The Cambridge Multiple-Choice Questions Reading Dataset

Cambridge University Press and Assessment

Contributors:

Andrew Mullooly, Cambridge University Press and Assessment

Øistein Andersen, Department of Computer Science and Technology, University of Cambridge

Luca Benedetto, Department of Computer Science and Technology, University of Cambridge

Paula Buttery, Department of Computer Science and Technology, University of Cambridge

Andrew Caines, Department of Computer Science and Technology, University of Cambridge

Mark J. F. Gales, Department of Engineering, University of Cambridge

Yasin Karatay, Cambridge University Press and Assessment

Kate Knill, Department of Engineering, University of Cambridge

Adian Liusie, Department of Engineering, University of Cambridge

Vatsal Raina, Department of Engineering, University of Cambridge

Shiva Taslimipoor, Department of Computer Science and Technology, University of Cambridge

Citing this paper: Mullooly, A., Øistein, A., Benedetto, L., Buttery, P., Caines, A., Gales, M. J. F., Karatay, Y., Knill, K., Liusie, A., Raina, V., & Taslimipoor, S. (2023). *The Cambridge multiple-choice questions reading dataset*. Cambridge University Press & Assessment.
https://doi.org/10.17863/CAM.102185

Introduction

Multiple-choice assessments involve selecting the best answer from options to complete sentences, fill in blanks, or to demonstrate comprehension of a given text (Brown & Hudson, 1998). Compared to true-false tests, multiple-choice assessments usually have less of a guessing factor, depending on the number of options. The versatility of this format makes it valuable for evaluating various learning and assessment objectives.

With the ease of marking and contribution to test reliability they offer, multiple-choice questions (MCQs) remain a popular and appropriate tool for use in large-scale language assessments. Widely considered to be an effective means of testing detailed understanding of a text, MCQs are seen as allowing for sophisticated elements of text content, such as opinion, argument, and inference, to be

tested. Questions can be written for required difficulty levels through careful text selection and crafting of options (Khalifa and Weir, 2009).

According to Rupp et al. (2006), in the MCQ format, the extent and intensity of reading comprehension can significantly vary across items. Analysing the structure and content of MC questions in a reading comprehension test typically reveals that different items assess distinct levels of reading comprehension. The process of responding to multiple-choice reading comprehension questions in standardized tests often involves problem-solving and relies on verbal reasoning (Rupp et al., 2006). In a typical multiple-choice test format, test-takers are required to grasp the instructions of the task, recognize the pertinent information within the text, process this information and ultimately make the appropriate selection based on it (Buck et al., 1997).

The primary objective of this dataset is to provide an open-access resource that allows researchers, test developers, and educators to explore the nuances of multiple-choice language assessments. Covering diverse proficiency levels and a broad range of reading comprehension topics, the dataset aims to serve as a fundamental tool for exploring language test theories, developing new assessment models, and ultimately enhancing the effectiveness of language testing.

Dataset Overview

This dataset consists of 120 4-option MCQ multi-item reading tasks. Almost half the tasks (58 in total) target Common European Framework of Reference for Languages (CEFR) B2 proficiency level, with the full set ranging between CEFR B1 and C2 level.

The CEFR is an international standard used to describe and measure language ability, ranging from A1 for beginners to C2 for those who have mastered a language. In this dataset, the CEFR levels are represented as follows: B1 corresponds to an intermediate level, B2 to an upper-intermediate level, C1 to an advanced level, and C2 to a mastery level. This diversification enables the dataset to meet a broad spectrum of research and assessment requirements.

Each task assesses comprehension of a written text. The length of text and number of questions, or items, per task varies according to the target level. All task content (input text, questions, and options) and accompanying statistics relate to pretesting versions of the tasks.

In a first for a release of this kind, for a large subset of the dataset (78 tasks in total), option-level values for two of the most important Classical Test Theory measures – facility (proportion correct) and discrimination (point biserial correlation) - are being made available. These figures can help identify not only whether a particular item is performing within expected parameters, but where within the

item the likely explanation for any under-performance may be found. Figure 1 shows the breakdown of tasks and questions/items by target level and the accompanying statistical information.

Figure 1

Target Level	**Task:** Difficulty, Facility **Item:** Difficulty, Discrimination, Facility **Option:** Discrimination, Facility	**Task:** Difficulty, Facility **Item:** Difficulty, Discrimination, Facility
C2	6 tasks (41 items)	3 tasks (20 items)
C1	12 tasks (83 items)	13 tasks (86 items)
B2	37 tasks (244 items)	21 tasks (158 items)
B1	23 tasks (115 items)	5 tasks (25 items)
Total	**78 tasks (483 items)**	**42 tasks (289 items)**

The dataset is provided in a JSON Lines (JSONL) format. JSONL presents each record as a single line of text, which simplifies the management of extensive data. Its hierarchical structure effectively represents connections among tasks, items, and options. For example, all information relating to the first task (I.e., a single reading passage and related set of questions) comes under the label:

- "id": 1

First, there is the task level information which includes:

- the "title" of the reading comprehension text
- the full reading comprehension "text" or passage
- the target CEFR "level" for the task
- task difficulty "Diff" and facility "fac" figures.

