Półrocznik Językoznawczy Tertium. Tertium Linguistic Journal 10 (1) (2025)
www.journal.tertium.edu.pl
https://doi.org/10.7592/Tertium.2024.10.1.322

Beyond the Bubble: Reconsidering the Role of Multiple-Choice Questions in Language Assessment

Louis Train
University of Calgary, Canada
louistrain@gmail.com

Abstract

This paper examines the role of multiple-choice questions (MCQs) in language assessment, combining a review of existing research with a focused case study from Uzbekistan. From a pedagogical and psychometric perspective, the paper explores both the affordances and limitations of MCQs as tools for assessing language knowledge, receptive skills, and aspects of discourse and pragmatics. Drawing on sources such as Phakiti and Leung (2024), Hughes and Hughes (2020), and Bachman and Palmer (2010), the analysis highlights the standardisation, efficiency, and diagnostic precision that MCQs can offer, particularly in large-scale and high-stakes testing environments. At the same time, it addresses critical concerns, including the inability of MCQs to assess productive language skills, their tendency to fragment language knowledge, and their potential to distort teaching and learning. The paper also evaluates alternative assessment formats (e.g., cloze tests, constructed-response tasks, portfolios) and outlines principles for writing effective MCQs. The second part of the paper presents a document-based case study of the Pedagogical Excellence and International Assessment Centre in Uzbekistan and its use of MCQs in national summative assessments. The research material includes government resolutions and sample items from the end-of-quarter summative exams (ChSB), supported by critical commentary from local media sources. Analysis of this material suggests that MCQs have been instrumental in supporting the country's transition to criterion-referenced, CEFR-aligned assessment, but that tensions may remain between goals of standardisation and concerns about test validity, teacher autonomy, and pedagogical washback. The paper concludes that MCQs are neither inherently flawed nor universally appropriate: their effectiveness depends on alignment with construct definitions, assessment purposes, and broader educational values.

Keywords: language assessment, multiple-choice questions (MCQs), test validity, Uzbekistan, education reform.

Streszczenie

Poza bańką: ponowna próba rozważenia roli pytań wielokrotnego wyboru w ocenianiu umiejętności językowych

Artykuł analizuje rolę pytań wielokrotnego wyboru (MCQ) w ocenianiu umiejętności językowych, łącząc przegląd istniejących badań z ukierunkowanym studium przypadku z Uzbekistanu. Z perspektywy pedagogicznej i psychometrycznej, artykuł bada zarówno możliwości, jak i ograniczenia MCQ jako narzędzi do oceniania wiedzy językowej, umiejętności receptywnych oraz aspektów dyskursu i pragmatyki. Odwołując się do prac takich autorów, jak Phakiti i Leung (2024), Hughes i Hughes (2020) oraz Bachman i Palmer (2010), analiza podkreśla standaryzację, efektywność i diagnostyczną precyzję, jakie mogą zapewnić MCQ, zwłaszcza w kontekstach oceniania na dużą skalę i w testach wysokiej stawki. Jednocześnie artykuł podejmuje krytyczne kwestie, takie jak brak możliwości oceniania produktywnych umiejętności językowych, tendencja MCQ do fragmentaryzacji wiedzy językowej oraz ich potencjał do zniekształcania procesów nauczania i uczenia się. Artykuł ocenia również alternatywne formaty oceniania (np. testy luk, zadania otwartej odpowiedzi, portfolio) oraz przedstawia zasady tworzenia skutecznych pytań wielokrotnego wyboru. Druga część artykułu prezentuje studium przypadku oparte na analizie dokumentów z Pedagogicznego Centrum Doskonałości i Międzynarodowej Oceny w Uzbekistanie, dotyczące wykorzystania MCQ w krajowych ocenach podsumowujących. Materiał badawczy obejmuje rządowe rozporządzenia oraz przykładowe zadania z egzaminów podsumowujących na koniec kwartału (ChSB), uzupełnione o krytyczne komentarze z lokalnych źródeł medialnych. Analiza tych materiałów wskazuje, że MCQ odegrały kluczową rolę w wspieraniu przejścia kraju na ocenianie oparte na kryteriach, zgodne z poziomami Europejskiego Systemu Opisu Kształcenia Językowego (CEFR), jednak wciąż mogą występować napięcia między dążeniem do standaryzacji a obawami dotyczącymi ważności testów, autonomii nauczycieli i efektu zwrotnego w procesie nauczania. Artykuł konkluduje, że pytania wielokrotnego wyboru nie są ani z natury wadliwe, ani uniwersalnie odpowiednie — ich skuteczność zależy od zgodności z definicjami, celami oceniania oraz szerszymi wartościami edukacyjnymi.

Słowa kluczowe: ocenianie umiejętności językowych, pytania wielokrotnego wyboru (MCQ), ważność testu, Uzbekistan, reforma edukacyjna.

1. Introduction

Multiple-choice questions (MCQs) are among the most widely used item types in language assessment. From large-scale proficiency tests like TOEIC and IELTS to classroom quizzes and textbook activities, MCQs have come to define what assessment looks like in many educational contexts. Yet their prevalence has not secured their reputation. For decades, language educators and assessment specialists have debated their pedagogical value, psychometric reliability, and overall validity. Critics argue that MCQs are unable to capture authentic language use, while defenders point to their scalability, objectivity, and diagnostic precision.

