# Intra-Text Gap-Filling in Multilevel English Language Tests: Design, Validity, and CEFR Alignment

**A.A. Abbosov**

*Agency for Assessment of Knowledge and Competences under the Ministry of Higher Education,*
*Science and Innovations of the Republic of Uzbekistan,*
*100084, Tashkent., Bogishamol str., 12*

**Abstract.** Gap-filling items are among the most common types of questions in language proficiency tests. They are particularly suitable for multilevel formats because they can target a broad range of difficulty levels and reflect authentic language processing. Intra-text gap-filling items, in which the input text serves as the word bank at the same time, are a type of test item that has been developed relatively recently. The article discusses the principles of designing intra-text gap-filling tasks, examines their validity, and evaluates how they can be aligned with the Common European Framework of Reference for Languages (CEFR).

**Keywords:** Multilevel English language tests, reading comprehension, testlet, gap-filling items, intra-text gap-filling items, validity, reliability, CEFR.

## Introduction

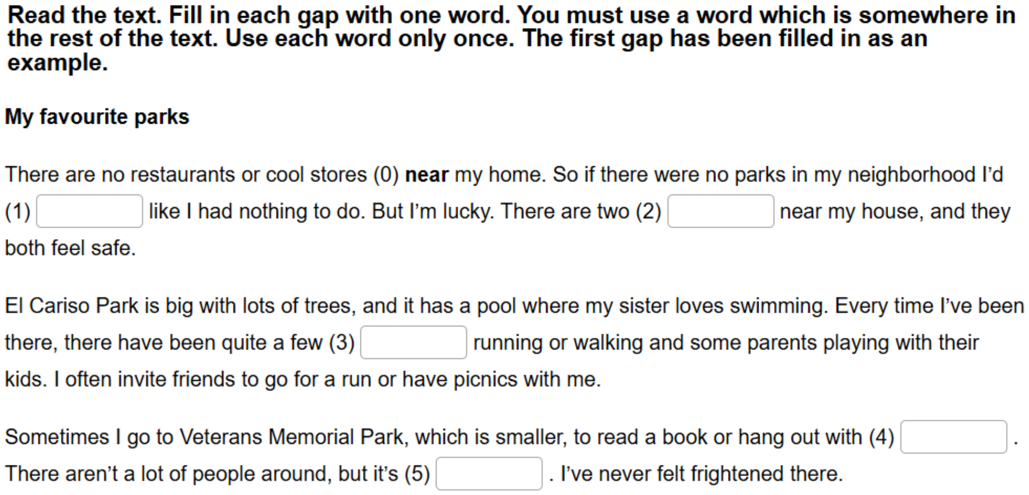
The multilevel English language proficiency test has been in use in Uzbekistan since 2022. A multilevel test is designed to assess language proficiency across a range of levels, from beginner to advanced, within a single exam. Developed as an alternative to one-level proficiency exams that were in use before 2022, this test gained popularity among test takers and other stakeholders due to its flexibility and inclusivity for assessing language proficiency across a broad spectrum of abilities. The test consists of four components, corresponding to each language skill. The listening and reading components are further subdivided into smaller sections, called testlets, each targeting a specific CEFR level. Several different types of questions are used to minimize construct-irrelevant variance – factors unrelated to language proficiency (e.g., test-taking strategy dependence), allowing test-takers to demonstrate their true language abilities. Alderson (2000) argues that high-quality reading tests use a variety of techniques, as this enhances the authenticity of the assessment [1]. Among the various question types employed, one notable example is the use of gap-filling questions, which are included in two of the testlets in the reading section.

## 1. Types of gap-filling items

Gap-filling tasks are considered to be particularly useful in testing reading [2]. Alderson (2000) states that they are often confused with cloze items [1]. The primary difference between cloze items and gap-filling tasks lies in the deletion process. In cloze items, as originally invented by Taylor (1953), words are removed randomly in the text, i.e., every n-th word is deleted [3]. In contrast, gap-filling tasks involve the removal of words based on a particular rationale, such as targeting content words that are essential to the main idea of the text or the words that carry the coherence of the text [1]. There are a few variations of gap-filling tasks:

- Open gap-filling tasks;
- Banked gap-filling tasks;
- Multiple-choice gap-filling tasks.

