An Evidence-Based Approach to Distractor Generation in Multiple-Choice Language Tests

David M. March
Cambridge Assessment English

Darren Perrett
Cambridge Assessment English

Christopher Hubbard
Cambridge Assessment English

The purpose of this project is to explore the feasibility of a new approach for producing evidence-based distractor sets. We use Common Wrong Answers (CWAs), and the associated performance data, generated by candidate responses to open gap-fill tasks, to produce distractor sets for multiple-choice gap-fill tasks based on the same texts. We then investigate whether these distractor sets are effective for use in language tests, in terms of empirical and qualitative review, and consider potential impacts on the production process for test material.

This project explores a new and innovative method of content development, and raises the possibility of a new approach to item production that can semi-autogenerate test items in shorter periods of time without affecting quality or reliability. Although the approach is specific to one task type, it is hoped that further research will expand on the applications of the approach to deliver a version that may be operationalised for use across different task types in the development of language assessments.

Keywords: English second-language, language education, language assessment, multiple-choice question, item generation, automated content production

INTRODUCTION

The process of producing tasks for English language tests is managed using principles of validity and reliability (Cambridge Assessment, 2017). In most situations, tasks are written by language experts to meet a pre-defined list of quality criteria and task specifications (validity), and performance data is collected and statistics reviewed by a panel of experts before a task is designated as fit for purpose (reliability). Two text-based tasks produced in this way, used in many language tests, are the focus of this study: the open gap-fill and the multiple-choice (MC) gap-fill. In this paper, we outline a proposal for shortening the production process for one of these tasks, based on evidence drawn from the performance data of the other.

LITERATURE REVIEW

Test Validity

A central component of any test is its validity, or construct, which includes a description of the things the test is supposed to measure. As Kane points out, any given observation can 'provide information about a variety of objects of measurement' (1982, p. 129), and it is up to the test developer to define exactly what object the test will measure. This is called the construct. A valid test will accurately measure only its given construct, and not measure anything that is not in its construct. For example, if a test construct is to measure reading in English, then a candidate's ability to read in English will be reflected in their final score, but their score will not be affected by any other factors, such as general knowledge. If a test is valid, and measures the construct reliably, then it will be possible to make inferences from test scores in order to make high-impact decisions (Capkova, Kroupova & Young, 2015, p.548; Messick, 1995, p.742; Weir, 2005, p.12).

Construct of Reading

This study considers two types of gap-fill task with different testing focuses within the construct of reading in English. In this study, the term gap-fill is used to refer to tasks or items that include a sentence or text from which a word or words have been removed in a non-random, rational, and strategic way (Khalifa & Weir, 2009, p.88). This contrasts with cloze tasks, in which words are removed from every *n*th position or another fixed ratio, or some types of automatically generated tasks where words are randomly removed. To complete a gap-fill task, candidates must repair the individual or text-based sentence(s) by replacing the missing word(s). Whether sentence-based or text-based, the items are delivered in a context which requires a candidate to read as part of completing the task. There are many types of gap-fill tasks, but the two that appear in this study are open and MC gap-fill tasks. Open gap-fill tasks require the candidate to recall a word that completes the sentence and so rely on a combination of receptive and productive elements. MC gap-fill tasks give the candidate a gap and a set of possible options to choose from, usually one key and two or three distractors, and so are more aligned to receptive knowledge. The distractors are known collectively as a distractor set. Examples of an open gap-fill task and an MC gap-fill task are given in the Appendix.

