

Targeting Local Benchmarks: A Framework for NLP-Based Automated Grading in Bangladesh's NCTB English Curriculum

Motia Mannan¹, Farhana Zabin^{2*}, Sabiha Akther³, Reshma Haque⁴, Mst. Sumaya Tasnim⁵,
Most. Hasna Banu⁶, Puspika Paul⁷, Ismat Ara Tabassum⁸, Mrityika Dey⁹, and Jannatul
Mawa¹⁰

¹Dept. of Education, University Tun Abdul Razak, Malaysia, Universal World Research Innovation Center, London, UK

²College of Education, Westcliff University,
Irvine, USA

³Department of Philosophy, Murari Chand College (M.C),
Sylhet, Bangladesh

⁴Department of Economics, Sorbonne University,
Paris, France.

⁵Dept. of English, First Capital University of Bangladesh,
Khulna, Bangladesh.

⁶Dept. of Politics & Public Administration, Islamic University,
Kushtia, Bangladesh.

⁷Dept. of Education, Noakhali Science & Technology University,
Noakhali, Bangladesh.

⁸Dept. of Education, Stockholm University Stockholm,
Sweden.

⁹Dept. of Statistics, University of Chittagong,
Chittagong, Bangladesh.

¹⁰Dept. of English Language & Literature, National University,
Chittagong, Bangladesh.

*Corresponding author: zabinfarhana4@gmail.com

Received: 25th June 2026; Accepted: 29th July 2026; Published online: 31st July 2026

Abstract

This study examines the methodological congruence between international standards and Natural Language Processing (NLP)-based assessment within Bangladesh's National Curriculum and Textbook Board (NCTB) English curriculum. Although AI-based grading systems are increasingly used in education, most remain built on Western language norms and tend to overlook local educational contexts and the influence of learners' first language (L1: Bangla). Using a qualitative research design, this study conducted content analysis of 200 student essays (100 at the SSC level and 100 at the HSC level) alongside semi-structured interviews with 15 local education professionals, including curriculum developers, teachers, and examiners. The findings reveal a high degree of contextual mismatch: existing NLP systems have no models calibrated to the NCTB level and tend to over-penalize non-standard but communicatively effective English. The absence of a locally digitized corpus further complicates the development of fair automated grading systems. By shifting the evaluative focus from native-speaker norms to locally competent communicators, this study underscores the importance of incorporating cultural and linguistic context into AI-based assessment

tools. It proposes a tentative policy framework to guide policymakers toward culturally responsive NLP systems that support fairer student assessment in Bangladesh.

Keywords: *NLP-based grading, English language assessment, automated grading systems, communicative competence, NTCB English*

INTRODUCTION

Educational AI assessment of learning is rapidly gaining importance in the world of education. The automated grading systems based on NLP (Natural Language Processing) minimize manual effort, expedite the grading process, and provide more uniform scores in large student populations (Ikäe et al., 2026). However, almost all mainstream NLP grading models were developed and trained using datasets from native English speakers in Western countries and are not easily used fairly in countries where English is a second or foreign language (Esomonu, 2025). Existing systems are primarily trained on native-style English writing. They see all variations from Western grammar as errors, even when they do not affect the meaning of the message. This limited approach does not allow for a fair assessment of learners whose English reflects the influence of their first language. It further establishes unfair criteria that privilege some learners over others. Such systems cannot really assess what learners can actually do with language.

The National Curriculum and Textbook Board (NCTB) is responsible for setting all national standards for English language teaching and assessment in Bangladesh. The main orientation of its core curriculum is toward oral and written communicative competence, which involves conveying meaning clearly and effectively, rather than imitating native speakers (Shabur et al., 2025). However, current AI grading programs do not align with this local educational objective. They tend to ignore sociolinguistic and cultural contexts, as well as the impact of the country's initial language (L1: Bangla) on how Bangladeshi learners write in English (Haque et al., 2025). This results in an urgent imbalance that jeopardizes fairness and equity in automated evaluation across the country.

Existing NLP-based grading systems are based on the Western attitude in which native-like English is the sole criterion for grading (Khan et al., 2026). This design has repercussions for those students whose first language is not English, such as non-native learners in Bangladesh, who speak English as a second language and Bangla as their primary language. L1 transfer (Lima, 2023) is a natural process that occurs when learners transfer Bangla language patterns to their English language writing. In some cases, students might avoid articles, use tense patterns in Bangla style, or use different words to convey meaning; such variations are not necessarily an impediment to communication, but they are not exactly what is expected in the West.

The NCTB framework prioritizes communicative effectiveness over grammatical perfection: writing that contains grammatical errors but still conveys the intended message should not be penalized, provided the message remains understandable (Rahman, 2025). Notably, however, such L1-influenced patterns are frequently classified as "errors" by automated systems, which deduct significant points for phrases like "*I go market yesterday*" or "*He give me book*," even though these are perfectly understandable (Afunugo & Molokwu, 2025).

Language Assessment in Bangladeshi English Education

The assessment of English in Bangladesh at the secondary and higher secondary levels is based mainly on the National Curriculum and Textbook Board (NCTB), and it focuses on communicative competence rather than native-like fluency. The term communicative competence refers to learners'

ability to make sense, organize their ideas, and communicate effectively in writing (Kabir, 2024). It is an orientation that contrasts with Western assessments of language, which focus more on conforming to native-speaker language norms (Mannan & Maruf, 2025). Fluency, coherence, meaning construction, and grammatical accuracy are therefore regarded as the criteria for writing in the rubrics of NCTB. However, attention to communicative teaching objectives while maintaining formal correctness may create tension between them (Arafat & Shamsuzzaman, 2025; Imam et al., 2025). The communicative competence of Bangladeshi learners becomes an issue, especially in the context of NLP-like grading systems that treat language correctness as the main focus of assessment, which may result in unfair grade awards. This tension is particularly significant when NLP-based grading systems prioritize language correctness over communicative competence, leading to unfair outcomes for Bangladeshi learners (Mita et al., 2024).