In open gap-filling tasks, the test taker is required to deduce the missing words, while in banked and multiple-choice gap-filling tasks, the words are provided before or after the text. A variation of banked gap-filling tasks is used in the Multilevel English language proficiency test. In this version, content words that are essential for maintaining the coherence of the text are deleted, and the remaining words of the text serve as a word bank from which test takers can choose the appropriate words to fill the gaps.



**Read the text. Fill in each gap with one word. You must use a word which is somewhere in the rest of the text. Use each word only once. The first gap has been filled in as an example.**

**My favourite parks**

There are no restaurants or cool stores (0) **near** my home. So if there were no parks in my neighborhood I'd (1) □ like I had nothing to do. But I'm lucky. There are two (2) □ near my house, and they both feel safe.

El Cariso Park is big with lots of trees, and it has a pool where my sister loves swimming. Every time I've been there, there have been quite a few (3) □ running or walking and some parents playing with their kids. I often invite friends to go for a run or have picnics with me.

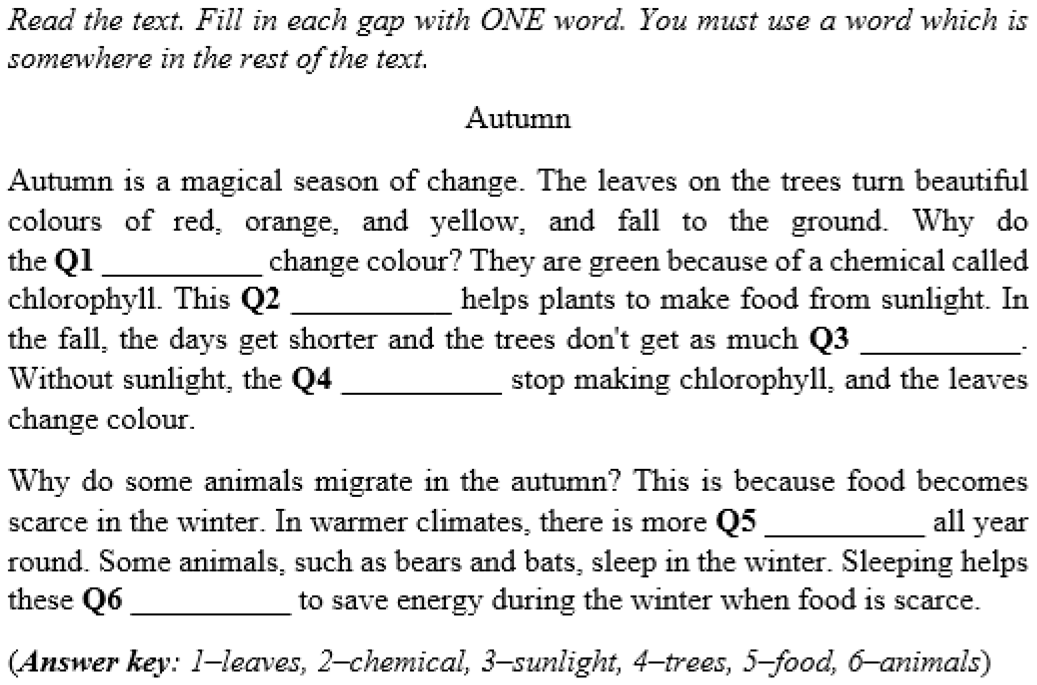
Sometimes I go to Veterans Memorial Park, which is smaller, to read a book or hang out with (4) □ . There aren't a lot of people around, but it's (5) □ . I've never felt frightened there.

*Fig.1. Intra-text gap-filling task*

## 2. Intra-text gap-filling items

Fig. 1 [4] shows an example of an intra-text banked gap-filling task. As can be seen from the example, gaps are to be filled with the words, which are somewhere in the rest of the text. They can be useful to test the overall understanding of the text as the content words are deleted, as well as to test the coherence of the text and building a mental model of the text because the deleted words are important to the internal links between the sentences in the text.

