P-ISSN 2708-6453	Pakistan Languages and Humanities Review	Jan-Feb 2026, Vol. 10, No. 1
O-ISSN 2708-6461	https://doi.org/10.47205/plhr.2026(10-I)14	[131-1343

Pakistan Languages and Humanities Review
www.plhr.org.pk

RESEARCH PAPER

Assessing Content Validity and Thematic Trends in CSS English Literature Question Papers: A Critical Study

[1] Uroosa Aurangzeb and [2]Dr. Khurram Shahzad

1. Lecturer, Department of English, University of Azad Jammu and Kashmir, Muzaffarabad, AJ&K, Pakistan
2. Assistant Professor, Department of English (GS), National University of Modern Languages, Islamabad, Pakistan

*Corresponding Author	uroosa.aurangzeb@uajk.edu.pk

ABSTRACT

This study examines the effectiveness of English literature question papers of Pakistan's Central Superior Services (CSS) examination from 2019 to 2023. The analysis is limited to literature papers, excluding exams from other disciplines. Designing assessments is crucial to ensuring test validity. Particularly, literature exams depend on the clarity and alignment of tasks. A qualitative research approach is utilized to evaluate the effectiveness of tests. The study is based on the Test Usefulness Model of Bachman and Palmer's (1996) and Shohamy's (2020) critical viewpoint on language testing, which focuses on reliability, validity, authenticity, and syllabus alignment. The study uses 50 exam papers, 10 from each year, obtained from the CSS website. The results indicate issues such as unclear instructions, inconsistent tasks, uneven difficulty levels, and misalignment of the syllabus. Language testing is an underexplored area of the CSS exam. Resultantly, clearer tasks and regular syllabus updates are required to improve exam design.

KEYWORDS CSS Examination, English Literature, Test Usefulness, Construct Validity, Assessment Design

Introduction

Assessment and evaluation in high-stakes tests are not merely measurement tools for assessing skills; they also play a significant role in designing the curriculum. In the domain of language and literature testing, the design of question paper plays a vital role, for it helps assess the literary and communicative competencies of the candidates. Instead of merely testing candidates' knowledge, the tests evaluate their knowledge and skills. In the context of high-stakes testing practices, the evaluation process informs teaching methods, preparation strategies for candidates, and the assignment of intellectual classifications to certain disciplines. This influence is notably visible in the CSS English literature examinations, where the design of the test guides candidates whether they should formulate the responses with memorized knowledge and just reproduce it, or they should critically analyze the literary text with deep textual interpretation. In this context, the quality of these tests is crucial because they bear on transparency, fairness, score interpretation, and the broader academic principles of literary and linguistic skills. Over time, the testing standard evolved from a focus on grammatical knowledge or linguistic skills to a focus on discourse competence and communicative linguistic competence. Language tests are very important for examining the structure of tests, the academic competence of test takers, and the effectiveness of assessment procedures.

The connection between what candidates learn and how they are evaluated is assessed through these tests. (Bachman & Palmer, 1996; Brown, 2010). Over time, test design has changed a great deal. It has changed from just evaluating grammar knowledge to assessing pragmatic competence, discourse, and purposive communication in different language settings (Weir, 2005; Canale & Swain, 1980; Hymes, 1972). Many tests require more detailed and longer answers from the candidates. In literature exams, higher-order skills such as literary competence, critical competence, cohesion and coherence, and interpretive skills are assessed rather than just surface-level skills such as basic vocabulary and grammar. In such tests, textual knowledge, cognitive competence, argument development, literary skills, and different perspectives on literature are assessed along with language skills. Research on academic writing highlights that a variety of language skills, such as planning and organizing ideas that help to turn thoughts into well-structured and clear writing, are evaluated through writing tests. (Cumming, 2001; Hyland, 2019; Kellogg & Whiteford, 2009). Researchers have pointed out that many problems, such as a disorganized structure and unclear test instructions, can reduce the fairness, transparency, and accuracy of the test. (Leki et al., 2008; Elander et al., 2006). Scholars agree that higher-level skills are assessed through detailed answers, but the practical implications depend on the clear design and structure of tests.