This paper aims to examine MCQs not only in terms of their format-specific strengths and weaknesses, but also in relation to broader educational contexts and assessment goals. It combines a critical review of research on the design, application, and critique of MCQs in

language assessment with a focused case study of their institutional use in Uzbekistan. The objectives are threefold: to map the main arguments in the literature concerning the validity and limitations of MCQs; to evaluate how MCQs compare with alternative formats in practice and principle; and to analyse a real-world implementation of MCQs as part of an ongoing educational reform effort. Through this approach, the paper seeks to provide a more nuanced understanding of when and how MCQs can be pedagogically and psychometrically appropriate.

This paper takes a critical yet balanced look at the role of MCQs in language assessment. Section 2 reviews how MCQs are used to assess various domains of language proficiency, outlines the format's unique affordances, and explores its well-documented limitations; this discussion draws on a targeted review of existing literature. Section 3 continues this literature-based analysis by situating MCQs within the broader landscape of assessment tools, comparing them to alternatives such as cloze tasks, short-answer questions, essays, and performance-based assessments. It also presents guidelines for writing high-quality MCQs that minimise construct-irrelevant variance and maximize diagnostic value. Section 4 offers a case study from Uzbekistan, where language assessments built on MCQs have been centrally deployed as part of a national education reform effort.

2. Understanding multiple-choice questions in language assessment

The discussion in this section is based on a targeted review of academic literature related to the design, application, and evaluation of multiple-choice questions (MCQs) in language assessment. Sources were identified through searches in Google Scholar and JSTOR using combinations of terms such as "language assessment," "multiple-choice questions," "test validity," and "construct-irrelevant variance." Peer-reviewed journal articles, book chapters, and published handbooks were prioritised, with a particular focus on texts that examined MCQs from psychometric, pedagogical, or theoretical perspectives.

This review is selective rather than systematic. It does not aim to exhaust all available research on MCQs, nor does it include unpublished studies, non-English sources, or research situated solely in content-area testing outside of language education. As such, the review reflects a curated synthesis of influential and illustrative sources, rather than a comprehensive mapping of the field. The goal is not to settle the debate around MCQs, but to clarify their pedagogical affordances, psychometric features, and limitations as they pertain to language assessment specifically.

2.1 Multiple-choice questions as a tool for valid language testing

Multiple-choice questions (MCQs) have long been a staple in language testing. While often criticised for being limited in their usefulness, they continue to play an important role in language assessment in a wide variety of contexts. This section explores how MCQs can be used to validly assess different elements of language proficiency.

2.1.1. MCQs and the assessment of language knowledge

One of the most effective uses of MCQs is in the assessment of linguistic knowledge, particularly knowledge of grammar and vocabulary. Selected-response formats like MCQs offer a practical way to measure learners' knowledge, which is often seen as a reliable indicator of broader linguistic competence (Phakiti & Leung 2024). For example, a vocabulary item might ask learners to choose the word closest in meaning to a given term in context. Hughes and Hughes (2020) contend that such items, when presented in authentic contexts, can assess both the breadth and depth of a candidate's vocabulary knowledge.

Grammar MCQs, likewise, can assess knowledge of a range of structures and forms, from verb tenses to syntactic choices. When distractors are carefully constructed to reflect common learner errors or misconceptions, such questions can diagnose specific areas of grammatical weakness or strength. As Phakiti and Leung (2024) contend, when MCQs go beyond rote recall and target higher-level syntactic processing, they can provide a valid picture of a candidate's grammar knowledge.

2.1.2. Reading and listening comprehension

MCQs have proven especially effective in assessing receptive skills, particularly reading and listening comprehension. In large-scale, high-stakes assessments such as TOEFL or IELTS, reading comprehension tasks often involve a passage followed by several MCQs designed to test understanding of the main ideas, supporting details, inferences, and vocabulary in context in that passage. According to Hughes and Hughes (2020), these tests have demonstrated a high degree of reliability and validity when subjected to rigorous item analysis and piloting.

Similarly, MCQs can be a valid format for listening assessments. For example, test-takers may listen to a conversation or a lecture and then answer comprehension questions. If the options are plausible and carefully worded, such tasks can effectively discriminate among different levels of listening proficiency. A study of the United Kingdom's GCE O-Level English exam, for example, found that MCQs in listening comprehension could assess

inferencing, discrimination of sound, and understanding of tone, but only if the audio was authentic and the questions well-aligned with the listening text.

2.1.3. Research on correlation with other measures

Studies have also demonstrated that MCQ-based tests can correlate well with more integrative or open-ended measures of language proficiency. Ajideh and Esfandiari (2009), for instance, found a significant correlation between candidates' results on a multiple-choice vocabulary test and a cloze test of lexical proficiency. This suggests that discrete-point MCQs can reflect language proficiency when well-constructed.

2.1.4. Pragmatic and discourse-level competence

Although less common, MCQs can also be adapted to assess pragmatic competence and discourse understanding. For instance, items might ask students to choose the most appropriate response in a given social context, or to select the correct interpretation of speaker intention in a dialogue. These formats have been incorporated into computer-delivered tests such as the TOEIC and the Cambridge English suite, often under the heading of "functional language" or "pragmatic use."

