Assessing History Assessment: Teacher and Student Perspectives on Standardized Testing and Instructional Reform in Bhutan

Thukten Tenzin[1] and Ugyen Tenzin[2]

[1]EAO, BCSEA, MoESD, Bhutan

[2]Teacher, Chapcha Middle Secondary School, MoESD, Bhutan

DOI: https://doi.org/10.68050/JAMS.2026.692

Received: 10 September 2026; Accepted: 22 September 2026; Published: 05 October 2026

ABSTRACT

Standardized assessment remains a central mechanism for educational accountability, yet its capacity to represent authentic historical understanding continues to be debated. This study employed a convergent mixed-methods approach to examine teachers' and students' perceptions of standardized History assessment in Bhutan across five domains: fairness and curriculum alignment; motivation and historical understanding; use of assessment data for instructional improvement; challenges and reform needs; and teacher assessment literacy. Quantitative data were obtained from 677 Grade X students and 59 History teachers using structured questionnaires, while qualitative evidence was gathered from twelve students and six teachers through a focus-group discussion and semi-structured interviews. The student questionnaire contained 29 closed-ended items and the teacher questionnaire 20 closed-ended items, rated on a five-point Likert scale. The instruments were reviewed by three History experts, piloted with 30 Grade X students and 29 teachers, and demonstrated an overall Cronbach's alpha of .80. Quantitative data were summarized using item-level means and standard deviations, and qualitative data were analysed thematically. Integration occurred during interpretation by comparing convergence, complementarity, and divergence between the two strands within the same five domains. Overall, respondents reported favourable perceptions of fairness, curriculum alignment, academic focus, feedback, and instructional use of assessment data. At the same time, participants perceived continuing concerns related to factual recall, limited inquiry opportunities, contextual inequities, examination pressure, and insufficient formal preparation in assessment literacy. These findings should be interpreted as stakeholder perceptions rather than direct evidence of psychometric validity or causal effects on learning. The study recommends a more balanced assessment system that retains common standards while expanding authentic historical inquiry, strengthening formative use of assessment evidence, and providing sustained professional learning in assessment literacy.

Keywords: Standardized assessment; History education; mixed methods; assessment literacy; student motivation; curriculum alignment; educational reform.

INTRODUCTION

Standardized assessment is an established feature of contemporary education systems and is commonly used to support comparability, accountability, certification, and instructional decision-making. In History education, however, the assessment challenge is particularly complex because disciplinary learning extends beyond the recall of factual information. Meaningful historical understanding also involves sourcing, interpretation of evidence, contextualization, corroboration, argumentation, and reasoned judgment. Consequently, an assessment may be procedurally standardized while still raising questions about the extent to which it represents the broader construct of historical thinking (Wineburg, 2001; Seixas & Morton, 2022).

In Bhutan, standardized History assessments play an important role in evaluating student achievement and maintaining common academic expectations. Stakeholders may nevertheless experience tensions between the accountability functions of standardized examinations and the pedagogical aims of History education. On the one hand, standardized assessments may be perceived as fair, objective, consistent, and aligned with

curriculum expectations. On the other hand, concerns may arise when test preparation privileges factual recall, when language or resource differences shape students' opportunities to demonstrate learning, or when feedback is not available quickly enough to inform subsequent instruction. Teacher capacity to interpret assessment evidence and to translate results into responsive teaching is equally important.

The present study therefore examines how teachers and Grade X students perceive standardized History assessment in Bhutan. Importantly, the study does not attempt to establish the actual psychometric validity of national assessments or to make causal claims about their effects on motivation or historical understanding. Rather, it investigates stakeholder perceptions of assessment quality, instructional usefulness, learning consequences, and reform priorities. This distinction is necessary because self-reported perceptions provide important evidence about experience and practice, but they do not by themselves demonstrate causal impact or criterion validity.

Problem Statement

Standardized assessment is a major mechanism for educational accountability and quality assurance in Bhutan, but there is limited empirical evidence concerning how History teachers and students experience its fairness, curriculum alignment, instructional usefulness, and capacity to support disciplinary learning. Existing concerns include possible overemphasis on factual recall, restricted opportunities for inquiry, contextual differences in language and access to learning resources, constraints in the timeliness and formative use of feedback, and uneven professional preparation in assessment literacy. These concerns point to a potential gap between the accountability purposes of standardized assessment and the pedagogical goal of developing higher-order historical reasoning. A systematic examination of stakeholder perceptions is therefore warranted to identify areas of convergence, tension, and possible improvement.

Research Aim

The study aimed to examine teachers' and students' perceptions of standardized History assessment in Bhutan, with particular attention to fairness and curriculum alignment, perceived effects on motivation and historical understanding, the instructional use of assessment data, challenges and reform needs, and teacher assessment literacy.

Research Objectives

i. To examine teachers' and students' perceptions of the fairness, objectivity, consistency, and curriculum alignment of standardized History assessments.

ii. To examine teachers' and students' perceptions of how standardized assessment relates to motivation, study behaviour, and depth of historical understanding.

iii. To investigate how standardized assessment results are perceived to be used for feedback, instructional planning, and targeted support.

iv. To identify perceived challenges in current standardized testing practices and stakeholder priorities for assessment reform.

v. To examine teachers' self-reported confidence, professional practices, and preparation in assessment literacy and the use of standardized assessment evidence.

Research Questions

i. How do teachers and students perceive the fairness and curriculum alignment of standardized History assessments in Bhutan?

ii. How do teachers and students perceive the relationship between standardized assessment, motivation, study behaviour, and historical understanding?

iii. How are standardized assessment results perceived to be used for feedback and instructional improvement?

iv. What challenges do teachers and students identify in the current standardized assessment system, and what reforms do they consider necessary?

v. What levels of confidence, professional practice, and formal preparation do History teachers report in relation to assessment literacy?