Another common phenomenon in Bangladeshi students' writing is L1 transfer from Bangla to English (Das et al., 2025). It could affect vocabulary, sentence construction, the use of articles, tense, and other Bangla translations and borrowings (Lima, 2023). Such features can be considered errors by automated systems but do not necessarily block meaning and could be a natural process of second-language acquisition (El Tayara et al., 2026). Thus, these features should be assessed for comprehensibility and task fulfillment, not automatically disallowed. This mismatch also helps justify the need for an assessment system based on locally grounded NLP that corresponds to the communicative goals of the NCTB curriculum.

So far, most research on AI-based assessment has been conducted in Western contexts; very little research has examined the alignment of AI with national curriculum objectives in the South Asian context, and almost none has proposed grading models suited to the requirements of the NCTB (Mannan et al., 2025). Without a locally adapted framework, automated grading will continue to generate biased results, misassess students' abilities, and further exacerbate inequities in assessment. What is needed, then, is an NLP-based grading system that aligns with NCTB pedagogy while also accounting for the language realities of learners in Bangladesh.

This study examines narrative and expository essay writing among students in Grades 10 and 12 (Secondary School Certificate (SSC) and Higher Secondary School Certificate (HSC), respectively) in two significant national secondary-level exams in Bangladesh. It focuses particularly on Essay 2 and Essay 3 of the official NCTB English syllabus, where Essay 2 is a personal experience or description essay in which the student describes events or shares personal experience; whereas Essay 3 is a descriptive/explanatory writing task in which students describe, explain, or present ideas, arguments, or procedures. These two formats are common in continuous assessment and final examinations at the national level. The principles and framework developed in this work can be further expanded for other types of writing and age groups in future research. This study goes beyond developing NLP for the technical side and also encompasses equity in assessment. It offers insights and guidance for designing AI tools aligned with local teaching objectives (Ahsan, 2025), which could be valuable for NCTB policymakers, teachers, and education technology developers. It is also used in South Asia and other regions of the Global South for global research in the field of inclusive NLP, in situations where English is not the native language.

The scholars provide a review of different AI assessment tools. It also examines access issues in the academic world, paying particular attention to educational equity. The focus is on NLP practices and models from a Western and South Asian perspective. This paper introduces the discussion of AI in Education within the linguistic, educational, and assessment contexts of Bangladesh. It acknowledges the increasing demand for an NLP-based assessment system for English writing that can assess students' writing according to NCTB learning outcomes, classroom practices, and the use of English in the local context. Instead of using native-speaker English as the only standard, the study

emphasizes the need for culturally and linguistically responsive AI systems that can distinguish between errors that negatively affect meaning and communicative language variations. This can help ensure fair and transparent evaluation of students from Bangladesh and minimize the chances of algorithmic bias. The suggested framework also extends the ongoing conversation about inclusive educational technology by providing an example of how locally developed benchmarks can be used to inform the creation of globally applicable AI tools in non-native English and multilingual settings.

RESEARCH OBJECTIVES

The objectives of this study are to examine how well global NLP-based automated grading criteria meet the communicative competence objectives of the English curriculum at the NCTB (SSC and HSC) level for 10th- and 12th-grade students in Bangladesh. In this study, communicative competence is defined as students' capacity to communicate meaning clearly, respond to the writing task, organize ideas coherently, employ contextually appropriate vocabulary and language, and produce comprehensible written communication even if minor grammatical errors appear in their writing. The first objective is to identify pedagogical and linguistic differences, particularly those arising from learners' first language (L1) influence on second language (L2) learning, and to examine how these differences shape written L2 production and affect the fairness of automated grading. The second objective is to propose a locally grounded grading framework that aligns with the NCTB's emphasis on communicative competence and reflects how Bangladeshi students learn and use English. To support these objectives, the study examines how existing automated grading systems respond to L1-influenced linguistic features, compares NCTB writing rubrics with NLP-based grading outcomes, and incorporates teachers' perspectives on fair assessment. Overall, the study aims to demonstrate how automated grading can be adapted to produce fairer, more contextually appropriate assessment outcomes for ESL learners in Bangladesh.

LITERATURE REVIEW

Automated Grading in NLP: Global Benchmarks and Approaches

NLP in the field of educational assessment has gained increasing interest in recent years due to its efficiency and scalability, largely because of its ability to automate grading processes (Vetrivel et al., 2025). Global benchmarks are standardized datasets, tasks, and evaluation measures used to compare the performance of NLP systems with state-of-the-art performance (the best results currently obtained by available models). These include the Conference on Computational Natural Language Learning (CoNLL) (Bozyigit et al., 2024), General Language Understanding Evaluation (GLUE) (Kobayashi et al., 2024), and the BEA Grammatical Error Correction Shared Task, which assesses systems for detecting and correcting errors in written language (Bohdal et al., 2026). The main focus of these benchmarks is on grammatical accuracy, coherence, language understanding, and error correction. Most of them, however, are developed and used in environments where English is the native language, and they tend to be influenced by Western/Standard English norms. CoNLL-related tasks are often seen as focusing on part-of-speech tagging and syntactic parsing (Neuberger, 2026); they are criticized for assigning too much weight to grammatical structure. For instance, Salifu et al. (2026) maintain that the CoNLL-based evaluation tasks tend to focus more on grammatical parsing. Zhao et al. (2026) have similar concerns about GLUE. However, even though GLUE includes several language-understanding tasks, Emara (2025) points out that language training may still perpetuate native-speaker expectations. The BEA Shared Task aims at error detection and correction in writing and can be used to support automated systems of error feedback. It is also often accompanied by expectations of "native-like" fluency, however. The benchmarks do not adequately take into consideration non-native language variation, which can lead to over-penalization or misclassification

of learner features such as article omission, tense variation, and pluralization errors (Pell et al., 2026; Asad, 2025). They also tend to neglect the first-language transfer phenomenon and may not be able to address pedagogical aspects like communicative competence (Kim, 2026). In general, current NLP benchmarks are too poorly suited to the non-native English context because the language learning and assessment goals may differ from those used in the context of Western education. The restrictions are especially relevant in Bangladesh, where English is studied as a second language and evaluated based on country-specific curricular requirements.

