accountability in Education: Meeting our Commitments*. United Nations iLibrary.
- University of Birmingham, & University College Birmingham (2020). *Moderation Code of Practice*. <https://www.ucb.ac.uk/about-us/external-examiners/>
- University of Edinburgh (2024). *Internal Moderation Guidance*. <https://registryservices.ed.ac.uk/academic-services/policies-regulations/new-policies/assessment-feedback>
- Williams, A. (2024). Delivering Effective Student Feedback in Higher Education: An Evaluation of the Challenges and Best Practice. *International Journal of Research in Education and Science*, 10, 473-501. <https://doi.org/10.46328/ijres.3404>
- Winstone, N. E., & Carless, D. (2020). *Designing Effective Feedback Processes in Higher Education*. Routledge.
- Winstone, N. E., & Pitt, E. (2025). Approaches to Feedback on Examination Performance: Research, Policy, and Practice. *Assessment & Evaluation in Higher Education*, 50, 876-896. <https://doi.org/10.1080/02602938.2025.2476622>
- Wollack, J. A., & Fremer, J. J. (2013). *Handbook of Test Security*. Routledge.