Literature Review

In Pakistan, the CSS exam is a way to get government jobs. It is very important that it is designed properly, has validity, and transparency. The English literature paper is designed to assess test takers' ability to analyze texts and understand literature. This problem has been backed by recent studies that point out problems such as unclear tasks, ambiguous instructions, and inconsistent cognitive demands, which lead test takers to memorize rather than analytical engagement with the material. (Shahzad et al., 2019). According to Hamp-Lyons (1991) written test should focus more on the candidate's cognitive skills rather than language use. Weigle (2002) also argues that unclear test design and lack of explicit scoring decrease scoring reliability. Shohamy (2020) highlights that the ideological perspectives in high-stakes selection tests reflect the test designers' ideological positions. This subjectivity in test design affects the choice of topics and task framing, and it can make it hard for test takers to assess. Overall, these issues in the evaluative design weaken the significance of testing and its intended construct. Although several studies address issues related to writing and their validity, little attention has been paid to the systematic evaluation of literature question papers, especially CSS. Most studies examined language competence and skills related to general essay writing; this specific area of descriptive response evaluation of literary competence remained underexplored.

Since research on this particular area remains limited, the current study evaluates CSS English literature question papers of the years 2019 to 2023 employing Bachman and Palmer's (1996) Model of Test Usefulness, which consists of six qualities of test usefulness, such as reliability, construct validity, authenticity, interactivity, practicality, and impact. The study also uses Shohamy's (2020) extension of the test usefulness model, which considers the power, fairness, consequences, and social impacts of tests. The CSS testing organization is responsible for conducting high-stakes tests for civil service selection. However, the test design of English literature question papers raises serious concerns regarding reliability, construct validity, and authenticity. This exam is highly important, but struggles to evaluate skills such as literary competence, discourse analysis, and textual analysis. Repeatedly occurring significant concerns, such as unrelated topics not covered in the course, repetition of specific question types, ambiguous instructions, and test items that do not match the syllabus, lead to an unreliable exam, which is subsequently harder to assess. Using Davidson and Lynch's (2008) model, Shahzad (2017) evaluated essay

exams of undergraduate students, but that study did not evaluate the CSS English literature exam questions. Moreover, that study did not use the framework of Bachman and Pakmer's (1996) test usefulness model. The discussion clearly displays the significance of CSS examination; however, despite their importance, the research on systematic English literature exams remains limited. Most contemporary studies focus on assessing general analytical skills based on grammatical competence rather than candidates' knowledge of construct representation, literary competence, and content validity. The present study fills that gap, which has diminished both the theory and practical reform in high-stakes literary assessments. By analyzing CSS English literature assessment papers, the research covers areas such as exam instructions, alignment of syllabus with exam, and overall test design.

Material and Methods

The study uses a qualitative research method to analyze the CSS English Literature exam papers, which are considered as an effective high-stakes assessment tool. The study focuses on examining the clarity of test, construct validity, structure of task, and alignment of questions with syllabus. The study does not focus on examining the literary knowledge and performance of the candidate in text. The exam papers are evaluated using standard language evaluation practices. To measure critical thinking, literary analysis, and the ability to interpret text, the study uses standard principles of language assessment.

Data Collection Procedure

The data for this study consist of CSS English literature question papers from the years 2019 to 2023, collected from the official website of the Federal Public Service Commission (FPSC). Ensuring the data is original and authentic, the papers were analyzed in their original form. Having collected them, they were reviewed to ensure that they matched the correct year and subject. The study primarily focused on Part II of the question papers -- the descriptive section. Part II was meant to analyze higher-level literary and analytical skills, while Part I consists of objective-type questions. 50 question papers, 10 questions from each year, were selected to make a dataset for deeper analysis. In addition to the questions, the instructions given in each paper are also included in the dataset, providing both the general and specific details for understanding how candidates should respond to the questions. These question papers helped to identify the recurring patterns in syllabus, structure of questions, and test instructions.

Method of Analysis

Textual and thematic analysis are used to evaluate the question papers. In textual analysis, issues such as unclear tasks, design of questions, given instructions, directive verbs, structure of sentences, and reductive language are analyzed. For recurring issues such as unclear formation of questions, alignment discrepancies between the test and the syllabus, and uneven cognitive levels across test questions from the past five years, thematic analysis is employed. The analysis is done step by step. In the first step, questions are examined with instructions, question design, and task presentation. Secondly, the questions are compared with the CSS syllabus to check the content, alignment, theoretical components, and different writing genres. In the final stage, test patterns are examined to check consistency and effectiveness in design over the years.

Analytical Consistency

For all 50 questions, no matter when they were designed, the same analysis tools are used, which shows that the analysis is reliable. The examples used to assess papers are

exactly as they are in original papers, and each example has the correct year listed to show where it came from. Same standards for every question make the comparison fair and free of personal opinions.

Ethical Considerations

The data involved in this study were obtained from the official website of the Federal Public Service Commission (FPSC). No human participants, no personal data, and no institutional data are involved in this study. As no human participation, personal data, or confidential record is involved in this study, ethical approval was not needed for this study.