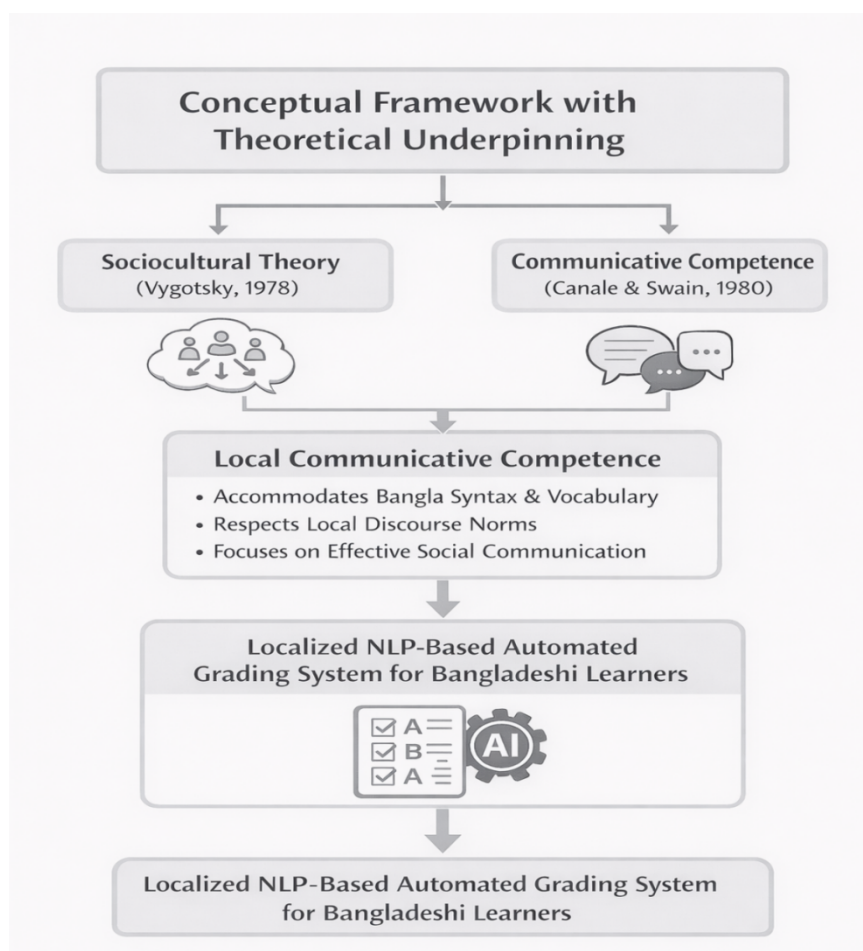
NLP Tools in Education

The use of NLP tools in the context of education has mainly been studied in Western countries, such as in the USA, which has implemented automated feedback systems, automated grammar or error-correction software, and automated essay-scoring systems. Although there is little research on non-Western varieties of English, the area of NLP tools for Indian English is even less explored. Indian English is influenced by a multilingual linguistic environment (Biju & Gayathri, 2025). There have been a few studies on the effects of multilingualism on NLP. For instance, Kaur et al. (2026) explore the effects of multilingual backgrounds on Indian learners' use of English as L1-influenced vocabulary, sentence construction, and code-mixing. These features can make NLP systems, which are trained primarily on standard English, less accurate and can lead to acceptable learner expressions being interpreted as mistakes. These issues have also been studied in Africa; for instance, Leuckert et al. (2026) explored the challenges of improving NLP performance for Nigerian English.

The localized NLP tools in Bangladesh are still limited despite these developments. This is one of the identified gaps by Islam et al. (2025) in the context of NLP applications in South Asia. While the literature on multilingual language variation and regional NLP development offers some insights into Bangla-English transfer patterns, NCTB writing rubrics, and the pedagogical and assessment methods of Bangladeshi classrooms, it rarely captures any of these aspects (Mannan et al., 2023). Surprisingly, there are not enough studies available in the literature regarding the application of NLP in the classroom environment in Bangladesh, including the creation of localized datasets and assessment tools. To fill these gaps, the present study proposes a framework grounded in Bangladesh's sociolinguistic context and the communicative purpose of the NCTB curriculum.

Conceptual Framework and Theoretical Underpinnings

The theory used for this study is Communicative Competence theory, developed by Canale and Swain (1980) (Bachman & Palmer, 2024). This theory shifts the meaning of language ability: it is no longer considered accuracy in language, but rather effective communication. It aims to apply language in real-world contexts for its intended purpose. Communicative competence in this study is defined functionally. It is the ability to express ideas clearly in a learner's language, regardless of differences in language norms from native English speakers. The study takes this concept a step further to encompass "local communicative competence." This concept reflects the linguistic reality of Bangladeshi learners. It explains the influence of L1 (Bangla) on the syntax, lexis, and discourse of L2 (English). The framework is also based on Sociocultural Theory (Vygotsky, 1978). This theory emphasizes the significance of social and cultural context in language learning (Sato et al., 2025). It affirms the notion that language is inseparable from the learning context. This study presents an NLP framework for automated grading based on these theories. The framework is in line with the pedagogical aims of the NCTB. It also takes into account local linguistic characteristics that are typically non-standard or non-conventional for traditional NLP systems. The proposed Automated Grading System (AGS) is designed to grade Bangladeshi learners more fairly. It is based on communicative competence and sociocultural theory. It aims to assess students' writing in a challenging, real-world, and meaningful manner.

Figure 1*Conceptual Framework***Synthesis and Research Gap**

While a lot has been achieved in automated grading and NLP benchmarks, there is still a significant research gap to be addressed. The current research is largely based on Western models of NLP and their technical deficits. However, little has been done on the issue of their acceptability to Bangladeshi learners. Neither the specific structure of the Bangladeshi education system nor the specific problems of the education system in Bangladesh are covered in most studies. In particular, there is a low level of integration of NCTB pedagogical goals with NLP tools. This leaves a mismatch between NLP development and local educational needs. In addition, the current study does not pay much attention to the influence of Bangla (L1), which causes linguistic variation. This deviation is occasionally considered an error in automated systems. This sparks doubts about the fairness of automated assessment. In light of this, the study fills the gap by suggesting a localized NLP framework for the Bangladeshi context. The framework is aligned with the NCTB pedagogical goals. It also takes into consideration the sociolinguistic and linguistic realities of Bangladeshi learners. This is designed to ensure fair and context-appropriate grading systems based on AI.