Here is another example, from the Multilevel English language tests:



*Read the text. Fill in each gap with ONE word. You must use a word which is somewhere in the rest of the text.*

Autumn

Autumn is a magical season of change. The leaves on the trees turn beautiful colours of red, orange, and yellow, and fall to the ground. Why do the **Q1** __________ change colour? They are green because of a chemical called chlorophyll. This **Q2** __________ helps plants to make food from sunlight. In the fall, the days get shorter and the trees don't get as much **Q3** __________. Without sunlight, the **Q4** __________ stop making chlorophyll, and the leaves change colour.

Why do some animals migrate in the autumn? This is because food becomes scarce in the winter. In warmer climates, there is more **Q5** __________ all year round. Some animals, such as bears and bats, sleep in the winter. Sleeping helps these **Q6** __________ to save energy during the winter when food is scarce.

(***Answer key:*** *1–leaves, 2–chemical, 3–sunlight, 4–trees, 5–food, 6–animals*)

*Fig.2. Sample testlet from Multilevel English language proficiency test bank*

In this example, the deleted words carry the coherence of the text. It is difficult to guess all the correct words without understanding the text. When test takers read the text, they deduce the word based on their understanding of the text and further verify it by linking the sentence with the preceding or the following sentence.

## 3. Item development process

This type of items in the Multilevel English language proficiency test targets lower B1 level. Therefore, texts are selected to suit this level of CEFR in terms of topic, length, lexical and grammatical complexity. After the text is chosen, it needs to be further edited to fit the level and specification requirements. The text should allow the deletion of six words as the testlet contain six items. A number of principles must be followed to maintain the quality of assessment:

- A few sentences at the beginning of the text must remain intact to provide enough context to understanding [1];
- The gaps must be spaced adequately, and the surrounding text should provide sufficient context and clues for the answers;
- Content words that are essential for maintaining the coherence of the text should be chosen for deletion;
- The selected words must appear elsewhere in the text, ensuring logical flow between sentences;
- Only content words should be deleted;
- Each word must be deleted only once;
- The answer key must include all valid answers, but synonyms or variations in phrasing are not acceptable.

## 4. CEFR alignment

Input texts as well as tasks in the Multilevel English language proficiency tests are aligned with CEFR according to the following parameters:

- Relevance of topic
- Nature of information
- Lexical and grammatical complexity
- Skill focus

A Core Inventory for General English [5], developed by British Council, provides core lists of discrete language points, topics and key lexis for each level of the CEFR. This document guides the development of test specifications, according to which the intra-text gap-filling task is aligned with the B1 level. During the text selection and editing process, item writers make sure that the text is within the capability of an average B1 learner in terms of topical, lexical and grammatical complexity.

According to the CEFR [6, 7], a B1-level reader can:

- understand the main points of clear standard texts on familiar matters regularly encountered in work, school, leisure, etc.;
- understand texts that consist mainly of high frequency everyday or job-related language;
- understand clearly written, straightforward texts on subjects related to their field and of personal interest.

This means the text should not be overwhelmingly long (around 150 words, according to the Multilevel English language proficiency test specifications [8]) and should not discuss abstract issues. These descriptors are embedded in the test specifications and monitored closely during the item design and review processes.

Table 1 demonstrates the CEFR alignment mapping in terms of task focus:

| **Aspect** | **CEFR B1 Relevance** | **How gap-filling task aligns with CEFR** |
|---|---|---|
| Reading for gist and detail | Learners must understand the main ideas and key details | Gaps placed at key points (e.g., content words) test understanding. |
| Contextual inference | Learners are expected to infer meaning from context | Choosing the correct word to fill a gap often requires using surrounding context. |
| Cohesion and coherence | Learners can follow the thread of a text and basic cohesion | Gaps involving reference words (e.g., “that,” “this”) test this skill. |

*Table 1. CEFR alignment mapping of intra-text gap-filling tasks*

In summary, intra-text gap-filling tasks at the B1 level strongly align with CEFR expectations if they:

- are based on texts accessible to B1 learners,
- focus on B1-level language forms and meanings,
- require comprehension of both sentence-level and discourse-level information,
- support integrative skills like vocabulary use, grammar, cohesion, and context interpretation.