Significance of the Study

This study has conceptual, policy, and classroom significance. Conceptually, it contributes context-specific evidence to discussions of assessment fairness, curriculum alignment, disciplinary validity, and the relationship between accountability-oriented testing and meaningful historical learning. By explicitly distinguishing perceived assessment quality from direct evidence of psychometric validity, the study provides a more appropriately bounded interpretation of stakeholder evidence.

From a policy perspective, the findings can inform review of standardized History assessment design, reporting practices, equity considerations, and professional support for teachers. At the classroom level, the study provides insight into how teachers and students experience assessment expectations, feedback, examination preparation, and the balance between content mastery and higher-order historical reasoning. These insights may support the development of more balanced assessment practices that combine common standards with formative, authentic, and inquiry-oriented tasks.

LITERATURE REVIEW

The literature relevant to the study is organized around five interrelated domains: fairness and curriculum alignment; motivation and historical understanding; use of assessment data for feedback and instructional improvement; systemic challenges and reform; and teacher assessment literacy. Together, these domains provide the conceptual framework for interpreting stakeholder perceptions of standardized History assessment.

Fairness and Alignment of Standardized History Assessments

Fairness and curriculum alignment are central attributes of high-quality assessment. Standardized procedures can promote consistency by reducing variation in administration and scoring, which is particularly relevant in subjects such as History where some responses require interpretation (Popham, 2017). Biggs and Tang's (2011) constructive alignment framework similarly emphasizes coherence among intended learning outcomes, teaching activities, and assessment tasks. Evidence from different educational settings suggests that stronger alignment can enhance the perceived credibility of assessment and teachers' confidence in the resulting information (Hsieh & Tsai, 2023; Muller & Colwell, 2022).

Nevertheless, procedural consistency should not be equated automatically with substantive fairness. Linguistic diversity, socioeconomic differences, school context, and unequal access to learning resources may influence students' opportunities to demonstrate what they know (Brookhart, 2018; Hopfenbeck & Throndsen, 2022). In History education, validity is also threatened when assessment disproportionately emphasizes factual reproduction rather than evidence evaluation, interpretation, and reasoning. Thus, stakeholder perceptions of fairness are educationally important, but they remain distinct from a direct technical evaluation of test validity.

Motivation and Historical Understanding

The relationship between standardized assessment and motivation is multidimensional. Self-Determination Theory suggests that external assessment can support engagement when expectations are perceived as meaningful, fair, and attainable (Deci & Ryan, 2000). Clear performance expectations can also strengthen study routines and academic focus (Harlen, 2014; Lee & Chong, 2023). Within History education,

standardized assessment may therefore provide structure by signalling priority content and expected levels of performance (Kazemi & Ashrafi, 2022).

At the same time, strong dependence on summative testing may encourage surface approaches to learning. Shepard (2000) cautions that assessment systems dominated by summative purposes can privilege short-term performance strategies. Wineburg (2001) and Seixas and Morton (2022) emphasize that historical thinking requires sourcing, contextualization, corroboration, interpretation, and reasoned judgment. These competencies may be only partially represented in conventional examination formats. Accordingly, the present study treats participants' reports about motivation and understanding as perceptions of learning experience rather than evidence that standardized assessment directly causes particular learning outcomes.

Use of Assessment Data for Feedback and Instructional Improvement

Assessment information has greatest instructional value when it is timely, interpretable, and actionable. Feedback can help students identify misconceptions and decide on appropriate next steps, while teachers can use assessment evidence to adjust instruction, differentiate support, and plan targeted interventions (Hattie & Clarke, 2019; Schildkamp, 2020). In History education, feedback may also support interpretive skill and historical literacy when it addresses the quality of reasoning rather than merely the correctness of recalled information (Wylie & Lyon, 2023). Delayed reporting, however, can reduce the formative usefulness of assessment results for immediate instructional decision-making (Dorji & Gyeltshen, 2023; Nusrat et al., 2022).

Challenges and the Need for Assessment Reform

High-stakes testing has frequently been associated with examination pressure, teaching to the test, and narrowing of the enacted curriculum. The OECD (2023) notes that strong accountability pressures may reshape instructional priorities, while Harlen (2018) argues that excessive reliance on summative assessment can weaken formative practice and reduce opportunities for creativity and inquiry. In response, contemporary assessment literature increasingly advocates balanced systems that combine summative, formative, performance-based, and authentic approaches (Fastre & Veenman, 2022; Wiliam, 2023). In History, such balance may include source analysis, document-based tasks, argumentation, narrative reasoning, and inquiry-based work (Seixas & Morton, 2022).

Teacher Competency, Assessment Literacy, and Professional Practice

Teacher assessment literacy is essential to the meaningful use of standardized assessment. Assessment literacy includes technical, conceptual, and pedagogical competencies such as aligning tasks with learning outcomes, constructing defensible assessment items, interpreting evidence appropriately, communicating results, and using feedback to support learning (Brown & Harris, 2023). International research indicates substantial variation in teachers' preparation for these roles and highlights the importance of sustained professional learning (DeLuca et al., 2022; Xu & Brown, 2023; Heritage, 2022). In History, professional learning must additionally address the discipline-specific assessment of evidence use, interpretation, contextualization, and historical reasoning (Levesque & Croteau, 2023).

METHODOLOGY

Research Design and Mixed-Methods Rationale

The study adopted a convergent mixed-methods design in which quantitative and qualitative evidence (see Figure 3.1) addressed the same substantive domains and were brought together during interpretation. The quantitative strand was used to describe broad patterns in teachers' and students' perceptions, whereas the qualitative strand was used to explore the meanings, experiences, and contextual explanations underlying those patterns. The rationale for combining the two forms of evidence was complementarity: survey data offered breadth across a comparatively large sample, while interviews and focus-group discussion provided depth and contextual nuance that could not be captured through closed-ended items alone.