METHODOLOGY

In this study, a qualitative comparative research design and methodological triangulation are used. This design was chosen because the study aims to interpret the alignment and differences between the practice of automated grading using a global NLP approach and the communicative assessment approach required by the English curriculum of NCTB, Bangladesh. It enables the researchers to

investigate the different understandings of language competence and fairness reflected in the outputs of automated writing assessment and teachers' assessment practice, and in the NCTB writing rubrics.

Two dimensions are studied, which are closely related to each other. The first is the technical dimension, which centers on the use of NLP-based grading systems for assessing students' writing, including L1-influenced language features, vocabulary use, task fulfillment, coherence, and grammatical accuracy. The second is the contextual and pedagogical dimension, which takes students' perceptions of Fair Grading, Classroom Assessment, and the NCTB curriculum into account. Considering these dimensions simultaneously can help identify points of alignment and misalignment between the criteria of the auto-grader and the educational setting of Bangladesh. Content analysis of students' essays, comparison of the grading results produced by students' writing and the ratings made by the NCTB with the grading results produced by the NLP, and semi-structured interviews with English teachers are used for methodological triangulation. Content analysis allows the identification of linguistic and communicative aspects of a student's writing, as well as the analysis of the evaluation of these aspects by automated systems. The rubric comparison is to determine whether the NLP grading outcomes match the NCTB assessment criteria. The interviews provide teachers' views on the fairness, accuracy, and context-appropriateness of automated grading. The findings from both sources support each other to draw a more robust conclusion and provide a fuller picture of automated evaluation of English writing in Bangladesh.

Study Participants

Researcher Description

The research is conducted by a total of 10 researchers who are specialized in education, linguistics, TESOL, English language, philosophy, economics, politics and public administration, and statistics. The first author specializes in multiple disciplines, including education and TESOL, and has more than 15 years of English teaching experience in Bangladesh and various countries abroad. This has made her familiar with both NLP and NCTB assessment practices and ESL teaching. The second author specializes in linguistics and has more than five years of English teaching experience in Bangladesh and the USA. Both the second and fourth authors specialize in TESOL and the English language field. These experiences help them understand the target learners' linguistic proficiency and learning gaps in a second language and propose solutions to mitigate those gaps. The third and ninth authors are currently teaching in Bangladesh, which helps them understand and analyze the current skills and limitations of ESL learners. The fourth, sixth, seventh, eighth, and tenth authors specialize in education, English language, economics, politics, and public administration and have studied in Bangladesh and abroad, which helps them collect data and view it through a broader lens by comparing their experiences in both contexts. Overall, the researchers' familiarity with both Bangladesh and the broader world informed the analysis of student essays and the interpretation of educator interviews. To minimize the influence of individual assumptions on data interpretation, the researchers applied the verification procedures detailed under methodological integrity, including inter-coder reliability and member checking.

Sources of Data

This study drew on two primary data sources: student essays and semi-structured interviews with education professionals. The essay corpus consisted of 200 essays, evenly split between 100 essays produced by Grade 10 students at the Secondary School Certificate (SSC) level and 100 essays produced by Grade 12 students at the Higher Secondary Certificate (HSC) level. The interview data consisted of 15 semi-structured interviews with three groups of education professionals: five English teachers, five curriculum developers, and five examiners. Because participants were selected on the basis of their professional role and area of work rather than personal characteristics, demographic data such as gender and age were not collected.

Sampling Procedures

Essay Sampling. Essays were selected using purposive sampling to ensure direct relevance to the NCTB English curriculum and the study's research objectives. To be included, an essay had to meet the following criteria: it was written by a Grade 10 or Grade 12 student enrolled in the NCTB curriculum writing program; it responded to one of the two designated writing tasks (Essay 2 or Essay 3) from the NCTB English curriculum; it contained sufficient written content to allow a judgment of communicative competence and language use; it was complete and legible; and it met the required minimum word count. Essay 2 involved personal-experience, narrative, or descriptive writing, while Essay 3 involved descriptive or explanatory writing involving the presentation of ideas, arguments, processes, or procedures. Essays were excluded if they were incomplete, off-topic relative to the selected tasks, submitted by students outside the two targeted grade levels, or produced under an international or other non-NCTB curriculum. These criteria were applied consistently using a researcher-developed essay-selection checklist.

Participant Sampling. Interview participants were also selected through purposive sampling, based on their first-hand experience teaching, developing, implementing, or evaluating English writing assessment under the NCTB curriculum.

DATA COLLECTION AND ETHICS

Data collection took place in May 2026 at secondary and higher secondary institutions in Bangladesh that follow the NCTB curriculum. These institutions were selected from among schools and colleges using the NCTB curriculum within the study areas. The 200 essays were collected as de-identified documents, consistent with the sampling criteria described above. Interviews were conducted during the same data collection period at the participating institutions, with each interview lasting approximately 45 minutes. A researcher-designed semi-structured interview guide was used to ensure uniformity across the three participant groups while allowing rich, detailed accounts of professional experience. The guide included open-ended questions addressing writing assessment practices under the NCTB, communicative competence, Bangla-influenced English, the fairness of NCTB grading principles, and the perceived need for a locally relevant automatic grading system. Follow-up questions were posed as needed to clarify or expand on participants' responses.

Permission to collect essays from SSC- and HSC-level students was obtained from the principals of participating institutions prior to data collection. Students' names, institutional identifiers, and other identifying information were removed from all essays before inclusion in the study. Prior to each interview, participants were informed of the study's purpose and procedures, its voluntary nature, and their right to withdraw at any stage. Each participant provided informed consent, including consent to audio-record, before the interview began. All identities—including those of students and interview participants—were anonymized, and all collected materials were stored in a confidential research data repository to ensure no individual could be identified.