The next level of information provided relates to the "questions" (also known as items). For each question, there is given:

- the number of the question
- the question "text" i.e., the wording of the question on the paper
- the correct response or "answer" - as all questions are 4-option multiple-choice, the possible answers are always "a", "b", "c" or "d"
- question difficulty "Diff", discrimination "disc" and facility "fac" figures (see below explanation of these measures).

Finally, for each of the four options:

- the option "text" i.e., the wording of the option on the paper
- where provided, the option discrimination "disc" and facility "fac" figures. Where not provided, the values are given as "null".

Pretesting

Usually employed as part of a quality control procedure in the assessment content production process, pretesting can be defined as "The administration of items to obtain information about the performance of the items, rather than the candidates." (UCLES, 2015). Tasks are trialled on a sample of students at an appropriate proficiency level and with a balance of L1s (first language speakers), providing statistical data on the performance of each item. Such pretesting is conducted not only for content intended for live test use but also for practice test materials, helping ensure that these are sufficiently representative of the relevant exam. At pretest review, if either the quantitative statistical data or qualitative expert judgement indicate a problem with the items within a task, this task will make it no further in the process without changes being made and subsequent re-pretesting before approval for use in a live exam.

To help ensure that this data release can be of most value to the research community, the decision has been taken to include items that would have been deemed unacceptable at the pretest review stage based on their statistical performance. A brief explanation of the types of statistical analysis conducted for pretests and how this information is represented in the dataset will now be given. For a more detailed discussion of these approaches to item appraisal, see Cambridge English Language Assessment Research Notes Issue 59, upon which this article draws.

Classical Test Theory

Classical Test Theory (CTT) is one commonly adopted approach towards appraising the performance of objectively marked items, such as MCQs. Some of the analyses that CTT yields are item analysis (e.g., item facility, item discrimination, and distractor analysis), reliability estimates (e.g., internal consistency, inter-rater reliability), standard error of measurement, and various validity analyses such as criterion-related and construct validity.

Often used in parallel with the more modern Rasch Measurement Theory and Item Response Theory (IRT), CTT produces conceptually straightforward results to interpret, for which the frame of reference is limited to the group of candidates taking the test (Corrigan and Crump, 2015).

Facility

The facility value for an item shows the proportion of candidates who answered it correctly. It is a useful measure to understand how well an item is matched to a specific group of test-takers. Represented as a decimal between 0 and 1, facility is calculated by dividing the number of candidates who select the correct answer for the item by the total number of candidates sitting the test. A

question with an item facility of 0.6 for example, would therefore indicate an easier question than one with an item facility of 0.5 on the same test when taken by the same group of test-takers. Were another group of candidates to sit the test, the proportion answering any item correctly may vary significantly and, as such, an equally correct but different facility figure could be generated (Corrigan and Crump, 2015). Where there is more than one mark available for an item, the facility is calculated by dividing the mean score for the item by the maximum available score to obtain a figure between 0 and 1.

For this dataset, facility figures are given not just for the correct response but also for the three distractor options for each item. This allows us to see the relative attractiveness of the different options. Assuming all test-takers selected one of the four options, the facility figures for each when added together would equal 1 (as figures are rounded to 2 decimal places, and some test-takers may have left a question unanswered, this will not always be the exact total).

Looking at the following example from the dataset (Figure 2), where option_a is the correct response, we can see that a slight majority (55%) of candidates went for this. All three distractor options are attracting some candidates, with option_c the most popular distractor and option_b the least at only 8% of test-takers.

Figure 2

Task_ID	Q1_option_a_fac	Q1_option_b_fac	Q1_option_c_fac	Q1_option_d_fac
1	0.55	0.08	0.26	0.11

As previously mentioned, facility values can help identify potential problems with an item. If one of the distractors attracted no candidates, for instance, then it would be contributing nothing towards the task and may require re-writing. To go one step further and understand more about which candidates are selecting each option, we must also look at item discrimination.

Discrimination

To ensure an accurate measurement of ability, high-quality items must discriminate effectively between strong and weak test-takers. The measure of discrimination given for the items in this dataset is the point biserial correlation coefficient, which is a version of the standard Pearson correlation used when one of the variables being correlated is dichotomous. In this case, it is a correlation between the score each candidate receives for an item, and their overall test score, which uses data from all candidates in its calculation. Figures for the point biserial correlation coefficient can range from −1 to +1 (Corrigan and Crump, 2015).