Integration occurred at the interpretation stage. Findings from both strands were organized under the same five domains, and qualitative evidence was used to clarify, elaborate, or qualify quantitative patterns. Convergence was identified when the strands pointed in a similar direction; complementarity was identified when qualitative accounts added explanation or contextual detail to a survey pattern; and divergence was acknowledged when the strands did not fully correspond. This integrative procedure is important because the study's conclusions are based on the combined meaning of both strands rather than on two independent sets of results.

Study Context, Participants, and Sampling

The quantitative sample comprised 677 Grade X students and 59 History teachers from Middle Secondary Schools across Bhutan (see Figure 3.1). The survey was distributed through the national MSS Principals' group to reach schools from all 20 Dzongkhags and included both urban and rural contexts. Responses were received from 50 schools out of 57 schools contacted. Eligible participants were Grade X students and teachers currently teaching Grade X. As participation was voluntary rather than probability-based, the sample provides broad national coverage but should not be regarded as statistically representative of the entire population.

The qualitative component included twelve students and six teachers selected purposively (see Figure 3.1). Three students were drawn from each of the western, eastern, southern, and northern regions to ensure geographical variation, while the teachers were selected based on their direct involvement in marking and evaluation. The qualitative sample was intended to generate diverse, experience-based insights into assessment practices rather than statistical generalization.

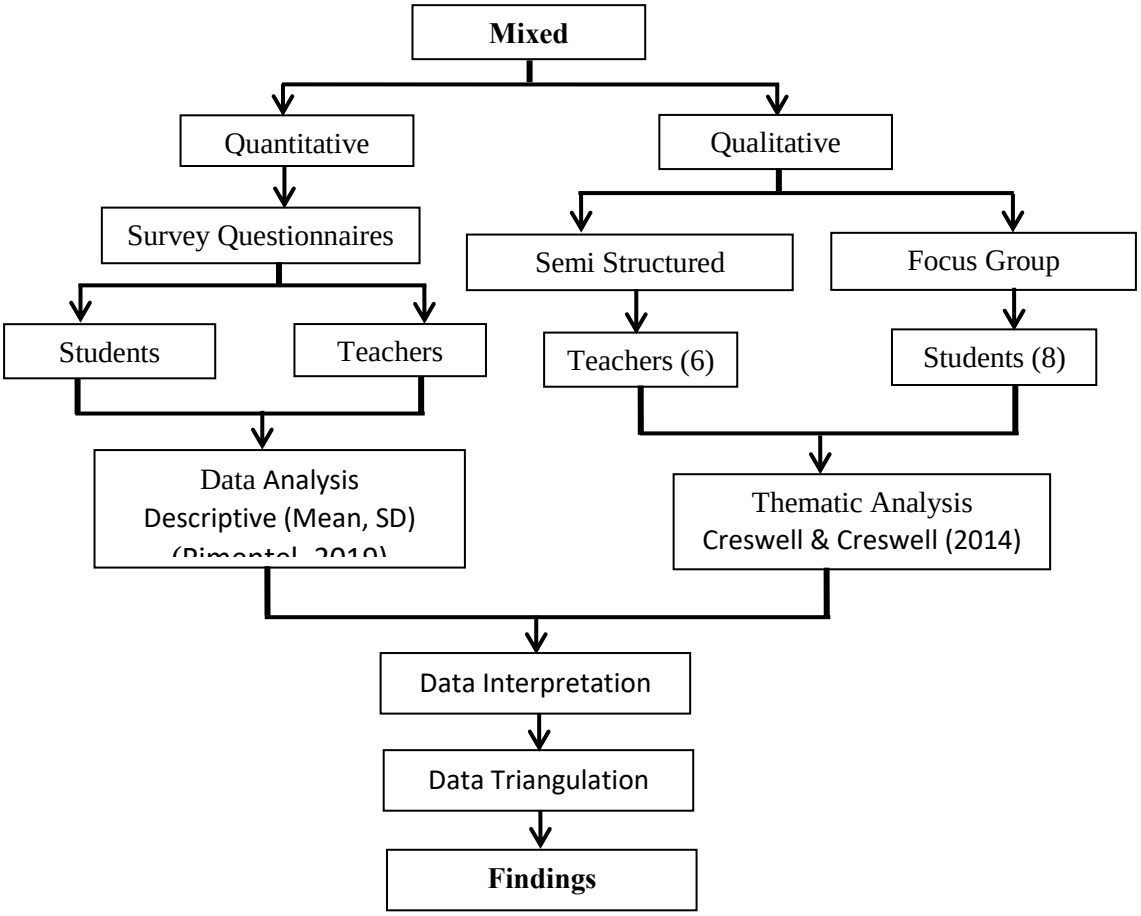


Figure 3.1 ***Sampling plan***

Quantitative Instrument

Two researcher-developed structured questionnaires were used to collect quantitative data: 20 item student questionnaire and 25 item teacher questionnaire. The instruments were developed in alignment with the Grade X History curriculum and the study objectives and covered five domains: fairness and curriculum alignment, motivation and historical understanding, use of assessment data for feedback and instructional improvement,

assessment challenges and reform needs, and teacher competency and assessment literacy. The teacher questionnaire addressed all five domains, whereas the student questionnaire included only those domains relevant to students' direct assessment experiences.

All closed-ended items were rated on a five-point Likert scale (see Table 3.1) ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), following the response format described by Pimental (2019). Numerical scores were assigned to each response category for quantitative analysis, and no reverse-scoring was applied because the instrument did not include negatively worded items requiring score reversal. The analysis focused on item-level means and standard deviations rather than treating the thematic domains as independently validated latent constructs; therefore, the findings are interpreted at the item and thematic levels without making claims regarding the construct validity of the domains.