DATA ANALYSIS

The students' essays were qualitatively analyzed using manual content analysis. The analysis was organized with a coding sheet developed by the researcher. The coding categories were based on the NCTB writing criteria, the study objectives, task fulfillment, content relevance, organization, coherence, vocabulary use, grammatical accuracy, clarity, and communicative effectiveness. The researchers also manually extracted features related to Bangla-to-English transfer, such as article omission, tense variation, lexical choice, and sentence structure. All essays were read using the same

categories to maintain consistency between the SSC and HSC samples. Thematic analysis was used to analyze the interview data. The recorded interviews were transcribed, and the researchers read the transcripts multiple times to become acquainted with the participants' answers. Initial codes were assigned to relevant statements, and similar codes were clustered into categories, with related categories forming broader themes. The themes that emerged included assessment practices in the NCTB, communicative competence, Bangla-influenced English, and perceptions of the fairness and restrictions of grading based on NLP. These analyses do not appear as findings, but are presented later in the Results section.

Methodological Integrity

To enhance the reliability of the manual analysis, a small sample of essays and the transcripts of the interviews were independently coded by two researchers using the same coding categories. Their coding decisions were compared, and any differences were discussed until they reached agreement. The agreed interpretations were then used to clarify the coding categories before the remainder of the data were analyzed. To informally test the preliminary interpretation of the interviews, summaries of the analysis were presented to a selected sample of participants, who were asked to verify that the summaries were accurate and captured the nature of their interpretations. Changes and clarifications were made to the final analysis.

FINDINGS

The results are reported in two ways: the examination of 200 student essays from manual analysis and thematic analysis of 15 educator interviews.

Findings from the Essay Analysis

The essay analysis revealed common features of Bangla influence in English writing, including the omission of articles, variation in tense marking, subject-verb agreement errors, lexical variation, and non-standard sentence structure. For example, article omission appeared in students' essays as "I go to market yesterday"; non-standard tense marking appeared as "He give me book every day"; and Bangla-style word order appeared as "Very beautiful the scenery was." These patterns were evident across both SSC- and HSC-level essays, although their impact varied by response. In many cases, students were able to convey their intended meaning, complete the task, and organize their ideas despite grammatical errors. In other cases, however, repeated linguistic difficulties reduced clarity and coherence. Nevertheless, the data indicate that the meaning-blocking error rate was only 11% overall, 13% at the SSC level, and 9% at the HSC level, indicating that most of the linguistic features observed in these essays still preserved the intended meaning.

The analysis therefore distinguished between errors that affected formal accuracy and errors that hindered communication. It also revealed that grammatical correctness is not the sole criterion for evaluating writing; task fulfillment, coherence, organization, vocabulary, clarity, and communicative effectiveness are equally important. These results address the first research question by demonstrating that L1-influenced language features can affect the fairness of automated grading.

Findings from the Educator Interviews

The interview analysis yielded three key themes. First, teachers, examiners, and curriculum developers viewed meaning, task fulfillment, coherence, and organization as central to the assessment process. In response to the question, "In your professional judgment, what takes priority: grammatical accuracy or clear, effective communication?", one teacher prioritized communication over

grammatical accuracy, stating, *“Clear communication is always first. Umm... grammar matters, but only if it actually stops the reader from understanding the message. I think, umm... a small mistake that does not confuse anyone should not lose marks.”* Similarly, in response to the question, “Could you describe how English writing assessment works under NCTB guidelines at SSC and HSC levels?”, an examiner pointed to the official mark distribution, explaining, *“The marking guide is actually clear.” Uh... task completion, coherence, and clarity together carry 80% of the total marks, and grammar is only 20%.* Although participants considered grammar important, they did not view all grammatical errors as equally serious. In response to the same question about priority between grammatical accuracy and effective communication, a curriculum developer expressed a similar position: *“I think... perfect grammar with no clear message is a failure. Uh... clear communication with minor errors is a success.”*

Second, participants identified Bangla influence in the articles’ tense, vocabulary, word order, and sentence construction. When asked the question, “How does students’ first language (Bangla) shape their English writing? What patterns do you see regularly?”, a teacher provided a common example from students’ writing and explained, *“Actually, tenses are the most common example. Uh... like many students wrote ‘I go to market yesterday’, and any local person understands this perfectly, but any AI will mark it as a major error.”* Again, asking the same question to a curriculum developer, he pointed out the pattern of the Bangla language and said, *“Bangla has no articles, so, yes... students naturally omit them. Bangla marks tense differently, so our students do too. Ah... these are consistent patterns across all regions.”* They noted that such features should be evaluated by their impact on meaning. Minor deviations were deemed to be less serious if there was no confusion about the meaning of the message. To answer the question, “Should these Bangla-influenced patterns be treated as errors? Why or why not?”, an examiner expressed, *“Yes, we treat these as acceptable learner language, not errors. And that is our official instruction.”*

Third, participants supported the development of an automated grading system aligned with the NCTB curriculum framework. In response to the question, “What key features must a fair, locally appropriate NLP grading system have for NCTB students?”, one curriculum developer stated, *“The system must align fully with NCTB standards, definitely, umm... not external native-speaker expectations. Yes, of course, it needs separate scoring dimensions for meaning and form.”* Participants proposed that such a system should measure communicative effectiveness alongside grammatical correctness, while also accounting for the typical patterns found in Bangladeshi learners’ English. They also identified the lack of local, teacher-annotated student writing data as a major obstacle to developing context-sensitive grading practices. In response to the question, “Is there enough annotated, local student writing data available to build such tools? What is missing?”, the same curriculum developer explained, *“Not really. No locally annotated corpus actually exists. Uh... I think this is the single largest barrier to developing fair, context-appropriate tools.”* Similarly, when asked the same question, an examiner explained why such data does not survive, stating, *“No organized collection exists. Ah... we only keep papers for a short time before they are destroyed.”*

Synthesis of Findings

Table 1 summarizes the key findings from the essay and interview analyses. Together, these results suggest a need to differentiate between language deviations that reduce formal accuracy and those that significantly affect meaning, particularly in the context of fair automated grading in Bangladesh.