Theoretical Framework

Bachman and Palmer's (1996) model for assessing the usefulness of tests serves as the framework for this study. This framework consists of six main qualities: reliability, validity, authenticity, interactiveness, impact, and practicality. Reliability refers to the consistency in the interpretation of test scores. Construct validity refers to the degree of assessing the specific skills it is intended to assess. Authenticity means how well the test tasks match language use in real life, especially in areas such as analyzing literature. Interactiveness refers to the degree of engagement of test takers with the tasks, using their strategic knowledge, cognitive processes, and their language skills. Impact considers how testing practices affect the education system and broader society, including how candidates learn, how transparent the testing is, and how well they prepare. Practicality means whether the test can be carried out authentically, given the limitations of an institution. This model is mainly used to assess the overall test design, instructions, and alignment of the CSS English literature question papers with the syllabus. Shohamy's (2020) critical viewpoint on language testing is also used to review data. This approach focuses on the social, political, ideological, and ethical effects of high-stakes exams, which makes it different from the traditional testing methods. The data analysis is based on both viewpoints, focusing on the technical aspect as well as looking on social impacts of testing.

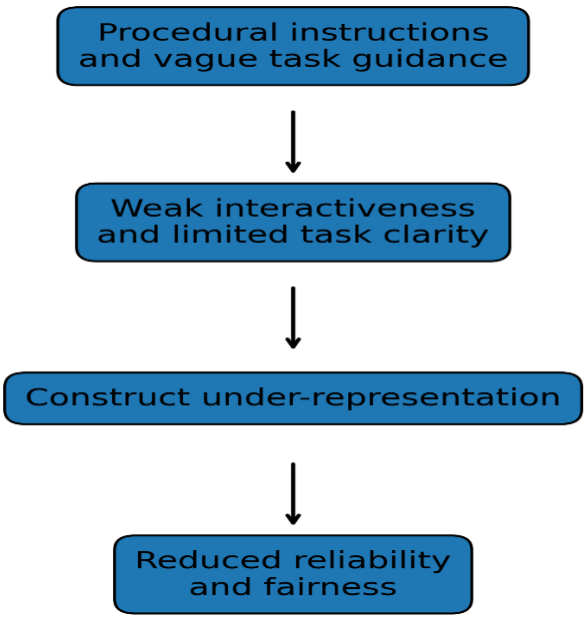

Figure 1. The relationship between test usefulness model and test.

Results and Discussion

Data Analysis of the CSS English Literature Question Paper

The CSS English literature question papers are analyzed from three perspectives to assess their effectiveness as a high-stakes assessment tool. First, the test instructions are deeply examined to see if they are clear, help candidates in understanding questions, and are aligned with international standards of testing. Second, the test content is assessed to make sure the questions match the syllabus that is prescribed. Third, the whole design of the test is assessed for its validity, reliability, impact, and practicality. Using Bachman and Palmer's (1996) Test Usefulness Model and Shohamy's (2020) critical view on language testing, the CSS English literature question papers are analyzed from all these perspectives.

Analysis of Instructions and Test-Taker Guidance

Procedural Orientation and Lack of Pedagogic Support

Although the instructions given in the CSS English literature question papers are on the proper format, they do not completely match with Bachman and Palmer's model of test usefulness, in terms of transparency, support for test-takers, authenticity, and clarity. On close evaluation of test, especially the instructions, there are some repeated issues found that affect how well the exam can measure the ability of a candidate to communicate. The first instruction on the question paper, Part II, is to be attempted on the separate Answer Book (*Question Paper attached in Appendix B*), which is part of the procedure. Such instructions reduce the authenticity of the test because they have administrative goal of addressing the task-related problems of performance or cognition. The key idea of the test structure is to give guidance to candidates about what is being tested from them and how their abilities are being assessed. In this format, the instructions do not give educational help; rather, they act more like rules made by an authority.

Limited Strategic Guidance and Reduced Interactiveness

The second instruction*:* Attempt ONLY FOUR questions from PART-II. ALL questions carry EQUAL marks. These instructions focus more on rule enforcement rather than providing guidance to candidates on decision-making. There is no clear information for candidates about what exactly is needed for each question, whether they require comparative analysis, textual arguments, or theoretical engagement. Due to unclear instructions, it becomes harder for candidates to choose the question that best suits their strengths, which affects their performance. It not only affects the performance of candidates but also weakens test interactivity. The instructions are also ambiguous: *All the parts (if any) of each Question must be attempted at one place instead of at different places*. The conditional phrase *if any* makes the internal structure of questions unspecified, creating an ambiguity about whether a question is divided into sections. The ambiguities make it difficult for candidates in time management, understanding of task and organizing answers.