2.2 Assessment features exclusive to multiple-choice formats

As the preceding discussion illustrates, well-designed multiple-choice questions (MCQs) can serve as a valid and reliable means of assessing various dimensions of language proficiency. Moreover, beyond their versatility across language domains, MCQs also offer unique benefits not presented by other assessment formats. In particular, MCQs offer distinctive advantages in terms of efficiency, standardisation, scalability, and diagnostic precision - advantages that are not easily replicated by open-ended or performance-based tasks.

2.2.1. Standardization and scalability

One of the most significant advantages of MCQs is their capacity for standardization. Unlike performance-based tasks, such as essays or interviews, MCQs ensure that all test-takers receive the same input, face the same constraints, and are scored against identical criteria. This uniformity is crucial in large-scale testing contexts, where consistency across thousands, or even millions, of test-takers is necessary. As Bachman and Palmer (2010) note, selected-response formats like MCQs can be scored consistently and eliminate the variability

introduced by human raters, thereby supporting more reliable comparisons across individuals and groups.

This standardisation also facilitates scalability. MCQs can be rapidly administered, scored, and analysed, even across international boundaries, making them uniquely well-suited for high-stakes contexts (Im & Cheng 2019). By contrast, constructed-response format are substantially more resource-intensive and require far greater resources to implement at scale, making them less feasible and, indeed, in some cases, impossible to implement.

2.2.2. Statistical precision and item-level diagnostics

MCQs also allow for a level of psychometric precision that is difficult to achieve in other formats. Each MCQ item can be pretested, calibrated, and statistically analysed to ensure appropriate levels of difficulty and discrimination. Through Item Response Theory (IRT), test developers can assess how well each item distinguishes between higher- and lower-proficiency learners and how well it fits the intended ability scale (Rauch & Hartig 2010). This enables the construction of tests that are not only valid but also fine-tuned to target specific levels of language proficiency.

Moreover, the diagnostic capabilities of MCQs extend to sub-skill analysis. Well-constructed MCQs can be designed to reflect discrete competencies, such as vocabulary breadth, grammatical sensitivity, inferential reading, or pragmatic comprehension, allowing score reports to offer nuanced feedback to test-takers and educators. For example, the TOEIC Listening and Reading test includes items specifically targeted at distinguishing between literal comprehension, inferencing, and recognition of organisational or rhetorical cues. Such targeted diagnostics are rarely possible with broader performance tasks, which tend to blur the boundaries between constructs.

2.2.3. Controlled measurement of specific constructs

Unlike open-ended tasks, MCQs offer unparalleled control over the linguistic and cognitive variables involved in test performance. In writing or speaking assessments, a wider range of uncontrolled factors, such as topic familiarity, anxiety, or even handwriting, can interfere with construct measurement. In contrast, MCQs allow test developers to isolate and examine specific language features under tightly controlled conditions. As Phakiti and Leung (2024) argue, this precision is particularly valuable when assessing language knowledge, as discussed earlier.

This level of control also supports construct validity. While open-ended responses may involve construct-irrelevant variance, MCQs can be deliberately structured to reflect only the targeted ability. For instance, a vocabulary item that asks test-takers to choose the synonym of a word in context can be carefully designed to avoid confounding influences like syntactic complexity or topical bias (Hughes & Hughes 2020). By tightly focusing on the construct being measured, MCQs reduce measurement error and improve the interpretability of test scores.

2.2.4. Efficient feedback loops for learning

Another distinct strength of MCQs lies in their pedagogical utility, particularly when used formatively. MCQs can be scored instantaneously and can provide immediate feedback, making them well-suited for iterative learning environments. As Paxton (2000) notes, multiple-choice formats, when accompanied by opportunities for explanation and revision, can foster meta-linguistic awareness and help learners distinguish between similar language forms.

Moreover, MCQs allow for the rapid collection of data on learner performance, enabling instructors to identify patterns of misunderstanding across cohorts. For example, if a significant number of students consistently select a particular distractor in a grammar MCQ, this can indicate a widespread misconception worth addressing in subsequent instruction. No other test format enables this kind of large-scale, real-time instructional feedback as efficiently as MCQs.

2.2.5. Resistance to scorer bias

Constructed-response tasks, such as essays and interviews, are inevitably subject to scorer interpretation. Even with detailed rubrics and rigorous training, human raters vary in their judgments, and inter-rater reliability can be difficult to maintain. By contrast, MCQs eliminate scorer bias entirely when computer-scored, ensuring that every test-taker is evaluated in exactly the same way. This objectivity makes MCQs especially valuable in contexts where fairness and equity are critical, such as admissions or employment screening.

This is not to suggest that MCQs are immune to design bias, but rather that, once items are finalized, they are applied uniformly. As Bachman and Palmer (2010) emphasise, eliminating rater subjectivity enhances the reliability and defensibility of decisions made on the basis of test scores.

2.2.6. Benchmarking and equating across test forms

Because MCQs lend themselves so well to statistical calibration, they are the only format that allows for precise test equating, that is, ensuring that scores from different test forms are comparable. This is essential in high-stakes settings, where multiple versions of a test must be administered over time. For example, the TOEIC maintains score comparability through extensive use of pre-equated MCQ items that anchor different test forms to the same scale (Im & Cheng 2019).