Table 3.1 *Criteria for Interpreting the Means of the Perception Level*

Likert-Scale	Likert Scale interval	Level of Interpretation	Level of Agreement
1	1.00 – 1.79	Very low	Strongly Disagree
2	1.80 – 2.59	Low	Disagree
3	2.60 – 3.39	Moderate	Neutral
4	3.40 – 4.19	High	Agree
5	4.20- 5.00	Very High	Strongly Agree

Note. Adapted from Pimental (2019, p.188).

Instrument Review, Pilot Testing, Validity, and Reliability

Prior to the main study, the student and teacher questionnaires were reviewed by three History experts for language appropriateness, clarity, relevance, and coverage of the intended thematic domains. The instruments were then piloted with 30 Grade X students and 29 teachers from Chapcha Middle Secondary School. This process provided evidence of content relevance and contextual comprehensibility within a Bhutanese school setting.

Internal consistency was examined in Jamovi using Cronbach's alpha, and an overall coefficient of $\alpha = .80$ was reported. This value indicates acceptable internal consistency at the overall instrument level. However, because the available study records report only a single alpha coefficient, reliability should not be generalized automatically to each individual thematic domain. Likewise, expert review and pilot testing provide evidence of content appropriateness, but they do not constitute a full psychometric validation of the constructs.

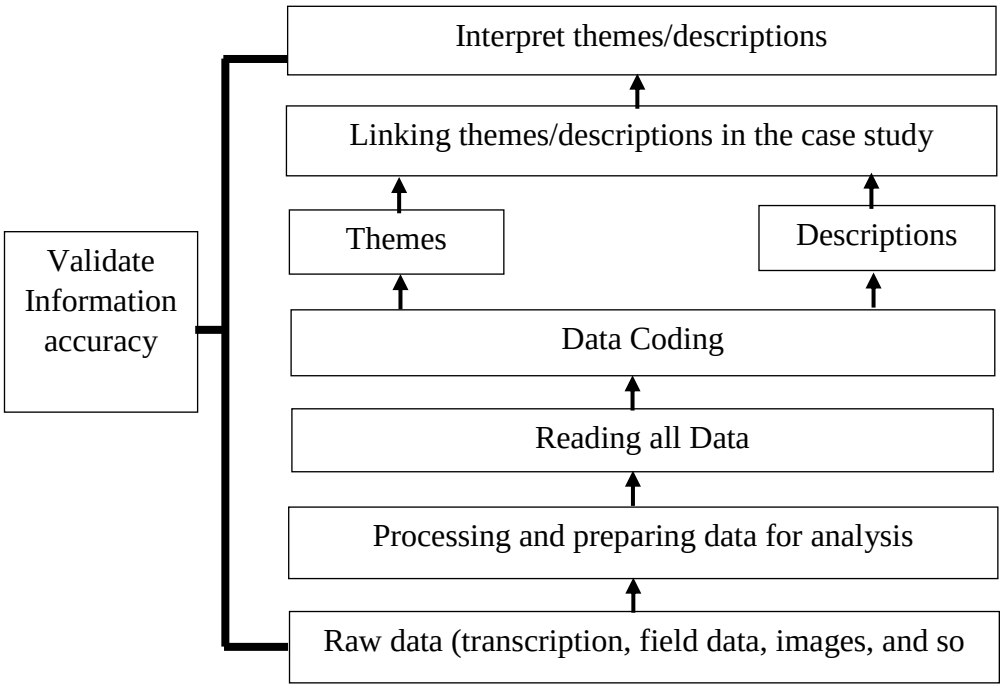
Qualitative Data Collection

Qualitative data were collected through four student focus-group discussions and six semi-structured teacher interviews. The discussion and interview guides were aligned with the five questionnaire domains, enabling participants to elaborate on survey-related issues and provide contextual insights beyond the closed-ended items. Each student focus group comprised one high, middle, and low-performing student to capture variation in academic achievement. All sessions were conducted in English in the school conference hall, lasted approximately 30 minutes, and were facilitated by the researcher, who also maintained field notes. With participants' consent, the discussions and interviews were audio-recorded and subsequently transcribed systematically following the qualitative data preparation procedures outlined by Creswell and Creswell (2014). Confidentiality was maintained by removing identifying information from transcripts and reports, and focus-group participants were requested to respect the confidentiality of views shared during the discussions.

Data Analysis

Quantitative data were analysed in Jamovi using descriptive statistics, principally means and standard deviations. Mean scores were interpreted using the five intervals (see Table 3.1) adapted from Pimental (2019): In this, the score range was divided into 5 categories: Very Low (1-1.79), Low (1.80 - 2.59), Moderate (2.60 -3.39), High (3.40 4.19), and Very High (4.20 -5.00). The lowest possible mean score was 1 and the highest score was 5.

For qualitative data analysis, the participants' responses to semi-structured interview were obtained using audio recording. By listening to the audio based data repeatedly, the data were transcribed for analysis. Further, the transcription was read line by line and coded the relative words for categorization as the main aim of the interview question was to obtain in-depth information. The categorization of qualitative data was analyzed thematically following the sequential steps suggested by Creswell and Creswell (2014) (see figure 3.2)



Note. Adopted from the book by Creswell (2014, p. 247).

Figure 3.2 *Thematic analysis steps*

Ethical Considerations

Ethical safeguards were applied throughout the study. Written permission was obtained from the relevant school administration. Participants were informed of their right to decline participation or withdraw, and anonymity was protected through the use of pseudonyms. Written informed consent was obtained from student participants before data collection. Confidentiality was maintained in the handling of participant information, and the study reported no foreseeable risks associated with participation.