Control-Oriented Language and Negative Washback

The analysis of test instructions exhibits a more control-oriented approach rather than one that supports candidates. The directive *Write Q. No. in the Answer Book in accordance with Q. No. in the Q. Paper* depicts rules without considering understanding of task, cognitive engagement, or comprehension. Similarly, instructions such as *No Page/Space be left blank between the answers and All the blank pages of the Answer Book must be crossed out* impose rules without illuminating their relevance to the organization of the answers. Such instructions create the feeling of anxiety among the candidates, especially

those who are not aware of the rules of the exam. They also decrease the positive washback effects that exams can have on learning and educational results. The final instruction: *an extra attempt at any question or part of the question will not be considered*, has a tone that seems to be harsher than supportive. It displays that revising answers is not allowed, making it difficult for candidates to adjust their responses. This rule is tough for candidates as they cannot change the question they chose, even if they believe they could write a better answer, demonstrating better literary competence. Instead of helping candidates to write what they know, these unclear, strict, and unsupportive instructions reduce their ability to show their actual linguistic skills.

Instructional Stagnation and Test Validity

The test instructions are structured in a highly strict, administrative-focused way, highlighting control, transparency, efficiency, and adherence to rules, rather than being more supportive of candidates. More focus on administrative rules ignores the aspect of clarity, how responses were evaluated, and the cognitive level. As Shohamy (2020) points out, the use of authority during testing can affect how well candidates perform in language skills. According to Bachman and Palmer's test usefulness model, effective test instructions should help candidates to comprehend and support them in showing their abilities meaningfully. This instructional format has remained unchanged for 5 consecutive years. This consistency in instructions indicates the testing agency's least concern with keeping up with new teaching approaches, test takers' needs, and recent research on evaluation. Although there is a global shift toward testing standards that emphasize test takers' communicative skills, as outlined in Bachman and Palmer's framework on task clarity, learner-focused instruction, and communicative purpose, the CSS examination is still based on outdated, static instructional patterns. This stagnation weakens institutional inertia, educational impact, and undermines construct validity in a high-stakes testing setting.

Table 1
Coding Framework for the Analysis of CSS English Literature Question Papers

Test Usefulness Dimension	Focus of Analysis	Coding Indicators
Construct Validity	Alignment with syllabus and intended literary competence	Prescribed texts, genre relevance, avoidance of construct-irrelevant knowledge
Authenticity	Relevance to literary analysis	Text-based interpretation, engagement with literary form and meaning
Interactiveness	Cognitive engagement required	Analytical depth, synthesis, interpretive reasoning
Reliability	Clarity and consistency of task wording	Defined concepts, absence of vague or binary framing
Impact	Educational consequences	Encouragement of critical study vs memorization
Practicality	Feasibility under exam conditions	Clear expectations, manageable scope

Syllabus Alignment and Content Representation

General Alignment with Persistent Construct Distortion

The CSS English literature question papers are critically analyzed and compared with the official syllabus (see Appendix A), employing the Common European Framework of Reference (CEFR) and Bachman and Palmer's Test Usefulness Model. It helped to present an overall effort, maintaining engagement with the themes of literary competence. Yet, the number of test items aligns only partially with the prescribed syllabus, influencing content relevance, reliability, and construct validity. The CEFR focuses on task clarity,

authentic language use, and goal orientation. Bachman and Palmer emphasize the reliability, construct validity, authenticity, and impact dimensions that are often not well addressed in the papers reviewed.

Disproportionate Text Emphasis and Syllabus Imbalance

One recurring issue concerns the framing and frequency of questions about Emerson's *Self-Reliance* in CSS English literature exams. Even though the text is part of the syllabus, the questions often treat it as the main focus instead of one of the prescribed texts. For example, Q.2 (2021): "What does Emerson mean by Whoso would be a man must be a non-conformist? Discuss" and Q.2 (2022): "Explain how both Bertrand Russell's The Conquest of Happiness and Ralph Waldo Emerson's 'Self-Reliance' focus on how the individual must develop and rely on his or her moral judgement", directing candidates' attention disproportionately toward Emerson. Although *Self-Reliance* is listed under the essay section of the syllabus, it is being emphasized more than it should. Over several exam years, three out of five papers give more weight to questions about Emerson, which throws off the balance of the syllabus. This creates a pattern that favours test takers who expect repetition rather than those who study all the required texts thoroughly. Such repetition weakens the representativeness of the content and supports strategic memorization rather than broad literary competence.