Without MCQs, this kind of equating would be much more difficult to implement. Constructed-response tasks are harder to standardise, and their subjective scoring makes statistical equating less stable. The result is that only MCQ-based assessments can provide defensible, longitudinal data suitable for benchmarking and large-scale policy decisions.

2.2.7. Subtle measurement of cognitive load and error patterns

Finally, MCQs offer a unique opportunity to capture subtle features of test-taker cognition. For instance, distractors can be crafted to reflect common interlanguage errors, allowing researchers to infer not only whether an answer was correct, but why a learner chose an incorrect option. As Ajideh and Esfandiari (2009) show, even simple MCQs can reflect deep lexical or syntactic knowledge when distractors are informed by corpus-based error analysis.

Additionally, MCQs can be embedded in adaptive testing systems, in which the difficulty of subsequent items is determined by the test-taker's previous responses. This type of complex, technology-based approach is only feasible when item performance is predictable and statistically modeled - something possible only with large banks of calibrated closed-format items such as MCQs. These systems allow for shorter tests with higher measurement precision, a format increasingly favored in digital learning environments (Phakiti & Leung 2024).

2.3 Limitations of multiple-choice formats in language assessment

While the preceding section has highlighted the distinctive advantages of MCQs, it is equally important to acknowledge the limitations that have led to persistent critique. This section examines these and other limitations, drawing on empirical studies and theoretical arguments to provide a balanced perspective on the role of MCQs in language assessment.

2.3.1 Inability to assess productive language skills

One of the most commonly cited limitations of MCQs is their inability to assess productive language skills. Writing and speaking, which require the candidate to generate original output, categorically cannot be validly measured through selected-response formats (Hughes & Hughes 2020). Even with scenario-based MCQs intended to simulate communicative tasks, the test-taker remains constrained to choosing among preformulated options. As such, the construct validity of MCQs is inherently limited when the assessment goal includes language production (Phakiti & Leung 2024).

2.3.2 Shallow processing and strategic test-taking

This limitation extends beyond speaking and writing. Paxton (2000) argued that MCQ-based tests often reward a candidate's ability to identify correct answers rather than to explain or justify them. In language learning contexts, this can distort learners' focus; rather than developing skills such as argumentation, synthesis, or the manipulation of complex grammatical structures, learners may be trained simply to recognize patterns and eliminate distractors. Paxton ultimately concluded that students who are assessed primarily via MCQs "are not given enough opportunity to develop communicative competence" (p. 110).

Another concern is that MCQs can encourage test-wise strategies that do not reflect genuine language ability. For instance, some learners become adept at identifying distractors based on their form rather than meaning. This raises the issue of construct-irrelevant variance, that is, variance in test scores attributable to factors other than the intended construct, such as familiarity with test formats or guessing skills (Bachman & Palmer 2010). Given that it is not uncommon for candidates to retake high-stakes exams multiple times, this can give an unfair advantage to students with more test-taking experience, thereby compromising the assessment.

2.3.3 Fragmentation and loss of integrative language use

Moreover, MCQs often fail to represent the integrative nature of real-world language use. As Argüelles Álvarez (2013) notes in her reflection on grammar assessment in large-scale testing, MCQs may simplify language knowledge into discrete, decontextualized units, which risks neglecting the interconnectedness of grammatical, lexical, and pragmatic knowledge. While efforts have been made to create more contextually grounded items, the format inherently lends itself to fragmentation, undermining the assessment of holistic language competence.

Empirical studies have also pointed to the limitations of MCQs in differentiating among higher-level cognitive processes. Rauch and Hartig (2010), for example, found that while MCQs could measure basic comprehension skills in reading, they were less effective at assessing complex inferencing or evaluative skills. This echoes earlier concerns by Oller (1979) and others, who argued that MCQs tend to prioritize recognition over active language processing, a distinction that bears heavily on the cognitive validity of such tasks.

2.3.4 Lack of authenticity

Even when MCQs are used to assess receptive skills, they can suffer from low authenticity. Simply put, in natural language use, people do not respond to multiple-choice prompts. This problem is particularly evident in listening tests, where the artificial nature of pausing for questions and offering choices can interfere with the flow of comprehension (Hughes & Hughes 2020). As a result, test-takers may not engage with the listening text in a way that mirrors real-world processing.

2.3.5 Educational consequences

Finally, from an educational perspective, MCQs can have negative washback on teaching and learning. When high-stakes exams rely heavily on this format, instructional practices may be skewed toward drilling test-taking strategies rather than fostering communicative or critical thinking skills. Paxton (2000) emphasizes that assessment methods powerfully shape what and how students learn, and warns that the dominance of MCQs can marginalize more meaningful forms of language use.

3. Beyond MCQs: contextualizing their role

This section takes a break from the narrow ‘what’ and ‘how’ of MCQs and attempts to situate them in the broader context of assessment practices. Section 3.1 compares MCQs with other common language assessment formats, while section 3.2 provides a review of best practices for teachers and test developers when drafting MCQs.

3.1 Strengths and limitations of other assessment formats

While MCQs have long been criticised for their perceived limitations in measuring communicative competence, the alternatives often touted as superior also require careful

scrutiny. Constructed-response formats, such as cloze tests, short answer questions, essays, interviews, and integrated tasks, do offer advantages, especially in assessing productive language use. However, they also come with significant trade-offs. This section examines the comparative merits and limitations of these alternatives, arguing that while alternative formats can address some of MCQs' weaknesses, they are not inherently, categorically superior.