RESULTS

The results are organized according to the five study domains. Quantitative findings are reported descriptively and are integrated with qualitative evidence in each subsection. Because the quantitative data are perception ratings, the results should be read as evidence of stakeholder views rather than as direct proof of assessment validity, instructional effectiveness, or causal impact.

4.1 Fairness and Curriculum Alignment

No.	Survey item	Teachers M (SD)	Interpretation	Students M (SD)	Interpretation
1	Standardized History tests provide a fair and objective measure of students' performance.	4.46 (.63)	Very High	3.93 (1.04)	High
2	The format and content of standardized tests align with curriculum learning objectives.	4.36 (.64)	Very High	4.12 (.99)	High
3	Standardized tests ensure consistency in assessing History across schools.	4.34 (.66)	Very High	3.95 (1.00)	High
4	The current assessment system maintains academic standards in History education.	4.02 (.84)	High	3.88 (.99)	High
5	The design of standardized tests allows students to demonstrate higher-order thinking skills.	4.37 (.67)	Very High	4.21 (.97)	Very High
6	Standardized tests make me take History more seriously.	—	—	3.88 (.99)	High
7	The test papers allow me to show what I really know and understand.	—	—	4.21 (.97)	Very High

Teachers reported very high agreement that standardized History assessments are fair and objective (M = 4.46, SD = .63), aligned with curriculum objectives (M = 4.36, SD = .64), and implemented consistently across schools (M = 4.34, SD = .66). Students similarly expressed favourable perceptions of assessment fairness (M = 3.93, SD = 1.04) and curriculum alignment (M = 4.12, SD = .99). These quantitative findings indicate generally strong stakeholder confidence in the procedural fairness and curricular relevance of standardized History assessment. This perception was reinforced by a teacher who explained, "The current standardized History tests are generally fair and objective in structure, as they are based on national curricula and use uniform marking schemes that ensure consistency across schools" (T_2). Similarly, one student stated, "I think standardized History tests are fair because everyone gets the same questions and is graded in the same way" (S_1).

Nevertheless, the qualitative findings also revealed important contextual qualifications. One student observed, "Some schools have better resources. Some students struggle with the test language. Not everyone learns the same way" (S_1), highlighting how differences in resources, language proficiency, and learning contexts may affect equitable opportunities to demonstrate achievement. Concerns regarding the nature of assessment tasks were also evident in the teacher data, with one participant noting that standardized tests "often limit higher order thinking by emphasizing recall over analysis, focusing on dates instead of evaluating historical significance" (T_6). Taken together, the findings suggest that participants generally perceived standardized History assessment as fair and curriculum aligned, while simultaneously identifying contextual inequities and an emphasis on factual recall that may constrain the assessment of higher order historical thinking. These findings represent stakeholder perceptions and should not be interpreted as direct evidence of psychometric fairness or full construct validity.

4.2 Motivation, Study Behaviour, and Historical Understanding

No.	Survey item	Teachers M (SD)	Interpretation	Students M (SD)	Interpretation
1	Standardized tests encourage students to focus on key historical concepts and skills.	4.17 (.67)	High	4.03 (.91)	High
2	Students are motivated to improve their	3.88	High	3.56	High

	performance after receiving their test results.	(.67)		(.95)	
3	The emphasis on standardized testing limits creativity and inquiry-based learning.	3.44 (1.19)	High	3.42 (1.02)	High
4	Standardized test outcomes accurately reflect students' historical understanding.	3.85 (.78)	High	4.03 (.95)	High
5	Students develop better exam strategies but not necessarily deeper comprehension.	3.81 (.94)	High	3.92 (.97)	High
6	Standardized tests help me focus on important topics in History.	—	—	4.03 (.91)	High
7	I study harder for standardized tests than for other assessments.	—	—	3.56 (.95)	High
8	These tests make me memorize facts rather than understand History deeply.	—	—	3.42 (1.02)	High
9	My results in standardized tests show how much I have really learned.	—	—	4.03 (.95)	High
10	I feel more confident in History after preparing for these tests.	—	—	3.92 (.97)	High

The survey results indicate that standardized assessment is perceived as providing academic focus and encouraging preparation. Teachers agreed that standardized tests encourage attention to key historical concepts and skills (M = 4.17, SD = .67), while students reported that such tests help them focus on important topics (M = 4.03, SD = .91). Students also reported relatively high confidence after preparing for standardized tests (M = 3.92, SD = .97). This positive function was reflected in the qualitative data, with one student explaining, "These tests motivate me to study more seriously and review what I've learned" (S_8). Similarly, a teacher noted that standardized tests "encourage students to study seriously and focus on key historical facts, dates, and themes" (T_1).

At the same time, the findings reveal a tension between examination preparation and deeper disciplinary learning. Teachers agreed that standardized testing can limit creativity and inquiry based learning (M = 3.44, SD = 1.19) and can promote examination strategies without necessarily deepening comprehension (M = 3.81, SD = .94). Students likewise reported a tendency toward memorization rather than deep understanding (M = 3.42, SD = 1.02). This concern was clearly expressed by one student who stated, "I mostly memorize because the test questions often need exact facts" (S_6), while a teacher observed that standardized assessments "often prioritize memorization over critical analysis and contextual comprehension" (T_6). Taken together, these findings suggest that standardized assessment is perceived to provide structure and motivation for learning while simultaneously encouraging examination oriented preparation and, in some cases, factual memorization at the expense of inquiry and deeper historical understanding. These findings represent participants' perceptions and do not establish a causal effect of standardized testing on either motivation or historical understanding.