Scaffolded Comparative Demands and Content Validity

The concern regarding syllabus misalignment becomes more prominent in scaffolded comparative questions. In Q.2 (2022), candidates were asked to compare Emerson's *Self-Reliance* with Russell's *The Conquest of Happiness*, even though the syllabus lists these texts separately and does not prescribe thematic pairing. Emerson, a 19th-century transcendentalist writer, and Russell, a 20th-century analytical philosopher, come from very different intellectual backgrounds. Asking candidates to compare their views on moral independence imposes a cognitive demand that is not clearly supported by the syllabus. The content validity is weakened because the question formulations demand analytical connections that are not taught to examinees specifically. Although the questions seem smart and intellectually stimulating, they ask for interpretations that go beyond what was covered in the course. This makes the answers vary in how well they are based on the text rather than on organized literary analysis.

Abstract Moralization and Construct-Irrelevant Variance

A similar issue of construct distortion appears in Q.7 (2022): "Contrast the concept of self-love in Somerset Maugham's The Lotus Eaters with the concept of love for others as reflected in Iris Murdoch's Under the Net". While both texts are mentioned in the syllabus, the question framing changes the focus from literary analysis to abstract moral philosophy. The stress on ethical concepts such as *self-love* and *love for others* tends to encourage general philosophical discussion rather than engagement with literary form, narrative structure, or characterization. As defined by Bachman and Palmer, such abstraction introduces construct-irrelevant variance and weakens the authenticity of the task. Misalignment is further intensified when external critics are involved. In Q.7 (2023): "It is we, says Hazlitt, who are Hamlet. Illustrate with textual examples, Hamlet as a universal character", the mention of Hazlitt introduces an external critic not listed in the prescribed syllabus. Test-takers who rely heavily on syllabus-based preparation are placed at a disadvantage, while others with access to supplementary critical material gain an unintended advantage. According to the CEFR view, these tasks reduce the test's transparency by violating the rules of clarity and accessibility.

Interim Synthesis of Syllabus-Related Issues

While the test questions demonstrate surface-level engagement with prescribed texts, they also demonstrate inconsistencies in reliance on external criticism, uneven weighting of content, comparative demands, and compromises in construct validity, authenticity, and fairness.

Table 2
Recurring Design Issues in CSS English Literature Papers (2019–2023)
Theory-Driven Prompts and Construct Drift

Identified Issue	Years Observed	Representative Examples
Binary question framing (Is/Does)	2019, 2020, 2023	Q.2 (2019); Q.7 (2020); Q.2 (2023)
Syllabus misalignment	2020–2023	Hazlitt (Q.7, 2023); Pinter (Q.5, 2020)
Scaffolded comparison	2022	Russell–Emerson (Q.2, 2022); Maugham–Murdoch (Q.7, 2022)
External critical knowledge	2022–2023	Hazlitt (2023); Classical sources in Ulysses (Q.8, 2022)
Over-general theory prompts	2020, 2022	Post-colonial Studies (Q.7, 2020); Marxism (Q.6, 2022)

A similar problem arises in Q.6 (2022): "Explain how the theories of Karl Marx are still relevant to literature today". Although Marxism is a part of the syllabus under *Literary Theory and Criticism*, the way the question is phrased encourages a generalized theoretical essay rather than text-based literary analysis. Since there is no instruction to apply Marxist ideas to specific texts, genres, or literary forms, the task's construct validity is weakened. Candidates are likely to write about broad socio-political issues rather than demonstrate how Marxist theory helps literary interpretation. As a result, the task fails to test the precise analytical competence it claims to evaluate. A comparable shift away from text-based analysis appears in Q.8 (2022): "How much of a literary debt does Tennyson's Ulysses owe to classical Greek and Roman tradition?" *Comment in detail*. While *Ulysses* is part of the syllabus, the question focuses on how much it is influenced by classical epics, requiring comprehensive knowledge of Homeric and Virgilian traditions. Such type of content is neither mentioned in the syllabus nor specified within the expected literary history section. This adds construct-irrelevant knowledge and introduces an unnecessary layer of intertextual complexity, disadvantaging otherwise competent candidates. The partial alignment is shown in Q.5 (2020) on Harold Pinter's *The Caretaker*. Although Pinter is a central figure in modern British drama, he is not mentioned explicitly in the prescribed syllabus. However, major themes such as post-war drama or absurdism justify his inclusion, but the question's focus on Mick's language shifts the task toward linguistic analysis rather than literary criticism. The inclusion of questions not explicitly listed in the syllabus shows examiners' discretion in selecting questions, which influences content selection. Misalignment in tasks is further shown in Q.7 (2020): "Is Post-colonial Studies the most flourishing sector of cultural studies today? Justify your arguments with the help of appropriate examples". Even though post-colonial criticism is listed in the syllabus, the question invites sociological commentary rather than literary analysis. The phrase *flourishing sector of cultural studies* shifts the focus from textual interpretation to disciplinary status. Candidates may discuss academic trends, university curricula, or historical development without engaging with post-colonial literary texts. Bachman and Palmer argue that such deviations lead to under-representation of the intended construct, increase construct-irrelevant variation, and reduce score interpretability.