3.1.1 Cloze tests and gap-fills

Cloze and gap-fill tests are often positioned as more integrative than MCQs, since they require test-takers to actively produce language elements based on contextual cues. Proponents argue that these tasks tap into both grammatical and discourse-level knowledge. Indeed, Ajideh and Esfandiari (2009) found a statistically significant correlation between cloze performance and vocabulary knowledge in Iranian learners, suggesting cloze tests can be useful proxies for integrative proficiency.

However, cloze tests have well-documented scoring and reliability issues. Responses may be semantically or grammatically appropriate but not match the scoring key. Hughes and Hughes (2020) caution that even carefully constructed cloze tests require intensive piloting and multiple acceptable answers to avoid unfair penalization. Furthermore, cloze tasks are vulnerable to construct-irrelevant variance caused by guessing or familiarity with specific lexical collocations.

3.1.2 Constructed-response questions

Short-answer and essay questions allow test-takers to demonstrate productive language skills and higher-order cognitive processing. These tasks are especially valuable in academic and professional contexts where argumentation, synthesis, and coherence are key competencies. Writing tasks in particular allow for the assessment of discourse organisation, lexical choice, and grammatical range, among other competencies.

However, these benefits come at a cost. Scoring is subjective and requires intensive rater training to maintain consistency. Inter-rater reliability remains a persistent challenge - even among trained examiners. As Bachman and Palmer (2010) note, variation in scoring due to rater bias or fatigue can undermine the fairness and validity of constructed-response formats, especially in high-stakes contexts. Moreover, open-ended tasks are difficult to standardise and time-consuming to administer, making them impractical for large-scale testing.

3.1.3 Performance-based tasks

Oral interviews and integrated tasks such as presentations or group discussions are lauded for their authenticity. They align closely with real-world communicative goals and offer rich data on fluency, pronunciation, interactional competence, and pragmatic appropriacy. Tests like the TOEFL iBT and the Cambridge suite have increasingly embraced these formats to capture a broader picture of learner proficiency.

Still, the logistical challenges are substantial. Speaking tasks require trained interlocutors, secure audio capture, calibrated raters, and often elaborate scheduling. Even computerised oral tests, such as the TOEIC Speaking, demand robust technical infrastructure and still grapple with questions of rating validity (Im & Cheng 2019). Moreover, performance-based tasks often yield less granular diagnostic feedback, making it harder to pinpoint specific areas of weakness compared to targeted MCQ formats.

3.1.4 Portfolios and self-assessments

Portfolio assessments offer a longitudinal view of learner progress and can showcase writing samples, projects, or oral performances. Advocates argue that portfolios promote reflective learning and allow students to demonstrate growth over time. Similarly, self-assessments foster learner autonomy and metacognitive awareness.

Yet these formats are fundamentally unsuitable for high-stakes decisions. They lack standardisation, are prone to subjectivity, and cannot be easily scaled. Moreover, students from different cultural or educational backgrounds may approach self-assessment tasks with varying levels of confidence or familiarity, reducing score comparability.

3.2 Principles for writing effective multiple-choice items

As the previous section has shown, alternative assessment formats offer valuable opportunities for measuring language proficiency, but they also present substantial challenges in terms of reliability, scalability, and standardization. Given these trade-offs, MCQs retain a prominent place in both large-scale and classroom-based language assessment. However, their effectiveness depends not on the format alone, but on the quality of the items themselves.

3.2.1. Align each item with a clearly defined construct

The first and most fundamental principle of MCQ design is construct alignment. Each item should target a specific, well-defined component of language proficiency. Without a clear

construct definition, it is impossible to determine whether an item measures what it claims to measure (Bachman & Palmer 2010). For instance, a vocabulary item should assess knowledge of word meaning in context, not just spelling or general world knowledge. Similarly, a grammar item should require the application of syntactic rules in realistic language use, not the memorisation of decontextualized forms.

3.2.2. Ensure the stem is clear, concise, and meaningful

The stem (the part of the question that presents the problem) of an MCQ should be linguistically simple but cognitively challenging. It must provide enough context to make the item meaningful, but avoid unnecessary complexity or ambiguity. According to Hughes and Hughes (2020), stems should avoid negatives unless absolutely necessary, and test-writers should beware of cueing correct answers through grammatical agreement, verb tense, or collocation patterns.

Consider the following poor stem:

> *Which of the following is not incorrect in the sentence below?*

This stem uses a double negative and vague language, which may confuse candidates who otherwise would be able to answer the question correctly. A better version would be:

> *Which of the following correctly completes the sentence?*

When testing grammar or vocabulary, stems should ideally be embedded in authentic or as-close-to-authentic-as-possible discourse. Paired with brief emails, dialogues, or signs, this increases the ecological validity of the task without sacrificing clarity.

3.2.3. Use plausible and functionally diagnostic distractors

The effectiveness of an MCQ depends as much on its distractors as on its key. Distractors (incorrect answer options) should be plausible and based on common learner errors, transfer patterns, or developmental sequences. As Phakiti and Leung (2024) explain, high-quality distractors provide insight into what kind of misunderstanding the test-taker may have.