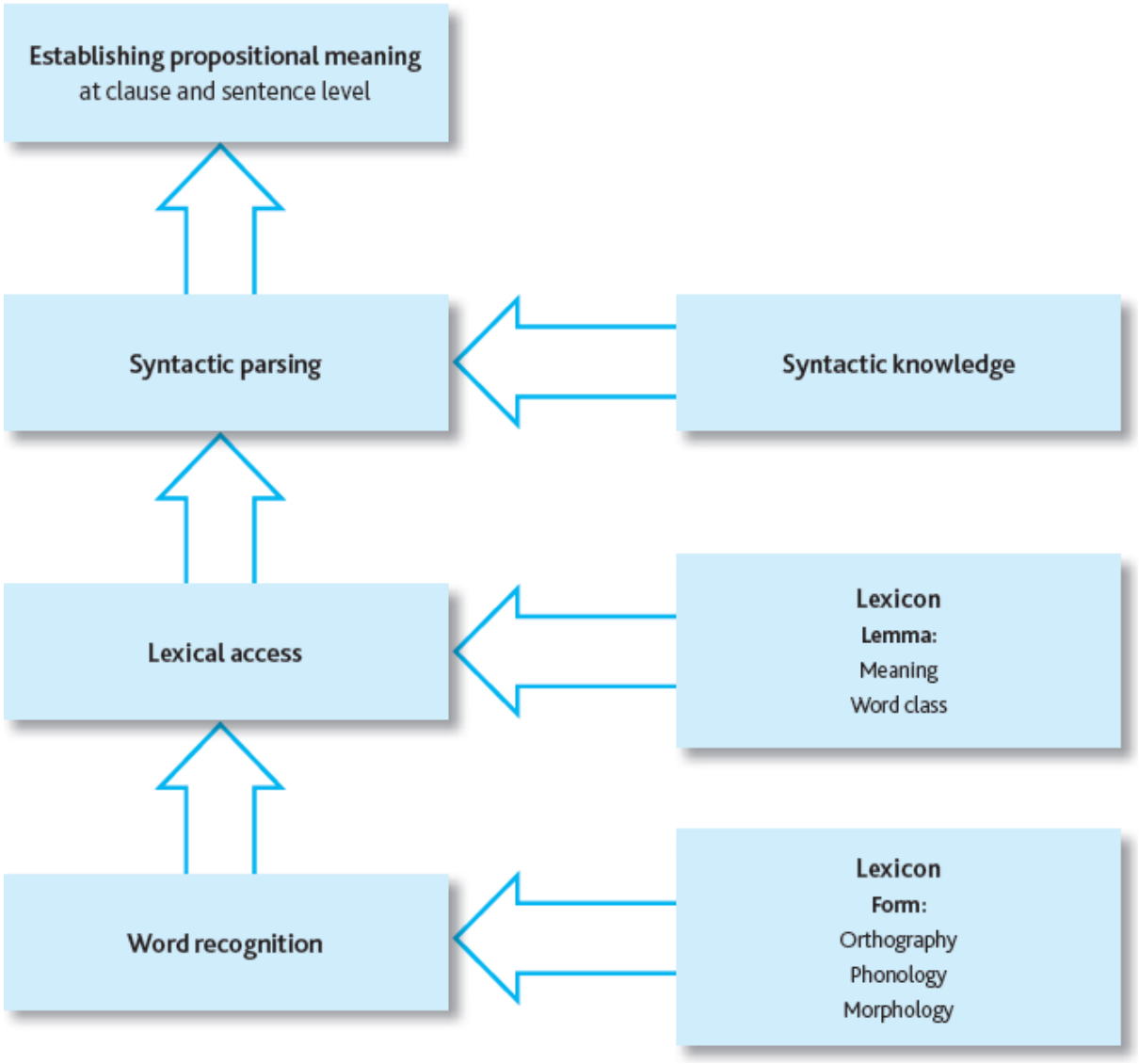
Evidence indicates that reading competence can be separated into metacognitive abilities (Table 1) and knowledge bases required (Figure 1) as highlighted in Khalifa and Weir's (2009) model of reading.