## Conclusion

Table 2 illustrates the analysis of the sample intra-text gap-filling task (Fig. 2) used in the Multilevel English language proficiency tests. The items are analyzed with classical test theory using SPSS, as well as the item-response theory. Each item in the testlet is given a special code, which indicates the year (2023) and month (November) the item is administered, as well as the session number and item’s location in the test booklet.

| Item | Mean (difficulty) | Standard deviation | Correlation (discrimination) | Cronbach's alpha if deleted | Cronbach’s alpha |
|---|---|---|---|---|---|
| *23No101* | 0.691 | 0.462 | 0.406 | 0.808 | 0.816 |
| *23No102* | 0.776 | 0.417 | 0.334 | 0.811 | |
| *23No103* | 0.626 | 0.484 | 0.410 | 0.808 | |
| *23No104* | 0.799 | 0.401 | 0.314 | 0.811 | |
| *23No105* | 0.657 | 0.475 | 0.497 | 0.804 | |
| *23No106* | 0.861 | 0.346 | 0.413 | 0.809 | |

*Table 2. Item performance parameters*

The table shows how well the testlet items are performing. The mean difficulty of each item represents its relative difficulty, with values closer to 1 indicating easier items and values closer to 0 indicating more difficult ones.

Correlation value shows whether the item is discriminating more knowledgeable test takers from less knowledgeable ones. The rule of thumb is that this value needs to be greater than 0.2 to consider it an acceptable item [9, 10, 11]. The analysis shows that all the items in the testlet are performing really well in terms of item discrimination. Higher-quality items contribute positively to internal consistency, whereas poor-quality items can decrease it. ‘Bad items’ are those that fail to discriminate well between higher and lower ability test-takers or fail to contribute to the overall reliability of the test.

Cronbach’s alpha is calculated for the whole reading section administered in November 2023. Cronbach’s alpha is a measure of the internal consistency of the test, or how reliably the items measure the same underlying construct. A value greater than 0.7 indicates high internal consistency [12]. The ‘Cronbach’s alpha if deleted’ column shows how the internal consistency of the test changes when an item is removed. If the alpha value decreases, it means the item is contributing positively to the test’s overall reliability. The table shows that all of the items in the testlet are contributing positively to the internal consistency of the test, which in turn supports the construct validity of the test [9, 10, 11].

While Cronbach’s alpha provides insight into the overall internal consistency of the test, item-response theory (IRT) offers a more detailed analysis of how well the individual items perform, particularly in terms of difficulty and fit. The calculations are done using dexter packet in R, a software environment for statistical computing and graphics [13]. The item difficulty from the Rasch model (Table 3) is represented by ‘beta’ values: negative values indicate easier items, while positive values indicate more difficult ones:

| item_id | beta | SE_beta |
|---|---|---|
| *23No101* | -0.780592 | 0.023149 |
| *23No102* | -1.224829 | 0.024546 |
| *23No103* | -0.363351 | 0.022502 |
| *23No104* | -1.375662 | 0.025204 |
| *23No105* | -0.434099 | 0.022569 |
| *23No106* | -1.833887 | 0.027857 |

*Table 3. Item difficulty from the Rasch model*

Item-response theory (IRT) provides a more nuanced view of item performance. Fig. 4 shows that response patterns for items 5 and 6 do not align well with the model's predictions, indicating potential issues with these items' performance or fit within the test structure. Fig. 5 shows the relationship between item difficulty and person ability. IRT describes item difficulty and person ability in the same continuum to show whether test takers were given the items within their capability or not. In the IRT model of the Multilevel English language proficiency tests, B1-level learners typically fall between -1.5 and 0. The graph in Fig. 5 shows that five out of six items fall within this range, indicating that the test items are appropriately calibrated for B1-level learners' abilities. This can also be considered as evidence of construct validity.