Consider, for example, an item testing the present perfect:

> *I _____ in London for five years.*
> a) lived

b) have lived
c) am living
d) live

Each distractor represents a plausible-looking but contextually inappropriate choice. A well-designed distractor set allows test developers to diagnose whether an error stems from, for example, tense confusion, aspect misunderstanding, or present-time reference errors. Distractors that are obviously wrong - grammatically, semantically, or typographically - should be avoided, as they lower the discriminative power of the item and invite test-wise guessing (Burton et al. 1991).

3.2.4. Avoid unintended cues and patterns

Test-writers must take care to avoid giving away the correct answer through superficial cues. These include:

- The correct answer being consistently longer or more detailed
- Grammatical mismatches between the stem and incorrect options
- Syntactic repetition or collocation "echoes" from the stem to the correct answer

As noted by Argüelles Álvarez (2013), statistical item analysis is essential to detect such flaws. If a distractor is never chosen, it may be too obviously incorrect. Conversely, if an item is too difficult or behaves erratically across groups, it may contain hidden ambiguities or construct-irrelevant difficulty.

3.2.5. Maintain cultural and linguistic fairness

MCQ items should avoid culturally specific references or topical content that could disadvantage test-takers from particular linguistic or ethnic backgrounds. For example, asking about baseball terminology or culturally specific holidays in an English proficiency test risks introducing construct-irrelevant variance. Test fairness demands that MCQs reflect general rather than culture-bound knowledge, unless the test explicitly aims to assess sociocultural competence.

3.2.6. Pilot, revise, and analyze

No matter how experienced the test writer, MCQs must be piloted before being operationalized. Empirical item analysis allows developers to assess item difficulty,

discrimination, and distractor performance. Poor-performing items can then be revised or removed. As Burton et al. (1991) argue, even items that appear well-written should be empirically evaluated through item analysis. The actual quality and functioning of an MCQ can only be determined by how test-takers respond to it, not by surface inspection alone.

In large-scale assessments like the TOEIC, item development is embedded within an evidence-centred design framework, and items undergo extensive pretesting, statistical analysis, and procedural review prior to inclusion in operational forms. Items that do not meet psychometric criteria, such as appropriate levels of difficulty and discrimination, are excluded from scored sections, regardless of their linguistic plausibility (Schmidgall, 2017). The validity argument for the TOEIC is supported not only by design documentation, but also by ongoing statistical monitoring and research publication review, all of which contribute to justifying the reliability, generalisability, and fairness of the test.

4. Case study: MCQ language assessment in Uzbekistan

The next section of this paper provides a brief case study of one organisation, The Pedagogical Excellence and International Assessment Centre in Uzbekistan, and its use of MCQs in language assessment. The study will begin with a brief overview of the history and mandate of the organisation, before going on to discuss how it uses MCQs to meet that mandate.

4.1 The Pedagogical Excellence and International Assessment Centre in Uzbekistan: its role and mission

4.1.1. Institutional establishment and legal mandate

The Pedagogical Excellence and International Assessment Scientific-Practical Centre, hereafter referred to simply as the Assessment Centre, operating under the Specialized Educational Institutions Agency of the Ministry of Preschool and School Education of Uzbekistan, is a state institution tasked with bringing about innovation in assessment and teacher development. The Assessment Centre was formally established by Cabinet of Ministers Resolution No. 741, dated December 7 2021, and later reorganised through subsequent resolutions to reflect its expanded remit in pedagogical training and large-scale assessment (O‘zbekiston Respublikasi Vazirlar Mahkamasi 2025).

The establishment of the Assessment Centre was part of a broader reform initiative set in motion by Presidential Decree PQ-5188 dated July 19 2021, which aimed to modernise the

operation of schools across Uzbekistan. According to official government documentation, the Assessment Centre's responsibilities include: designing and piloting national curriculum and assessment materials; conducting final certification of graduates; assessing and certifying teacher knowledge and skills; and implementing international standards in collaboration with leading global research institutions (O'zbekiston Respublikasi Vazirlar Mahkamasi 2025). Broadly speaking, it can be said that the Assessment Centre is the leading body in Uzbekistan for the development, implementation, and analysis of large-scale assessments for both students and their teachers.

4.1.2 Innovation, international alignment, and MCQ development

One of the largest initiatives of the Assessment Centre has been the introduction of a criterion-referenced, 100-point assessment system in over 1500 general education schools, replacing the legacy 5-point system. This pilot aims to enhance transparency and equity by aligning classroom assessment more closely with modern pedagogical principles, including higher-order cognitive skills, critical thinking, and communicative competence (Gazeta.uz 2023). The Assessment Centre explicitly frames its mission around the so-called "4C" competencies - creative thinking, critical thinking, communication, and collaboration - as foundational goals of education reform in Uzbekistan.

The Assessment Centre's work purports to reflect an effort to harmonise national education policy with international practices observed in systems such as Singapore and Finland, while remaining grounded in the local Uzbekistani context. As part of this mission, it has spearheaded the development of assessment instruments, including those using multiple-choice questions (MCQs), for use in both classroom settings and high-stakes national evaluations. These efforts form the basis for the following analytical section, which examines how MCQs are publicly deployed in language assessments produced by the Assessment Centre.