4.3 Use of Assessment Data for Feedback and Instructional Improvement

No.	Survey item	Teachers M (SD)	Interpretation	Students M (SD)	Interpretation
1	Standardized test results are released in a timely manner for instructional planning.	4.22 (.72)	Very High	3.85 (1.14)	High
2	I use test data to identify students' learning gaps and provide targeted support.	4.31 (.77)	Very High	—	—
3	Students receive meaningful feedback based on their test performance.	4.24 (.84)	Very High	4.13 (1.00)	High
4	Standardized tests help me modify teaching strategies for better outcomes.	4.44 (.60)	Very High	—	—

5	Assessment feedback leads to tangible learning improvement in subsequent lessons.	4.31 (.62)	Very High	—	—
6	I receive my History test results on time.	—	—	3.85 (1.14)	High
7	Teachers explain the results clearly and show me how to improve.	—	—	4.13 (1.00)	High
8	I understand my mistakes after getting test feedback.	—	—	4.28 (.91)	Very High
9	The feedback helps me do better in the next test.	—	—	4.18 (.91)	High
10	My teachers use our test results to plan extra support or revision.	—	—	4.16 (.97)	High

Teachers reported very high perceived use of standardized assessment data for instructional purposes, indicating that test data help identify learning gaps (M = 4.31, SD = .77), support modification of teaching strategies (M = 4.44, SD = .60), and inform subsequent learning support (M = 4.31, SD = .62). Students also viewed feedback positively, particularly in relation to understanding mistakes (M = 4.28, SD = .91) and improving performance in subsequent assessments (M = 4.18, SD = .91). One student explained that "the feedback shows my mistakes, so I can study those parts better next time" (S_3), illustrating how feedback was perceived as supporting reflection and targeted improvement.

An important point of mixed methods integration concerns the timeliness of feedback. Survey respondents generally rated reporting positively, with teachers indicating very high agreement that results are released in time for instructional planning (M = 4.22, SD = .72) and students reporting high agreement that they receive results on time (M = 3.85, SD = 1.14). However, the qualitative findings revealed that delays occurred in some contexts. As one teacher noted, "Standardized test results in Bhutan are often delayed, limiting their immediate usefulness for lesson planning" (T_4). Taken together, these findings suggest that assessment feedback is generally perceived as useful and timely, while its instructional value may be reduced when result dissemination is delayed. This divergence highlights the value of integrating quantitative breadth with qualitative evidence of contextual variation.

4.4 Challenges and Priorities for Reform

No.	Survey item	Teachers M (SD)	Interpretation	Students M (SD)	Interpretation
1	Standardized testing places too much pressure on students and teachers.	3.37 (.87)	Moderate	—	—
2	The current system focuses more on rote learning than critical historical inquiry.	3.32 (1.00)	Moderate	—	—
3	There is a lack of balance between formative and summative assessment approaches.	3.46 (1.00)	High	—	—
4	The assessment system requires reform to make it more learning-oriented.	4.12 (.74)	High	—	—
5	Teachers and schools should have greater input in designing History assessments.	3.90 (.94)	High	—	—
6	Standardized tests cause a lot of stress for students.	—	—	3.23 (1.10)	Moderate
7	The tests do not fully show what I know about History.	—	—	2.90 (1.10)	Moderate
8	I prefer assessments that include projects,	—	—	3.30	Moderate

	discussions, or essays.			(1.00)	
9	History tests should focus more on understanding than memorization.	—	—	4.09 (.98)	High
10	The testing system should be improved to make learning more interesting.	—	—	3.91 (.94)	High

Teachers reported moderate perceptions of excessive pressure (M = 3.37, SD = .87), a stronger emphasis on rote learning than critical historical inquiry (M = 3.32, SD = 1.00), and a perceived imbalance between formative and summative assessment (M = 3.46, SD = 1.00). Qualitative accounts reinforced these concerns. One teacher explained, "The most significant is the pressure to ensure high student performance, which often shifts the focus from deep learning to exam preparation. This can lead to teaching to the test, limiting opportunities for creativity, discussion, and inquiry-based learning that History naturally promotes" (T_6). Students likewise reported moderate examination stress (M = 3.23, SD = 1.10) and high agreement that History assessment should emphasize understanding rather than memorization (M = 4.09, SD = .98).

Participants also expressed clear interest in a more authentic and inquiry oriented assessment system. One teacher recommended "incorporating project-based and practical tasks that reflect real-life applications" and using "varied formats like quizzes, presentations, and group work to keep students engaged and encourage deeper understanding rather than rote memorization" (T_2). A student similarly suggested that History assessment should include "open-ended questions, source analysis, group projects, or presentations" and "real-life examples, stories, or visuals" so that assessment encourages thinking and creativity rather than factual recall (S_{12}). Taken together, the findings indicate stakeholder support for a more balanced assessment approach that retains standardized assessment while expanding formative, authentic, and inquiry oriented opportunities for demonstrating historical understanding.

4.5 Teacher Assessment Literacy and Professional Practice

No.	Item	N	M	SD	Interpretation
1	I am confident in analyzing and interpreting test data to guide instruction.	59	4.03	0.615	High
2	I have received adequate training on assessment design and data use.	59	2.49	1.12	Low
3	I collaborate with colleagues to address issues highlighted in test performance reports.	59	4.05	0.753	High
4	Continuous classroom assessment complements standardized testing in improving learning.	59	4	0.809	High
5	Assessment results are discussed with students to foster self-reflection and accountability.	59	4.39	0.644	Very High

Teachers reported high confidence in analysing and interpreting assessment data (M = 4.03, SD = .615) and in collaborating with colleagues to address issues identified in test performance reports (M = 4.05, SD = .753). They also agreed that continuous classroom assessment complements standardized testing (M = 4.00, SD = .809) and reported very high agreement that assessment results are discussed with students to encourage self-reflection and accountability (M = 4.39, SD = .644). However, formal preparation emerged as a clear concern, with adequate training in assessment design and data use receiving a low mean rating (M = 2.49, SD = 1.120). This concern was reflected in the qualitative data, as one teacher stated, "Hardly any training on assessment design but still need ongoing professional support in data analysis and effective strategies for utilizing assessment results" (T_6). Taken together, these findings suggest a distinction between teachers' self-reported confidence in routine assessment practices and their perceived need for more systematic professional learning in assessment design, data interpretation, and effective use of assessment evidence. Accordingly, the findings support sustained professional development as an important component of assessment reform, while not assuming that self-reported confidence alone constitutes evidence of comprehensive assessment literacy.