Question Framing and Construct Clarity

A recurring structural weakness in the dataset is the frequent use of binary interrogative framing, especially questions beginning with *Is* or *Does*. For example, Q.2

(2023):"Is happiness possible in the modern world? Explain in the light of Russell's The Conquest of Happiness" and Q.5 (2023): "Is there a significant relation between language and political thought? Explain the rules of writing good prose in light of Orwell's Politics and the English Language". While the questions are grammatically correct, such formulations are pedagogically reductive. The task's interactivity is weakened because the binary framing of questions encourages yes/no responses rather than an analytical focus and interpretive depth. The same issue is present in Q.4 (2020): "Does Eliot's own poetry also depict the same quality in poems The Waste Land and The Love Song of J. Alfred Prufrock?" Terms such as *same quality* and *depict* are not defined explicitly, leaving test takers uncertain about the required analytical construct. Without clear direction, responses vary widely in focus, scope, and emphasis, leading to inconsistent grading.

Lexical Ambiguity and Under-Specified Tasks

Several questions exhibit lexical or conceptual vagueness, weakening construct clarity. In Q.5 (2020): "Discuss Mick's exploitation of language for accosting Davies in Harold Pinter's play The Caretaker", the phrase *exploitation of language* is ambiguous, while *accosting* sounds semantically forceful yet contextually ambiguous. Test takers interpret the task through linguistic, psychological, or thematic lenses, leading to responses that are hard to assess in a standard way. Likewise, Q.6 (2021): "Eugene O'Neill's Long Day's Journey into Night is a modern tragedy. Explain", does not clearly specify what constitutes *modern*, whether it refers to dramatic structure, form, theme, or psychology. This lack of clarity leads to diffused, unfocused answers, violating the principles of construct validity because the question does not clearly guide a specific form of literary engagement.

Unscaffolded Comparative Demands

Unclear scaffolding also characterizes some comparative prompts. In Q.2 (2022): "Explain how both Bertrand Russell's The Conquest of Happiness and Ralph Waldo Emerson's Self-Reliance focus on how the individual must develop and rely on his or her moral judgement", and assume an analytical linkage between the two texts originating from different intellectual backgrounds. Without specific guidance on comparative parameters, candidates struggle to form a coherent analytical framework. A similar issue arises in Q.7 (2022): "Contrast the concept of self-love in Somerset Maugham's The Lotus Eaters with the concept of love for others as reflected in Iris Murdoch's Under the Net". The imposed philosophical contrast creates an artificial opposition rather than encouraging text-based literary analysis, thereby increasing cognitive load without improving construct representation.

Pragmatic Weakness and Construct Drift

Some questions seem conceptually open but lack pragmatic clarity. In Q.8 (2020): "How far do you think psychoanalysis is appropriate for understanding literary texts"? The phrase *how far* suggests evaluative scaling without defining criteria. Test-takers may invoke trauma theory, Jung, Freud, or Lacan theory, leading to a range of interpretations that undermine scoring reliability. Similarly, Q.6 (2022) on Marxism elicits responses that range across literary criticism, social theory, ideology, and again weaken construct control.

Better-Constructed Exceptions

Despite these issues, some questions are well-structured and align more closely with literary assessment principles. Q.3 (2020): "How successfully do you think Chesterton manages to employ the technique of narrative within a narrative in his short story A

Somewhat Improbable Story" specifies an evaluative task, a critical concept, and a text. Likewise, Q.4 (2022): "What major tragic aspects of humanity are incorporated into Thomas Hardy's Far from the Madding Crowd and D.H. Lawrence's Sons and Lovers", and how do their frames set up a comparison around a clearly defined thematic focus? The above questions are aligned with Bachman and Palmer's model of test usefulness.