TABLE 1
MODEL OF METACOGNITIVE ABILITIES

	Careful reading	Expeditious reading
Local	Understanding sentence	Scan/search for specifics
Global	Comprehend main ideas/overall texts	Skim for gist; Search for main idea and important detail

Adapted from Khalifa and Weir, 2009, p.43

FIGURE 1
MODEL OF KNOWLEDGE BASES



Adapted from Khalifa and Weir, 2009, p.43

For both of the aforementioned gap-fill task types, the candidate's metacognitive competence is categorised as *understanding sentence* or *careful local reading*, as it has been shown that candidates completing such tasks 'rely on the immediate constituents surrounding the gap to fill it in' (Weir, 2005, p.122, citing Alderson, 1978). It is possible to test either grammar or lexis using gap-fill items (Khalifa & Weir, 2009, p.88), although it is rare for a single item to test both grammar and lexis (Fulcher 2010, p.172). According to Khalifa and Weir's (2009) model of knowledge bases in reading (Figure 1), a grammar gap-fill item is categorised as *syntactic parsing*, while a lexis item is categorised under *lexical access*, meaning the two task types have two different testing focuses. For an open gap-fill task, the target words for removal are often lexico-grammatical in nature. For MC gap-fill tasks, the target words are often lexical and work together with the distractor set to create the challenge.

Effective Distractors

A range of guidelines exist in the literature on how to write effective MC gap-fill tasks (see Fulcher, 2010, p.172; Haladyna, Downing & Rodriguez, 2002). Cambridge English *Item Writer Guidelines* stipulate that each option, including the key and distractors, should belong to a single, coherent lexico-grammatical category; i.e. if the key is 'the' then the distractor set may include other determiners such as 'a' or 'an' (Cambridge Assessment English, 2018a). The reasons for this are twofold: to prevent one distractor from standing out, and because it is assumed that words from the same category are likely to be effective distractors. Guidelines may also give measures of distractor effectiveness based on the percentage of selection by candidates. Kilgour and Tayyaba (2015, p.576) suggest that a distractor should be chosen at least 5% of the time to be considered effective, while Green (2013) suggests a higher threshold of 7%.

RESEARCH QUESTIONS

This study is comprised of two research questions (RQs):

RQ1: *How can we use Common Wrong Answers generated from items in an open gap-fill task to create distractor sets for an MC gap-fill task based on the same reading text?*

RQ1 looks at a potential process for using data collected through pretesting, including the CWAs submitted by candidates, to generate distractor sets.

RQ2: *Are the Common Wrong Answer distractor sets effective for use in language testing?*

RQ2 considers what measures are available to determine an effective distractor for MC gap-fill tasks and applies these measures to the distractor sets generated in response to RQ1.

METHODOLOGY

Distractor Generation Process

To answer RQ1, a process was developed to use data collected from candidate responses to open gap-fill tasks to produce distractor sets for MC gap-fill tasks. The resulting MC gap-fill tasks share the same texts as the original open gap-fill tasks, and the gaps are in the same places, but they have a different task type and testing focus. This process had three steps:

- task selection
- CWA report preparation
- distractor set creation.

Step 1 – Task Selection

In order to have a representative sample, tasks were selected across a range of target difficulties and batched into six groups, each aligned to one of the six levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Each task is a calibrated, multi-item, open gap-fill type task, comprised of a paragraph of text with five gaps. Each gap represents one item, and each item takes one word to complete the sentence. No sentence contained more than one item.