INTEGRATED DISCUSSION

The integrated findings reveal a pattern of broad support for the procedural and instructional functions of standardized History assessment alongside persistent concerns about the depth and breadth of disciplinary learning represented in current practices. Across both strands, respondents generally perceived standardized assessments as fair, curriculum-aligned, and useful for focusing academic effort. Teachers also reported extensive use of assessment results for instructional planning, while students described feedback as useful for understanding mistakes and preparing for subsequent assessments.

These positive perceptions should be distinguished from evidence of actual assessment validity. A perception that an assessment is fair, aligned, or representative of historical understanding is not equivalent to a technical validation based on test specifications, item analysis, criterion evidence, construct representation, or differential performance across groups. The present study therefore contributes evidence about stakeholder experience and perceived quality, not a psychometric evaluation of the standardized assessment system.

A second integrated theme concerns the tension between academic focus and disciplinary depth. Respondents perceived standardized testing as motivating preparation and helping students focus on important content, yet they also reported memorization, examination strategy, and reduced opportunities for inquiry. This apparent tension is consistent with scholarship indicating that assessment can simultaneously clarify expectations and shape learning toward the forms of knowledge most visibly rewarded by the test (Shepard, 2000; Wineburg, 2001). In History, the issue is therefore not simply whether standardized testing motivates students, but what forms of learning it encourages them to prioritize.

A third theme concerns the formative use of assessment information. Quantitative responses were strongly positive regarding feedback and instructional use, while qualitative evidence indicated that delayed reporting still occurs in some contexts. The mixed-methods interpretation is consequently conditional rather than categorical: assessment data appear to be perceived as useful when they are returned and discussed promptly, but their instructional value can be reduced where reporting is delayed. This finding reinforces the importance of timeliness, diagnostic clarity, and structured opportunities to act on feedback.

Finally, the teacher data reveal a meaningful gap between confidence and formal preparation. Teachers reported confidence in interpreting assessment information and collaborating around results, yet they rated formal training in assessment design and data use as low. This discrepancy suggests that professional competence may have developed partly through experience and collegial practice rather than through systematic professional learning. Sustained development in validity, reliability, item design, data interpretation, feedback, and discipline-specific assessment of historical reasoning may therefore strengthen both technical and pedagogical dimensions of assessment literacy.

CONCLUSION

This mixed-methods study examined teachers' and students' perceptions of standardized History assessment in Bhutan. The findings indicate generally favourable perceptions of fairness, objectivity, curriculum alignment, academic focus, feedback, and instructional use of assessment data. At the same time, participants identified concerns related to memorization, examination-oriented learning, limited inquiry opportunities, contextual inequities, occasional delays in feedback, and insufficient formal training in assessment literacy.

The central conclusion is not that standardized History assessment is either wholly effective or fundamentally deficient. Rather, stakeholders appear to value the consistency and accountability functions of common assessment while also expressing a need for broader opportunities to demonstrate historical reasoning. Because the study relies primarily on self-reported perceptions, its conclusions concern perceived assessment quality

and educational experience rather than direct evidence of test validity or causal effects on student learning. A balanced reform agenda would therefore preserve common standards while strengthening authentic historical inquiry, formative feedback, contextual fairness, and teacher professional learning.

RECOMMENDATIONS

The study recommends that national assessment authorities and curriculum agencies revise standardized History examinations to increase the proportion of higher-order and inquiry oriented tasks, including source analysis, evidence interpretation, comparison, argumentation, and contextual reasoning, while also reviewing assessment language and task design for clarity, accessibility, and contextual fairness. School leaders and History departments should strengthen the formative use of assessment by ensuring that results are discussed promptly, diagnostic information is shared with teachers and students, and structured follow-up activities are conducted to address identified learning gaps. The Ministry of Education, assessment authorities, and teacher professional-development providers should offer sustained training in assessment literacy, particularly in item construction, validity, reliability, data interpretation, feedback, and the assessment of historical thinking.

At the school level, History teachers should complement standardized examinations with projects, essays, discussions, source-based tasks, and authentic inquiry activities, while subject departments and professional learning communities should regularly analyse assessment evidence and share effective practices. Assessment-review committees should also incorporate teacher and student feedback when revising assessment policies and instruments. Finally, researchers and assessment agencies should undertake technical studies of actual examination papers and score data to examine construct representation, item functioning, fairness, and validity alongside stakeholder perceptions.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

Several limitations should be considered when interpreting the findings. First, the quantitative results are based on self reported perceptions and therefore cannot establish the actual validity of standardized History assessments or their causal effects on motivation and historical understanding. Future studies should address this limitation by directly examining assessment papers, item specifications, student work, classroom practices, and longitudinal learning outcomes. Second, the qualitative sample was intentionally small and suited to indepth exploration, but future research should include a broader range of participants and school contexts to strengthen qualitative representation. Third, subsequent studies should document the sampling frame, recruitment procedures, response rates, and school characteristics more explicitly to allow clearer judgments about representativeness. Fourth, because only an overall Cronbach's alpha of .80 was available, future research should report domain-specific reliability and examine the dimensionality of the questionnaire using appropriate psychometric analyses. Finally, as the present quantitative analysis was descriptive, future studies should employ inferential statistics and effect-size estimates to test differences across teachers, students, regions, and school contexts.