Participants also advocated for incorporating NCTB's assessment criteria, Bangla-influenced linguistic patterns, and teachers' professional judgment into the design of a locally informed grading system.

Table 1

Summary of Key Findings and Implications for Automated Grading

Finding	Evidence	Implication
Some Bangla-influenced features were found in SSC-level and HSC-level essays.	Variation in the use of articles, tenses, agreement, vocabulary, and sentence structure was detected through manual analysis.	Common patterns of local learner languages should be acknowledged in automated grading.
Formal errors did not always interfere with meaning.	A few essays were clear, coherent, and task-relevant, although with variations in grammar.	Communicative competence should be evaluated separately from grammatical accuracy.
Educators valued both meaning and accuracy.	The interview participants spoke about tasks fulfilled, clarity, coherence, and comprehensibility.	The communicative impact of an error should be scored.
Grading criteria must be in keeping with the NCTB.	The findings from the essay and interview revealed that learners' writing went beyond conformity to standard English.	There is a need to use the NCTB criteria and educators' professional judgment in automated grading.
Writing data are annotated locally, and data are limited.	Participants pointed to a lack of context-specific SSC- and HSC-level writing data.	It is recommended to build an NCTB-aligned learner-writing corpus.

Note. SSC = Secondary School Certificate; HSC = Higher Secondary Certificate; NCTB = National Curriculum and Textbook Board.

DISCUSSION

The results of this study reveal that grammatical accuracy and communicative effectiveness are not necessarily correlated in Bangladeshi students' writing. Manual analysis identified consistent problem areas in the use of articles, verb tenses, subject-verb agreement, vocabulary, and sentence construction. However, many essays remained understandable, coherent, and relevant to the required task despite these errors. Writing quality should therefore be judged based on task fulfillment, organization, clarity, coherence, meaning, and grammatical accuracy.

The results of the interviews confirm this. Grammar was important to teachers, examiners, and curriculum developers; they did not consider minor forms as errors that hindered communication. Their views show the communicative thrust of the NCTB curriculum and indicate that Bangla-influenced forms should be judged based on their contribution to meaning and should not be given a one-size-fits-all treatment as evidence of lower proficiency. The results of this study are relevant to the first research question because they produce a list of language and pedagogical factors that could contribute to unfairness in automated grading.

The study was not conducted with a specific automated grading model. However, the results point to possible weaknesses in grading models that focus mainly on standard or native-language English for Bangladeshi learners. The following aspects should be evaluated in an NCTB-aligned system: task fulfillment, coherence, organization, vocabulary, clarity, communicative effectiveness, and grammatical accuracy. It should also be capable of identifying meaningful variation in the local context from errors that significantly disrupt meaning.

The findings address the second research objective of developing a locally informed grade structure. It should incorporate criteria from NCTB assessment, teachers' professional judgment, and common traits of student writing that Bangla influences. An annotated corpus of SSC- and HSC-level essays will further facilitate the training and validation of a context-sensitive automated essay grading system. These measures may help to achieve fairer and more educationally appropriate assessment in Bangladesh.

Consequences of Localizing NLP Models

The directions suggested in this study are for the development of NLP tools aligned with the NCTB curriculum. Based on the results, an automatic grading system for the local area is proposed. This framework covers contextual criteria related to Bangladeshi learners. It is based on three pillars: rubric mapping localization, the creation of a localized annotated corpus, and educator feedback integration.

Rubric Mapping: Localization

Existing NLP systems follow Western models of the way people learn language. These models are based on grammaticality and naturalness. However, NCTB focuses on communicative competence (Mannan & Maruf, 2024). Thus, an NLP system should incorporate rubric-based contextual mapping. It needs to reflect the meaning and effectiveness of communication. It should also identify the English forms influenced by Bangla and those that assist communication. This helps minimize inappropriate penalization of non-conforming but meaningful use of language.

The Creation of a Localized Annotated Corpus

To improve NLP systems for Bangladeshi English, it is necessary to have a localized corpus. There are currently only limited datasets like these. A corpus of students' essays written in Bangla should be annotated. Common linguistic patterns and errors should be included in this corpus. This should also contain L1-influenced language structures. Using this corpus to train NLP models will help adapt the models to local language usage. It will also improve the accuracy of detecting context-specific writing errors.

Educator Feedback Integration

Assessment is a crucial part of education and is directly involved with educators. They are key to the fair design of NLP systems. The results of this study indicate that teachers value communicative competence more than grammatical correctness. Thus, educators' views need to be incorporated into system design. This will ensure that learners are not disadvantaged by automated grading systems. Additionally, this enhances alignment between teacher assessment and AI systems.

POLICY AND PRACTICAL IMPLICATIONS

The findings of this research have policy and practical implications for education in Bangladesh. The NCTB calls for integrating EdTech into assessment policy based on this research. The NCTB emphasizes the need to integrate EdTech into its policies. It also emphasizes the importance of culturally responsive, AI-based evaluation systems. Local corpora should be encouraged. This will help develop more precise NLP models for Bangladeshi learners. AI systems need to be developed in the context of local education. Otherwise, they will not be able to work effectively in Bangladesh. NLP system development must be done in a participatory manner. Close cooperation between developers and teachers is critical. Teachers need to be part of the design and assessment of the

system. This will make ESL classroom instruction relevant and equitable. Further, teachers need training in AI-based assessment. They should be knowledgeable about the proper use of these automated tools. Scoring should be handled by humans, with the help of AI.

LIMITATIONS AND FUTURE RESEARCH

This study has several limitations. First, the analysis was based on 200 essays; a larger dataset would improve the generalizability of the findings. Second, the study employed a qualitative research design, which added interpretive richness but may also have introduced some degree of subjective bias. Future studies should therefore incorporate quantitative validation approaches, using larger and more diverse datasets to further test the proposed framework and compare the results against standardized grading systems (Mahmud et al., 2024). Third, this study focused solely on the assessment of English writing; future work should extend the framework to other subject areas and grade levels, as well as examine regional variations in language use across Bangladesh to enhance the model's flexibility. Overall, future research should aim to benchmark the framework through both empirical and computational studies.