4.2 MCQs in Uzbekistan's End-of-Quarter Summative Assessment (ChSB): A Document-Based Analysis

4.2.1 Role and structure of the ChSB

Among the various assessment instruments developed or endorsed by the Assessment Centre, the *chorak bo'yicha summativ baholash* (ChSB), or end-of-quarter summative assessment, has

emerged as one of the most widely implemented. Designed to support the shift toward a standardised, criterion-referenced system, the ChSB reflects many of the Centre's stated priorities, including alignment with curriculum outcomes, use of scalable formats, and the application of MCQs in both language and content-area testing.

As of 2025, the ChSB is administered in approximately 1,500 state schools, with plans for continued expansion. It offers a revealing case study of how MCQs can be deployed in large-scale assessment at the national level. This section analyses the design and implementation of the ChSB using official test documents, focusing on the structure and function of its MCQ components and the extent to which they reflect (or depart from) established principles of valid language assessment.

In the subject of English, ChSB exams are administered from Grades 5 to 11. Each test consists of 15 questions, all of which are multiple-choice questions with four options and a single correct answer. The format is used consistently across grade levels to assess a range of competencies including grammar, vocabulary, and reading comprehension. These items are developed centrally and distributed nationwide to support the broader aims of comparability, fairness, and curricular alignment.

4.2.2 Curriculum alignment and item competencies

A distinctive feature of the ChSB English assessments is their competency-based structure. Each test item is explicitly aligned to a defined skill statement, known as a *kompetensiya*, which corresponds directly to the national English curriculum. These competencies serve as operational goals drawn from CEFR-linked learning outcomes. For example, Grade 9 ChSB sample items include competencies such as:

- *Can complete passive sentence with correct verb form in present tense.*
- *Can distinguish between first conditional and other tenses or forms. (citation needed*)

This structure enhances the validity of the assessment by ensuring that each MCQ targets a curriculum-relevant language skill. Moreover, it provides a clear pedagogical link between classroom instruction and summative testing, enabling teachers to align their lessons to the expected outcomes and interpret results more diagnostically.

4.2.3 Item design and skill coverage

In terms of item design, the majority of ChSB English MCQs are sentence-level or discrete-point items. For instance:

Please _____ the glass with water.

A) fill

B) cover

C) boil

D) dry

This item corresponds to a vocabulary competency on kitchen actions and tests for the appropriate verb in context. The distractors are relatively plausible but incorrect, and the key is clearly supported by the syntactic structure and semantic fit. Such items exemplify the use of MCQs to efficiently assess receptive knowledge of lexis and structure.

Other items test grammar, such as:

She won't go to the party if _____.

A) she feels sick

B) she will feel sick

C) she feeling sick

D) she felt sick

Here, the item assesses knowledge of the first conditional and corresponds to a Grade 9 grammar competency. Again, the distractors reflect typical learner errors, and the stem is simple and clear, contributing to the item's overall functionality.

Notably, the ChSB also includes integrative reading comprehension tasks. The Grade 9 sample ends with a short biography-based text and a four-option MCQ asking test-takers to identify the incorrect statement about the passage. This item requires close reading, synthesis of details, and inference, reflecting a move beyond surface-level processing.

4.2.4 Local critiques of the ChSB

Jalilov (2024), writing in Uzbekistani news site *Gazeta.uz*, asserted that the new assessment system undermines the pedagogical role of teachers and narrows the scope of what is being measured. He argued that by centralising the design and marking of summative assessments, the system excludes teachers from key aspects of the evaluative process and reduces assessment to a form of external control. According to Jalilov, this shift limits the capacity of schools to provide formative feedback or adapt instruction to learner needs. He further warned that the overemphasis on standardised test outcomes for high-stakes decisions creates systemic pressures that distorted both teaching practices and student learning goals. Drawing on

Campbell's Law, he contended that the system's heavy reliance on summative scores incentivizes superficial performance rather than meaningful educational engagement.

Shortly after the publication of Jalilov's article, the Ministry of Preschool and School Education, of which the Assessment Centre is a part, issued an official response in the same newspaper (*Gazeta.uz* 2024), defending the new approach as consistent with international best practices. The Ministry maintained that the new structure ensures objectivity and fairness, and that formative tasks, though limited in weight, retained pedagogical value. It argued that such a model is necessary to establish comparability across schools and regions while the national education system continues to modernise. While the Ministry expressed an eventual aim to restore greater autonomy to teachers, it suggested that current conditions do not yet permit a fully decentralised model.

4.2.5 The role of BSB as a complementary assessment

It is worth noting that the ChSB is only one component of the Assessment Centre's assessment system. The corresponding *bob bo'yicha summativ baholash* (BSB), or end-of-unit summative assessment, is also aligned to the national curriculum but typically employs a broader range of item formats. In contrast to the ChSB's exclusively MCQ structure, BSBs often include matching tasks, gap-fills, sentence transformations, and short-answer items. This complementary design enables the system to strike a balance: ChSBs prioritise objectivity, consistency, and scalability, while BSBs offer greater flexibility and are better suited to capturing productive and integrative aspects of language competence.