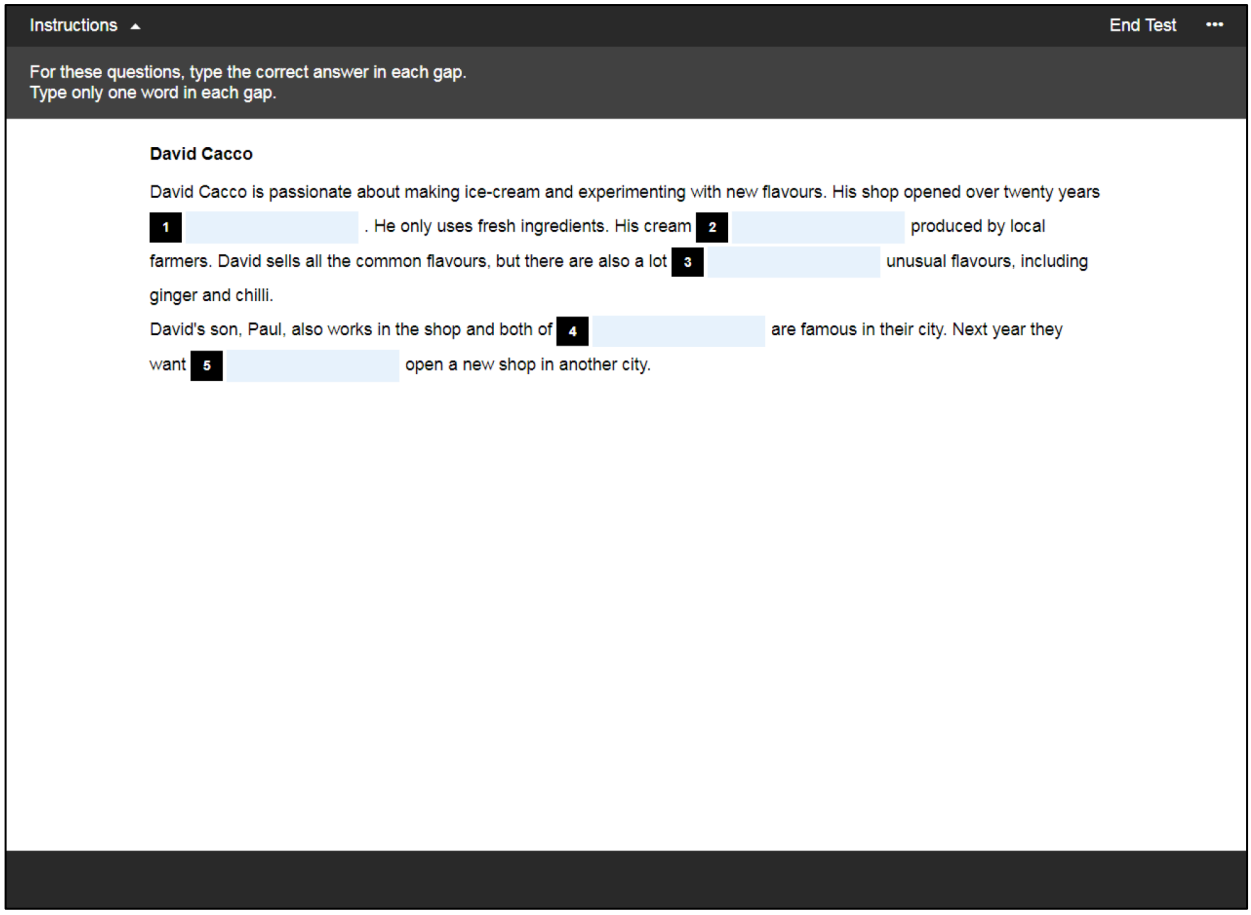
27 tasks were chosen, five from each of the levels A1 to C1, and two from C2 (Table 2). Only two tasks were selected at C2 level due to the availability of suitable tasks.

Figure 2 shows an example of an open gap-fill task at A2 level used in this study.

TABLE 2
NUMBER OF TASKS SELECTED BY CEFR LEVEL

A1	*A2*	*B1*	*B2*	*C1*	*C2*
5 tasks	5 tasks	5 tasks	5 tasks	5 tasks	2 tasks

FIGURE 2
AN A2-LEVEL OPEN GAP-FILL TASK USED IN THIS STUDY



Step 2 – CWA Report Preparation

Each task was pretested on 250 candidates, with first languages (L1s) from at least three different language families, for example, Romance, Slavonic, and Semitic. This generated a range of statistical data and a CWA report for each task. A CWA report shows each unique answer that was entered by candidates as they attempt to repair the gaps, in order of the frequency returned. To prepare the reports for generating distractor sets, each report was reviewed and amended by removing some results:

- misspelled words or words from languages other than English
- words separated by a space.