CONCLUSION

This paper seeks to address the gap between global NLP benchmarks and the NCTB English curriculum in Bangladesh, examining its implications for the computerized assessment of L2 writing. The findings reveal a clear disconnect between NCTB rubrics and global NLP evaluation standards, largely driven by the strong influence of L1 (Bangla) on students' writing. This contextual mismatch stems from a fundamental tension: NLP models are developed according to Western language norms, whereas the NCTB curriculum prioritizes communicative competence over native-like accuracy. Many NLP systems are capable of detecting deviations and mistakes in sentence structure, such as article omission and non-standard word order. However, these deviations often stem from L1 transfer and do not necessarily reduce communicative effectiveness, even though automated systems frequently penalize them regardless. This illustrates the need to localize NLP models: evaluation systems should prioritize meaning and communicative effectiveness over conformity to native-speaker norms. Figure 2 presents a proposed model for localizing NLP grading systems to better align with the NCTB English curriculum based on the results of this study.

Figure 2
Localizing NLP Grading Model for NCTB English



NOVELTY, CONTRIBUTIONS, AND IMPLICATIONS

The novelty of this study lies in the design of a context-specific NLP-based grading framework in the context of Bangladeshi ESL education and assessment. The framework is based on NCTB's pedagogical goals and learners' linguistic realities. It shifts from native-like fluency to communicative competence. It also helps to create culturally responsive AI systems in the field of learning. The results reveal the significance of AI localization in education beyond Bangladesh. There are other contexts in the Global South where English is an L2 language and where such problems are found. This study suggests balancing global NLP design with local educational needs. It highlights the importance of fairness in AI-based assessment systems. It also contributes to educational systems thinking in AI design. The aim is to minimize disparities in automated grading among linguistically diverse populations.

FUNDING

There is no specific grant or funding from any public, commercial, or profit or non-profit organization.

REFERENCES

- Afunugo, K. N., & Molokwu, G. C. (2025). AI-driven early detection and prevention of youth cult recruitment in Awka metropolis. *Iaqjah-igwebuike: An African Journal of Arts And Humanities*, 11(7). https://www.igwebuikeresearchinstitute.org/journal/igwebuike_1755521646.pdf
- Ahsan, T. (2025). Mainstreaming the concept of equity and inclusion into Bangladeshi education: What can we learn from the experience? In *Conversations and Key Debates on Inclusive and Special Education* (pp. 7–22). Routledge.
- Arafat, E., & Shamsuzzaman, M. (2025). Understanding students' perceptions and abilities of second language (L2) writing in English across disciplines at the tertiary level in Bangladesh. *MEXTESOL Journal*, 49(2). <https://doi.org/10.61871/mj.v49n2-11>
- Asad, M. (2025). The impact of mother tongue interference on English syntax: A case study of EFL students at Technical College, Saudi Arabia. *International Journal of Language, Linguistics, Literature and Culture*, 4(3), 13–30. <https://doi.org/10.59009/ijlllc.2025.0118>
- Bachman, L. F., & Palmer, A. S. (2024). The construct validation of some components of communicative proficiency 1. In *The Writings of Lyle F. Bachman* (pp. 41–59). Routledge.
- Biju, P. R., & Gayathri, O. (2025). Indic approach to ethical AI in automated decision-making system: implications for social, cultural, and linguistic diversity in native population. *AI & Society*, 1–26. <https://doi.org/10.1007/s00146-025-02381-z>
- Bohdal, O., Ceritli, T., Ozay, M., Moon, J., Lee, K. H., Ko, H., & Michieli, U. (2026). Data-driven Clustering and Merging of Adapters for On-device Large Language Models. *arXiv preprint arXiv:2601.17441*. <https://doi.org/10.48550/arXiv.2601.17441>
- Bozyigit, F., Bardakci, T., Khalilipour, A., Challenger, M., Ramackers, G., Babur, Ö., & Chaudron, M. R. (2024). Generating domain models from natural-language text using NLP: a benchmark dataset and experimental comparison of tools. *Software and Systems Modeling*, 23(6), 1493–1511. <https://doi.org/10.1007/s10270-024-01176-y>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1 to 47. <https://doi.org/10.1093/applin/I.1.1>
- Das, D., Guha, S., & Semaan, B. (2025). How do datasets, developers, and models affect biases in a low-resourced language? *arXiv preprint arXiv:2506.06816*. <https://doi.org/10.48550/arXiv.2506.06816>
- El Tayara, H., Milad, M., & Farran, M. (2026). Pedagogical Application of Artificial Intelligence to English Language Teaching Contexts in Arab World Higher Education. In *Higher Education in the Arab World: Artificial Intelligence* (pp. 493–515). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-99068-7_19
- Emara, I. F. (2025). A linguistic comparison between ChatGPT-generated and nonnative student-generated short story adaptations: a stylometric approach. *Smart Learning Environments*, 12(1), 36. <https://doi.org/10.1186/s40561-025-00388-z>
- Esomonu, N. P. M. (2025). Utilizing AI and big data for predictive insights on institutional performance and student success: A data-driven approach to quality assurance. In *AI and Ethics, Academic Integrity and the Future of Quality Assurance in Higher Education* (pp. 625–630). Sterling Publishers.
- Haque, M. N., Dahal, N., Hasan, M. K., Manandhar, N. K., & Sharmin, S. (2025, November). Impact of activity types and speaking hours on spoken English fluency: a comparative study in Bangladeshi tertiary classes. In *Frontiers in Education* (Vol. 10, p. 1639602). Frontiers Media SA. <https://doi.org/10.3389/feduc.2025.1639602>
- Ikae, C., Ben Souissi, S., Bieri, J. S., Müller, T. J., Feuz-Schlunegger, M. C., & Golz, C. (2026). A scoping review of natural language processing for detecting work-related stress among health professionals. *Discover Computing*, 29(1), 14. <https://doi.org/10.1007/s10791-025-09886-7>
- Imam, M. H., Pathan, M. A. I., & Farzana, S. (2025). Bridging the Policy–Practice Divide in English Language Teaching and Education: Challenges and Strategic Pathways for Bangladeshi