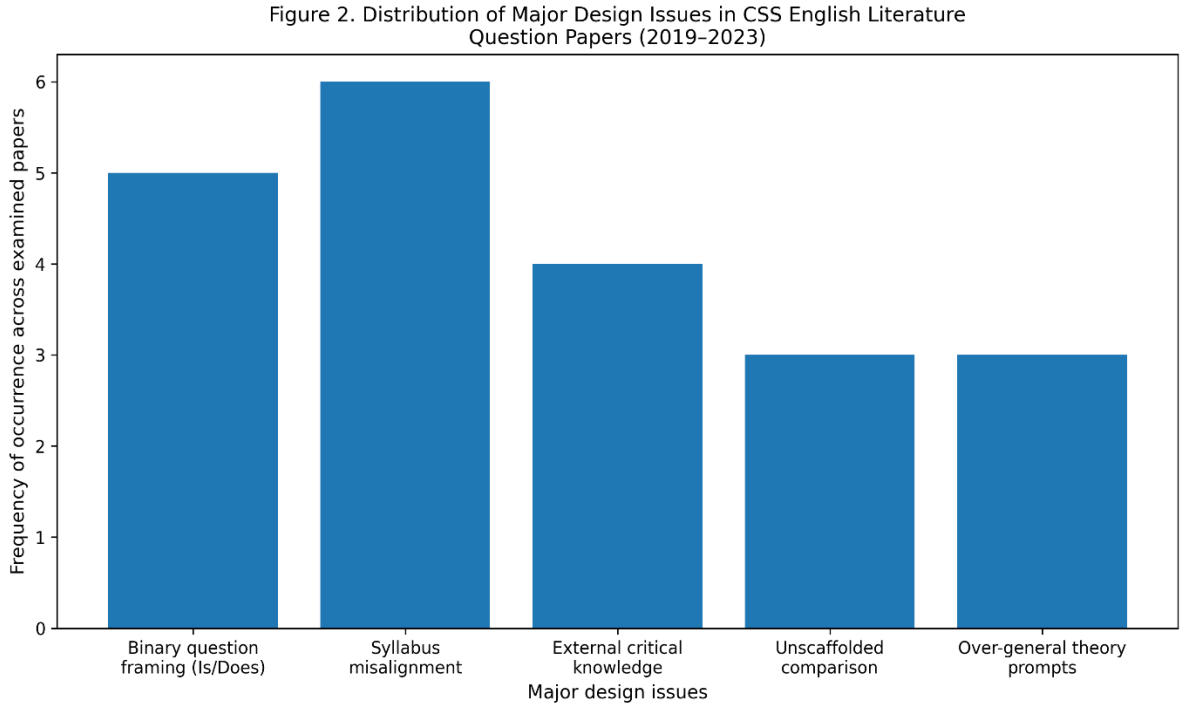

Figure 2. Distribution of Major Design Issues in CSS English Literature Question Papers (2019–2023)

Interim Synthesis

Evaluating the CSS English literature exam papers closely, there are clear problems with how the questions are designed. The tasks are not always clear, and they do not always align with the syllabus. There are also issues with the extent to which the questions measure what they are supposed to, and some are poorly constructed. Not all the prompts display these issues because some of them are well-focused and demonstrate a literary competence, some others present issues related to ambiguous comparisons, simple language, unclear tasks, and a demand for knowledge that is not part of the syllabus. According to Bachman and Palmer's Test Usefulness Model, effective test items should be relevant to the course and easy to understand. The problems occur across all 5 years, diminishing the tests' reliability and validity.

Discussion

The CSS English literature question papers are examined under the lens of Bachman and Palmer's (1996) test usefulness model and from the viewpoint of Shohamy (2020) on language testing. These language testing models focus on areas such as syllabus alignment, test design, task formulation, construct validity, and test instructions. The presence of such issues weakens the overall effectiveness and quality of test design as the validity and transparency are crucial for such high-stakes testing organizations. The first research question is addressed under the test usefulness qualities such as reliability, validity, and authenticity. The evaluation of 5 consecutive years of papers shows recurring issues related

to unclear test tasks that are designed in a way difficult to understand for the literary analysis. Further, the tasks push candidates to rely more on interpretation than on actual literary competence. This ambiguity of task design forces candidates to rely more on personal interpretation, leading them to produce their tasks based on memorization instead of actual literary competence. According to Weigle's (2002) standpoint on language assessment, the lack of clarity in task description results in score variation that is not constructed according to the intended test skills. The results support other researchers who maintain that clear test instructions have a strong impact on how candidates respond. Hamp-Lyons (1991) argues that effective written assessments are not only about how the tasks are designed, but also how clearly expectations are explained. The instructions provided by the test organization are in detail, but they focus more on the procedures and administrative control. The general and specific instructions fail to help candidates in understanding the difficult tasks that are important in the exam. As discussed by Bachman and Palmer (1996), having issues with task design and proper guidance lead to negative washback.