A discrimination figure close to 0 for the correct response would show that an item is not discriminating as it should between stronger and weaker candidates. A negative figure would indicate that the item is disadvantaging stronger candidates. Again, looking at distractor statistics is helpful in diagnosing where such a problem might be situated. For example, were one of the distractors to demonstrate a positive discrimination figure, it would be important to look at the text and see if there were anything that might make this option a potential double key (i.e., a second correct response).

Looking at the values for the same example item from the dataset as for facility, we can see that it is discriminating well. Candidates scoring higher on the pretest overall are more likely to select the correct option (option_a) than are low-scoring candidates and it therefore has a positive discrimination of 0.4. The distractors (or incorrect options) show the reverse pattern, with negative discrimination figures indicating lower-scoring candidates are more likely to select them than higher-scoring candidates – this is as would be expected with a well-performing item.

Figure 3

Task_ID	Q1_option_a_disc	Q1_option_b_disc	Q1_option_c_disc	Q1_option_d_disc
1	0.4	-0.18	-0.22	-0.17

The dataset also includes facility and discrimination values for each item. It should be noted that these values are the same as those for the correct option but are given to 3 decimal places.

Item Difficulty

The dichotomous Rasch model is the simplest form of Item Response Theory (IRT) in that it features only one item parameter, item difficulty. Although more complex IRT models will typically fit a given data set better, Rasch models share mathematical properties with measurement models such as those used in the physical sciences, which makes them more appropriate for identifying anomalous behaviour since they do not simply incorporate the anomalous behaviour into the model parameters (Andrich 1988). Unlike the CTT measures discussed above, Rasch difficulty estimates are sample independent, meaning that they should not depend on the distribution of abilities of the candidates in the data set. For dichotomous items such as those in this dataset "the probability of a candidate achieving a correct response in the Rasch model is a function of the difference between the candidate's ability and the difficulty of the item in question." (Elliott and Stevenson, p17, 2015). A key contrast with

facility values therefore is that the difficulty of an item can be determined irrespective of how strong or weak the candidates are.

The difficulty figures in this dataset have been scaled. This is done so that the difficulty of most items falls between 0-100, allowing for greater ease of use when evaluating item performance and constructing tests to set levels of difficulty. Taking the same previous example as for facility and discrimination, this item has a scaled difficulty of 83.03. This would make it suitably challenging for a learner at C2 proficiency level. Looking at another item from the dataset with the same facility value, helps illustrate the limitations of facility as a measure and why we also need to look at item difficulty. After calibration has occurred, we can make a direct comparison between their difficulty values and say that Q1_Task_1 is a harder item than Q1_Task_19. This is despite the fact the two items would have been pretested in different tests and with different groups of candidates around their respective target proficiency levels. In both cases, 54.8% of the cohort got the item right, but this tells us nothing about the relative strengths of the candidates or difficulties of the items.

Figure 4

Target_Level	Task_ID	Q1_diff	Q1_fac
C2	1	83.03	0.548
B2	19	65.88	0.548

Conclusion

This report presents a comprehensive overview of a dataset consisting of 120 4-option MCQ multi-item reading tasks, designed to assess varying levels of language. Ranging from B1 to C2 levels of the CEFR framework, the dataset provides rich analytical value through the inclusion of CTT measures, such as item facility and item discrimination. Moreover, the dataset makes a pioneering contribution by providing option-level values for facility and discrimination, allowing for a nuanced understanding of item performance. The dataset is anticipated to be an invaluable resource for academics, test developers, and educational institutions, serving not only as a rich empirical base for future research but also as a practical guide for those involved in the design and evaluation of multiple-choice assessments in language testing.

References

- Andrich, D. (1988). *Rasch models for measurement* (Vol. 68). Sage.
- Brown, J.D., & Hudson, T. (1998). The alternatives in language assessment. TESOL Quarterly, 32, 653-675. https://doi.org/10.2307/3587999
- Buck, G., Tatsuoka, K. and Kostin, I. (1997), The Subskills of Reading: Rule-space Analysis of a Multiple-choice Test of Second Language Reading Comprehension. *Language Learning, 47,* 423-466. https://doi.org/10.1111/0023-8333.00016
- Corrigan, M. and Crump, P. 2015. Item analysis. Cambridge Research Notes Issue 59, UCLES 2015.
- Docherty, C. and Corkill, D. 2015. Test construction: The Cambridge English approach. Cambridge Research Notes Issue 59, UCLES 2015.
- Elliott, M. and Stevenson, L. 2015. Grading and test equating. Cambridge Research Notes Issue 59, UCLES 2015.
- Khalifa, H. and Weir, C. J. 2009. Studies in Language Testing 29: Examining Reading, UCLES 2009.
- Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language testing, 23*(4), 441-474.