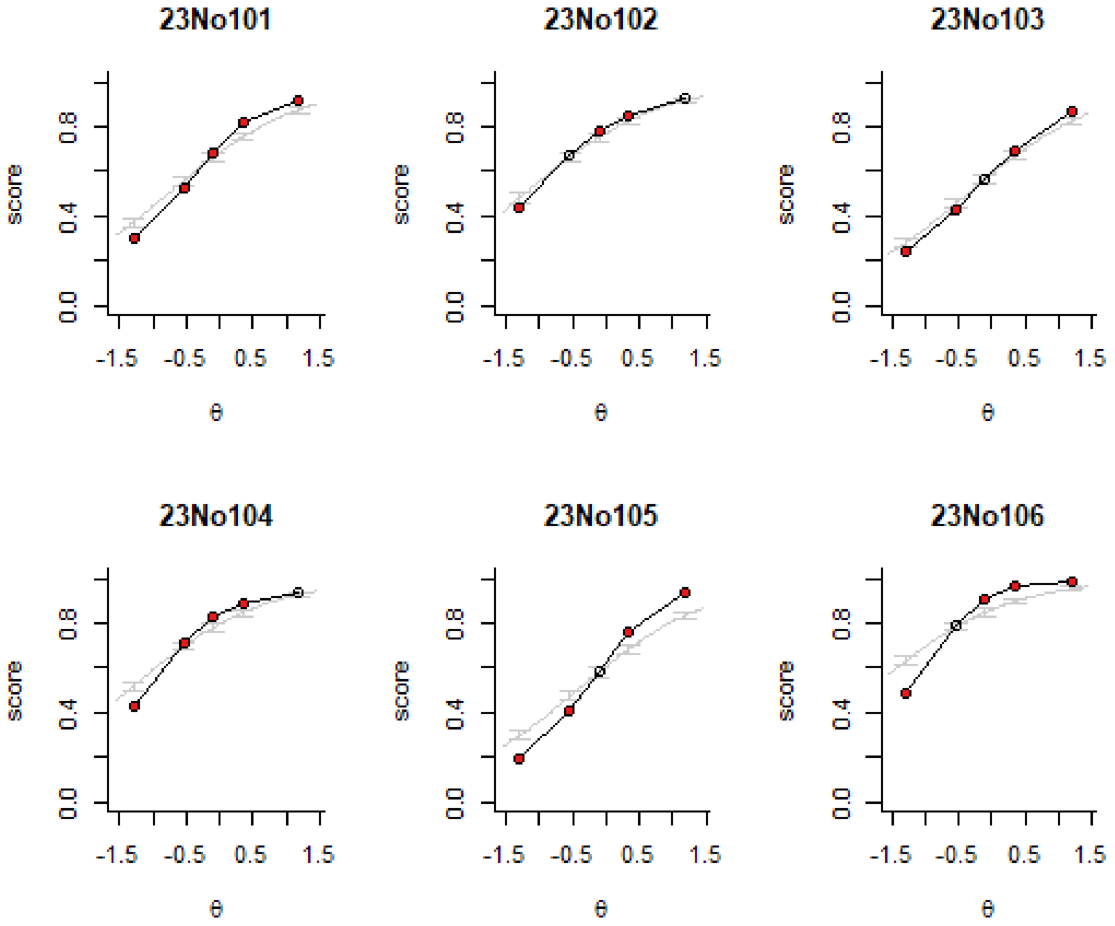


*Fig. 4. Test of fit graphs*

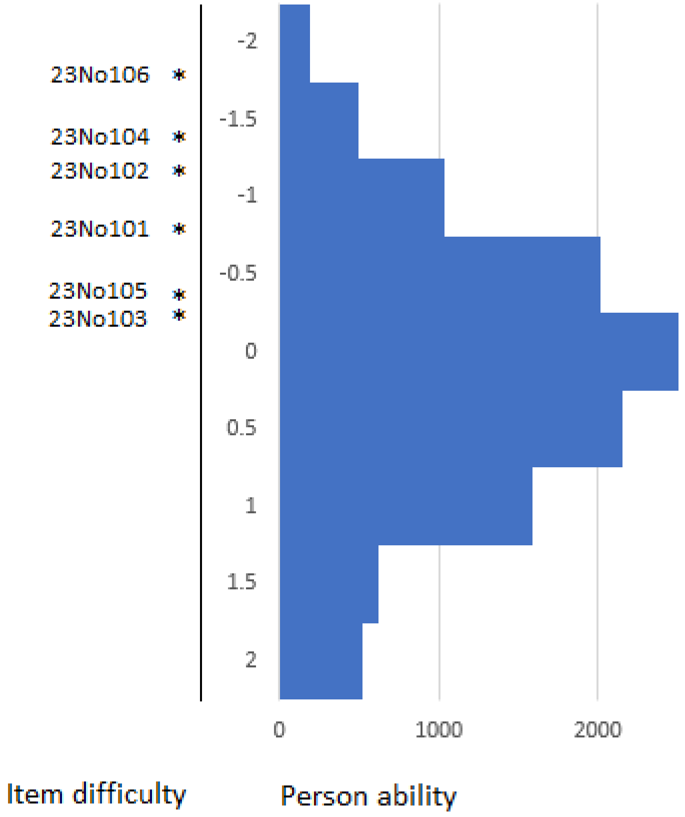


*Fig. 5. Item difficulty and person ability map*

## Conclusion

The study shows the strong empirical evidence of intra-text gap-filling items to test reading comprehension at B1 level. Unlike traditional cloze formats, intra-text gap-filling tasks allow for more controlled and meaningful deletions, targeting content words that contribute to textual cohesion and comprehension. Carefully designed intra-text gap-filling items can enhance the reading component of language proficiency tests that reflect real-world reading processes.

## References

1. Alderson, J.C. (2000). *Assessing Reading*. Cambridge University Press, Cambridge.
2. Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge University Press.
3. Taylor, W.L. (1953). Cloze procedure: a new tool for measuring read ability. *Journalism Quarterly*, 30. pp 414-438.
4. https://h5p.org/h5p/embed/503852
5. North B., Ortega A., & Sheehan S. (2015). A Core Inventory for General English. British Council
6. Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press.
7. Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume. Council of Europe Publishing. https://www.coe.int/lang-cefr
8. https://uzbmb.uz/page/test_sinovlari_formati

9. Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

10. Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

11. Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.
12. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(*3*), 297-334. doi: 10.1007/BF02310555.
13. Maris, G., Bechger, T., Koops, J., & Partchev, I. (2018) dexter: Data management and analysis of tests. URL: https://CRAN.Rproject.org/package=dexter

# Ko‘p darajali ingliz tili testlarida matn-bankli bo‘shliqlarni to‘ldirish topshiriqlari: dizayn, ishonchlilik va CEFR bilan muvofiqlik

**A.A. Abbosov**

*O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Bilim va malakalarni baholash agentligi,*
*100084, Toshkent sh., Bog‘ishamol k., 12*

**Qisqacha mazmuni.** Bo‘shliqlarni to‘ldirish topshiriqlari til bilish darajasini baholash testlarida eng ko‘p uchraydigan savol turlaridan biridir. Ular ayniqsa ko‘p darajali test formatlari uchun mos keladi, chunki ular turli murakkablik darajalarini qamrab olishi va tildan autentik foydalanishni aks ettirishi mumkin. Matn-bankli bo‘shliqlarni to‘ldirish topshiriqlari —bunda matnning o‘zi bir vaqtning o‘zida so‘z banki sifatida xizmat qiladi — yaqinda ishlab chiqilgan topshiriq turiga kiradi. Ushbu maqolada matn-bankli bo‘shliqni to‘ldirish topshiriqlarini ishlab chiqish prinsiplari, ularning ishonchliligi va Umumiy Yevropa Til Mezonlari (CEFR) bilan qanday muvofiqlashtirilishi muhokama qilinadi.

**Kalit so‘zlar:** Ko‘p darajali ingliz tili testlari, o‘qish tushunish, testlet, bo‘shliqni to‘ldirish topshiriqlari, matn-bankli bo‘shliqlarni to‘ldirish, ishonchlilik, validlik, CEFR.