The second question investigated the issues related to ambiguous task design, syllabus misalignment, and irregular content weighting influences the reliability, validity, and fairness of test questions. The test tasks show surface-level understanding of the text and include the contents that repeatedly appear in the test every year. Shohamy (2020) presents the idea of ideological biasness of high-stakes testing organizations that are relevant in this research, as most test items are ideologically biased because of misalignment with the syllabus. The design of topics selected for the test disadvantages those candidates who prepare for the test only from the syllabus and favors those candidates who have access to the broader critical discourse. The structure of tests like this makes it difficult to interpret scores against the principles of CEFR-based evaluation (Lillis, 2002; Inoue & Poe, 2012). Shohamy (2020) states that the design of these testing organizations indicates how the tests are used as tools of power with such strict instructional tones. It has advantages for one class of candidates and disadvantages for others who just depend on the syllabus. In the descriptive test design, the use of question format with yes/no binary forms, as discussed by Davidson and Lynch (2002), questions the validity of the test construct. The test is supposed to assess the skills that require descriptive responses, not just these binary constructions. The evaluation of a literary task should include control over the discussion, comprehension, and sensitivity to context.

Test tasks designed with such a yes/no items undermine the construct validity of the test. Hyland (2019) supports this argument that descriptive response-based literary tasks add a positive washback on the exam when they are designed properly with clear guidance on relevance through instructions that are task-specific. The analysis of the question papers with multiple perspectives shows that the problems highlighted in the study are not random to be ignored. The test organization is required to consider these issues to improve the effectiveness of the test. These issues weaken the reliability, validity, authenticity, practicality, and impact of the tests. The incorporation of clear test instructions, syllabus alignment, and alignment with modern assessment standards improve the efficacy of tests. The common problem of testing memorization instead of real literary skills also impacts the positive backlash on the education system.

Conclusion

Overall, the evaluation of CSS literature papers exhibits partial alignment with the testing standards. The remaining parts of questions papers depict the issues across multiple areas that are important to consider. The overall design of the question paper is deficient in areas such as reliability, construct validity, and authenticity as presented by Bachman

and Palmer's model of test usefulness. Designing a question paper according to the modern testing standards is crucial for the testing organization. Furthermore, the question paper is evaluated from the standpoint of Shohamy's (2020) critical view of language testing. Under the guidelines of his model, the issues highlighted related to power, transparency, and social impact on the candidate. The comparison of the test questions with standard testing practices reveals issues related to higher-order cognitive skills, lack of scoring rubrics, misalignment with syllabus, language accuracy issues, uneven selection of themes, and textual analysis. The reliability and construct validity of the test are compromised by these recurring issues.

Recommendations

The test design, structure, and execution need to be changed in order to fix these issues. The syllabus has not changed for a long time, so it should be updated regularly to include up-to-date topics and current trends. Resolving these issues makes the test more valid and authentic, thereby justifying the selection of candidates on the basis of communicative competence and literary skills. The overall design and structure of tests are required to be changed to meet the updated standards of testing practised around the globe. The alignment of test design with the frameworks of language testing, such as Bachman and Palmer's test usefulness model and the CEFR, could make the tests more valid, authentic, and transparent. The valid design of the test is mostly based on the academic competence of the examiners, so training examiners about the changing trends of testing could be fruitful for a valid and reliable test design. Along with this, the regular feedback on testing practices would result in continuous improvement, adaptability, and new changes in the literature and testing practices.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, H. D. (2010). *Language assessment: Principles and classroom practices*. Pearson Education.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies, 1*(2), 1–23.

Davidson, F., & Lynch, B. K. (2008). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Elander, J., Harrington, K., Norton, L., Robinson, H., & Reddy, P. (2006). Complex skills and academic writing: A review of evidence about the types of learning required to meet core assessment criteria. *Assessment & Evaluation in Higher Education, 31*(1), 71–90.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Ablex Publishing Corporation.

Hyland, K. (2019). *Second language writing*. Cambridge University Press.

Hyland, K., & Jiang, F. K. (2019). *Academic discourse and global publishing: Disciplinary persuasion in changing times*. Routledge.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin.

Inoue, A. B., & Poe, M. (2012). *Race and writing assessment*. Peter Lang.

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist, 44*(4), 250–266.

Leki, I., Cumming, A., & Silva, T. (2010). *A synthesis of research on second language writing in English*. Routledge.

Lillis, T. M. (2002). *Student writing: Access, regulation, desire*. Routledge.

Shahzad, K., Janjua, F., & Asghar, J. (2019). Assessing testing practices with reference to communicative competence in essay writing at undergraduate level in Pakistan. *Journal of Research in Social Sciences, 7*(1), 1–18.

Shohamy, E. (2020). *The power of tests: A critical perspective on the uses of language tests*. Routledge.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation*. Palgrave Macmillan.