- Colleges. *ICRRD Journal*, 6(3), 159–174. <https://icrrd.com/public/media/29092025050213-English-Language-Teaching.pdf>
- Islam, M. T., Johab, M. A., Ahmed, M. R., & Aman, N. (2025). Hybrid BiLSTM-BiGRU Model for Classification of Bangladeshi Dialects. In *International Conference on Data Science, AI and Applications* (pp. 1-16). Springer, Cham.
- Kabir, S. A. (2024). Policy and Practice of Listening in Secondary English Education: A Critical Reflection from Bangladesh. *MEXTESOL Journal*, 48(3), n3. https://mextesol.net/journal/index.php?id_article=377439
- Kaur, J., Agarwal, S., Kaur, S., & Mogaji, E. (2026). Namaste Alexa: The Impact of Non-Native Language Queries on Voice Assistant Usage Intentions. *Journal of Consumer Behavior*. <https://doi.org/10.1002/cb.70108>
- Khan, T. F., Arslan, M., Alturise, F., Almutairi, A. H., & Raza, M. T. (2026). UQA-RAG: Enhancing Urdu Question-Answer Retrieval With Semantic Similarity and LLMs. *IEEE Access*, 14, 2868–2882.
- Kim, T. Y. (2026). English Education During the Seventh National Curriculum Period (1997~ 2007). In *Historical Progression of English Education in Korea* (pp. 409–484). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10084-9_11
- Kobayashi, M., Mita, M., & Komachi, M. (2024). Large language models are state-of-the-art evaluators for grammatical error correction. *arXiv preprint arXiv:2403.17540*. <https://doi.org/10.48550/arXiv.2403.17540>
- Leuckert, S., Rüdiger, S., & Bernaisch, T. (Eds.). (2026). *Evolving Englishes: Varieties through Space and Time*. Taylor & Francis.
- Lima, S. A. (2023). *The Role of L1 (First Language) in L2 (Foreign or Second Language) Learning: An Inquiry at the Secondary Level* (Doctoral dissertation, BANGLADESH UNIVERSITY OF BUSINESS AND TECHNOLOGY).
- Mahmud, F., Hayath, T., Rahman, M. S., Kowsar, A. M., Mannan, M., & Siddique, M. M. (2024). The Role of Community Clinics in Shaping Health Outcomes in Rural Bangladesh: A Qualitative Study. *Malaysian Journal of Medical and Biological Research*, 11(1), 21–30. <https://mjmbmr.my/article/view/691>
- Mannan, M., & Maruf, T. I. (2024). Perceived Usefulness and Perceived Ease of Use in the Worth of Online Education System in Bangladesh. In *Second International Conference on Innovations in Management, Science, Technology and Automation in Sports (ICIMSTAS-2023)*. <https://doi.org/10.5281/zenodo.13763573>
- Mannan, M., & Maruf, T. I. (2025). Exploring the Role of AI-Driven Prompt Engineering in Enhancing ESL Pedagogy and Bridging the Digital Divide in Higher Education.
- Mannan, M., Maruf, T. I., & Ali, M. L. (2025). Evaluation of the DDR Method for Developing Technology-Enhanced Learning Frameworks in Higher Education: A Bangladeshi Context. *Citra Journal of Innovative Education Practices*, 3(1), 13–24. <https://citralestari.my/index.php/cjiep>
- Mannan, M., Mustafa, Z. B., Aziz, S. F. B. A., & Maruf, T. I. (2023). Technology adoption for higher education in Bangladesh—development and validation. *Journal of Education and Social Sciences*, 24(1), 1–9.
- Mita, M., Sakaguchi, K., Hagiwara, M., Mizumoto, T., Suzuki, J., & Inui, K. (2024, June). Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA, 2024)* (pp. 251–265). <https://aclanthology.org/2024.bea-1.21/>
- Neuberger, J. (2026). *Extracting Formal Process Model Information from Natural Language Text Descriptions using Machine Learning* (Doctoral dissertation). <https://epub.uni-bayreuth.de/id/eprint/8435/>
- Pell, M. D., Cui, H., Mori, Y., & Jiang, X. (2026). Speak or shout? Nonverbal vocalizations promote rapid detection of emotions in vocal communication. *PLoS One*, 21(1), e0327529. <https://doi.org/10.1371/journal.pone.0327529>

- Rahman, K. A. (2025). *Effects of school-based assessment in achieving the objectives of the intended English language curriculum at grades VI to VIII* (Doctoral dissertation, © University of Dhaka).
- Salifu, A., Mensah, H. N., Tchao, E. T., Ibrahim, A. M., Acheampong, F. A., Kponyo, J. J., & Agbemenu, A. S. (2026). A systematic review of accent classification techniques and datasets for inclusive speech recognition. *International Journal of Data Science and Analytics*, 21(1), 12. <https://doi.org/10.1007/s41060-025-00954-1>
- Sato, M., Thorne, S. L., Michel, M., Alexopoulou, T., & Hellermann, J. (2025). Language, people, classrooms, world: Blending disparate theories for united language education practices. *Modern Language Journal*, 109 (Supplement 2025), 15–38. <https://doi.org/10.1111/modl.12976>
- Shabur, M. A., Jahan, J., & Barua, P. (2025). Assessing the success of Education 4.0 in enhancing English proficiency for engineering students in Bangladesh. *Discover Education*, 4(1), 61. <https://doi.org/10.1007/s44217-025-00453-2>
- Vetrivel, S. C., Arun, V. P., Ambikapathi, R., & Saravanan, T. P. (2025). Automated grading systems: Enhancing efficiency and consistency in student assessments. In *Adopting Artificial Intelligence Tools in Higher Education* (pp. 41–61). CRC Press. <https://doi.org/10.1201/9781003470304-3>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Zhao, Y., Yao, Q., & Zhu, W. (2026). ShapLoRA: Allocation of Low-rank Adaptation on Large Language Models via Shapley Value-Inspired Importance Estimation. *arXiv preprint arXiv:2601.17921*. <https://doi.org/10.48550/arXiv.2601.17921>