5. Conclusion

Despite long-standing critiques, MCQs continue to occupy a central position in language assessment. Their enduring relevance lies not in their universality, but in their versatility. When carefully designed, MCQs can validly and reliably assess a wide range of linguistic competencies, from grammatical knowledge to inferential reading and pragmatic comprehension. Their efficiency, standardisation, and diagnostic precision make them particularly well-suited to large-scale assessment contexts.

At the same time, their limitations must be acknowledged. MCQs cannot assess productive language skills, often oversimplify complex communicative processes, and may encourage superficial test-taking strategies. These drawbacks underline the importance of

situating MCQs within a broader, mixed-methods approach to assessment: one that balances their strengths with the depth and authenticity offered by constructed-response and performance-based tasks.

As the case study of the ChSB in Uzbekistan demonstrates, the institutional use of MCQs can reflect broader goals of educational reform, international alignment, and system-wide comparability. But it also raises questions about test purpose, teacher agency, and the risks of over-reliance on a single format. Ultimately, the question is not whether MCQs are good or bad, but whether they are fit for purpose in a given context. Their value lies not in the format itself, but in the quality of the thinking behind their use.

References

Ajideh, Parviz, Reza Esfandiari (2009) "A Close Look at the Relationship between Multiple-Choice Vocabulary Test and Integrative Cloze Test of Lexical Words in Iranian Context." *English Language Teaching*, 2(3); 163–170. https://doi.org/10.5539/elt.v2n3p163

Argüelles Álvarez, Inmaculada (2013) "Large-scale Assessment of Language Proficiency: Theoretical and Pedagogical Reflections on the Use of Multiple-choice Tests" *International Journal of English Studies*, 13(2); 21-38. https://doi.org/10.6018/ijes.13.2.185861

Bachman, Lyle, Adrian Palmer (2010) *Language Assessment in Practice.* Oxford: Oxford University Press.

Burton, Steven J., R. R. Sudweeks, P. F. Merrill, B. Wood (1991) *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty.* Provo: Brigham Young University Testing Services and the Department of Instructional Science. https://testing.byu.edu/handbooks/betteritems.pdf

Educational Testing Service (ETS) (2015) *TOEIC Listening and Reading Test: Examinee Handbook.* https://www.ets.org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf

Gazeta.uz (2023, August 15) "O'zbekistondagi 500 ta maktabda mezonga asoslangan (100 ballik) baholash tizimi joriy etildi. Nima o'zgaradi? " ["A criterion-based (100-point) assessment system has been introduced in 500 schools in Uzbekistan. What will change?"]. https://www.gazeta.uz/oz/2023/09/30/schools/

Hughes, Arthur, Julia Hughes (2020) *Testing for Language Teachers.* 2nd edn. Cambridge: Cambridge University Press.

Im, Gwan-Hyeok, Liying Cheng (2019) “The Test of English for International Communication (TOEIC®)” *Language Testing* 36(2). https://journals.sagepub.com/doi/10.1177/0265532219828252

Oller, John W. (1979) *Language Tests at School: A Pragmatic Approach.* London: Longman.

O‘zbekiston Respublikasi Vazirlar Mahkamasi [Cabinet of Ministers of the Republic of Uzbekistan] (2025) *O‘zbekiston Respublikasi Maktabgacha va maktab ta’limi vazirligi huzuridagi Ixtisoslashtirilgan ta’lim muassasalari agentligi huzuridagi Pedagogik mahorat va xalqaro baholash ilmiy-amaliy markazi faoliyatini tashkil etish to‘g‘risida* [“On the Organization of the Activities of the Scientific-Practical Centre for Pedagogical Excellence and International Assessment under the Specialized Educational Institutions Agency of the Ministry of Preschool and School Education of the Republic of Uzbekistan”]. Resolution No. 130, 28 February 2025. https://lex.uz/uz/docs/-7409825

Paxton, Moragh (2000) “A Linguistic Perspective on Multiple Choice Questioning.” *Assessment & Evaluation in Higher Education*, 25 (2); 109–119. http://dx.doi.org/10.1080/713611429

Phakiti, Aek, Constant Leung (2024) *Assessment for Language Teaching.* Cambridge: Cambridge University Press.

Rauch, Dominque P., Johannes Hartig (2010) “Multiple-choice versus Open-ended Response Formats of Reading Test Items: A Two-dimensional IRT Analysis” *Psychological Test and Assessment Modeling*, 52 (4); 354-379.

Schmidgall, Jeffery E. (2017) “Articulating and evaluating validity arguments for the TOEIC® tests.” ETS Research Report Series 2017 (1); https://doi.org/10.1002/ets2.12182

Джалилов, Комил [Jalilov, Komil] (2024, November 25) “«Стало невозможно ни учиться, ни работать». К чему привело внедрение «критериальной» системы оценивания в школах?” [“It Became Impossible to Study or Work”: What Has the Introduction of the ‘Criterion-based Assessment System in Schools Led To?”]. *Газета.uz.* https://www.gazeta.uz/ru/2024/11/25/evaluation/

Министерство дошкольного и школьного образования [Ministry of Preschool and School Education] (2024, November 30) “В МДШО ответили на статью о новой системе оценивания в школах” [“The Ministry Responded to the Article on the New School Assessment System”]. *Газета.uz.* https://www.gazeta.uz/ru/2024/11/30/evaluation/