Future research should also move beyond perception data by evaluating whether assessment tasks adequately measure core dimensions of historical thinking, including sourcing, contextualization, corroboration, interpretation, and argumentation. Researchers should examine whether different groups of students have comparable opportunities to demonstrate these competencies and whether language, resources, or school context influence assessment performance. Comparative and longitudinal studies across regions and school settings would further strengthen understanding of equity, implementation, and the conditions under which standardized History assessment most effectively supports learning.

REFERENCES

1. Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
2. Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.
3. Brookhart, S. M. (2017). *How to give effective feedback to your students* (2nd ed.). ASCD.
4. Brookhart, S. M. (2018). *How to create and use rubrics for formative assessment and grading.* ASCD.
5. Brookhart, S. M. (2022). *Grading for formative purposes.* Routledge.
6. Brookhart, S. M., & Guskey, T. R. (2019). *Feedback for learning*. Routledge.

7. Brown, G. T. L., & Harris, L. R. (2023). Teacher assessment literacy: What do teachers know, believe, and do? *Assessment in Education, 30*(1), 1–20.
8. Carless, D., & Winstone, N. (2023). *Teacher feedback: A learning-focused practice.* Routledge.
9. Carretero, M., & van Alphen, F. (2024). Historical thinking and learning in contemporary classrooms. *Educational Psychology Review, 36*(1), 1–25.
10. Creswell, J.W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage Publications.
11. Datnow, A., & Hubbard, L. (2022). *Data-driven leadership*. Harvard Education Press.
12. Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268.
13. DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2022). Reconceptualizing teacher assessment literacy: A review of international evidence. *Educational Assessment, 27*(2), 165–188.
14. Dorji, T., & Gyeltshen, S. (2023). Timeliness and utility of assessment data in Bhutanese schools: Implications for instructional decision-making. *Journal of Himalayan Education Studies, 5*(1), 45–60.
15. Fastre, G., & Veenman, M. (2022). Designing assessment systems that promote deep learning: A review of contemporary evidence. *Assessment in Education, 29*(4), 512–530.
16. Harlen, W. (2014). *Assessment, standards and quality learning in primary education*. Routledge.
17. Harlen, W. (2018). The role of summative assessment in supporting learning. *Assessment in Education: Principles, Policy & Practice, 25*(3), 301–318.
18. Hattie, J., & Clarke, S. (2019). *Visible learning*: Feedback. Routledge.
19. Heritage, M. (2022). Strengthening formative assessment practice through assessment literacy. *Educational Measurement: Issues and Practice, 41*(4), 25–35.
20. Hopfenbeck, T., & Throndsen, I. (2022). Educational equity and assessment. *Assessment in Education, 29*(5), 601–615.
21. Hsieh, W., & Tsai, Y. (2023). Curriculum alignment and assessment fairness in Asian school systems. *Asia Pacific Journal of Education, 43*(2), 256–272.
22. Kazemi, Y., & Ashrafi, S. (2022). Student motivation and standardized testing. *Educational Studies, 58*(3), 345–363.
23. Kim, J., & Park, H. (2022). Barriers to effective data-driven instruction in Asian school systems. *Asia Pacific Education Review, 23*(4), 679–693.
24. Lee, Y., & Chong, S. (2023). Assessment motivation and learning behavior in high-stakes systems. *Learning and Instruction, 84*, 101702.
25. Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705–717.
26. Looney, J. (2021). Building capacity for assessment for learning. OECD Education Working Papers.
27. Levesque, S., & Croteau, J. (2023). Historical thinking and classroom assessment. *Journal of Curriculum Studies, 55*(6), 789–808.
28. Muller, C., & Colwell, J. (2022). Standardized assessment alignment in social studies. *Curriculum Perspectives, 42*(1), 45–58.
29. Nusrat, A., Rahman, M., & Chowdhury, T. (2022). The impact of assessment feedback timing on secondary students' learning engagement. *Assessment in Education, 29*(5), 632–648.
30. OECD. (2023). *Assessment and learning: Rethinking high-stakes examination systems*. OECD Publishing.
31. Pimental, J. (2019). Some biases in Likert scaling usage and its correction. *International Journal of Science and Applied Research, 45*(1), 183-191.
32. Popham, W. J. (2017). *Classroom assessment: What teachers need to know (*8th ed.). Pearson.
33. Schildkamp, K. (2020). Data-based decision-making for school improvement: Research insights and gaps. *Educational Assessment, Evaluation and Accountability, 32*(1), 1–26.
34. Schildkamp, K., & Poortman, C. L. (2023). Data teams for school improvement: A decade of research insights. *School Effectiveness and School Improvement, 34*(1), 1–22.
35. Seixas, P., & Morton, T. (2022). *The big six historical thinking concepts* (2nd ed.). University of Toronto Press.
36. Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

37. Stiggins, R. (2021). *Revolutionizing assessment: Empower students to own their learning*. Corwin.
38. Torres, A., & Lee, S. (2023). Memorization versus historical reasoning in secondary assessment. *Journal of Social Studies Research, 47*(4), 289–302.
39. Wiliam, D. (2023). Embedding formative assessment: Practical techniques for K–12 classrooms (2nd ed.). *Learning Sciences International*.
40. Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past.* Temple University Press.
41. Wylie, E. C., & Lyon, C. (2023). Formative assessment and timely feedback: What matters most for student learning? *Review of Educational Research, 93*(2), 267–298.
42. Xu, Y., & Brown, G. T. L. (2023). Teacher assessment literacy in practice: Evidence from classroom assessment studies. *Studies in Educational Evaluation, 76*, 101219.