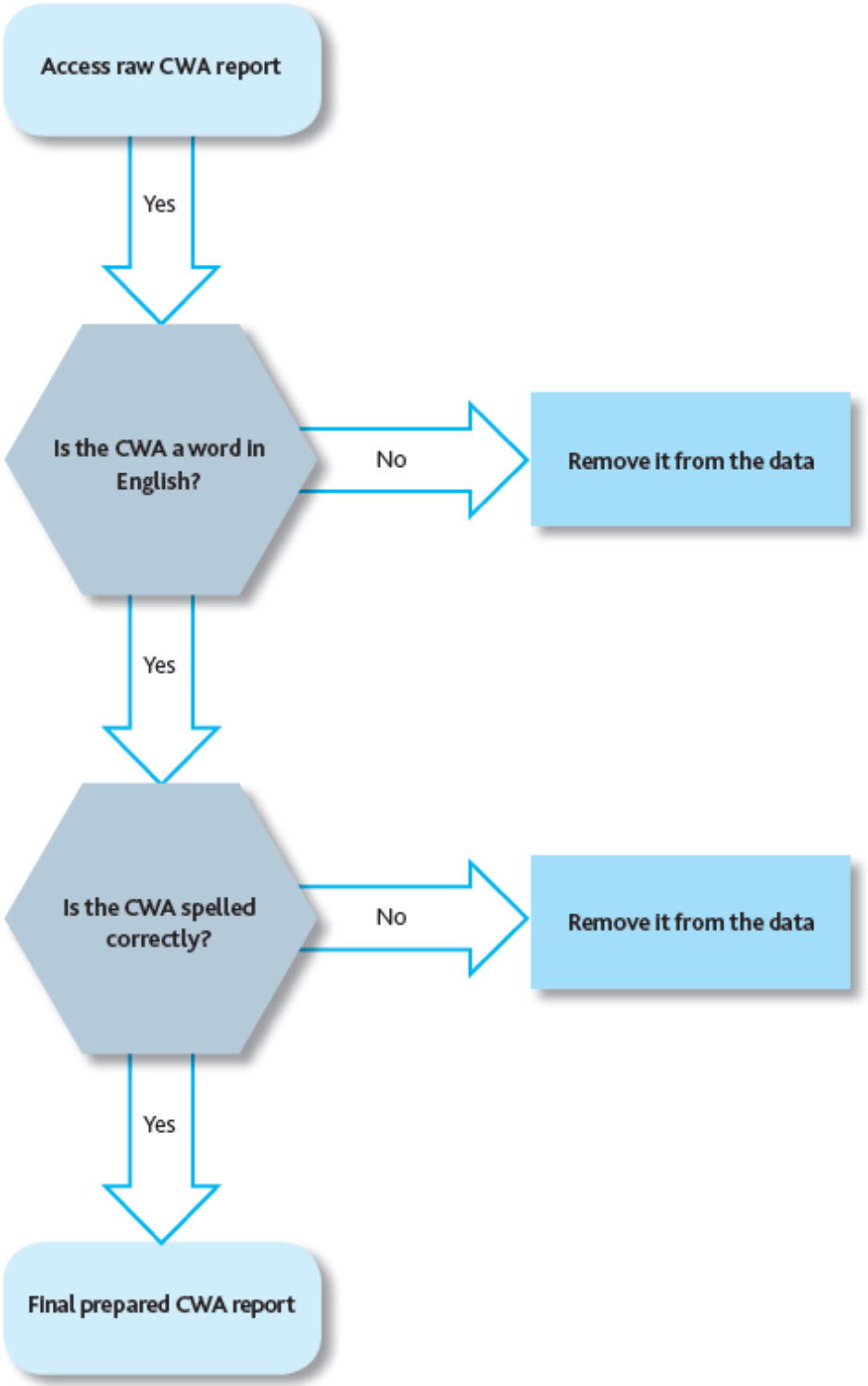
Misspelled words were removed because it would affect the content validity and face validity of high-stakes English exams to include distractors which were spelling variations of the key. Words separated by a space were removed because the task instructions allowed for one-word responses only.

The rest of the results made up the final CWA report and consisted of two groups:

- any correctly spelled English words, including proper nouns
- two words separated by an apostrophe or hyphen.

Figure 3 gives a visual model of this process.

FIGURE 3
PROCESS FOR AMENDING CWA DATA SETS



The final prepared CWA report for the sample gap-fill task shown in Figure 2 is given in Table 3. The key and five CWA columns are shown in order of the frequency returned, however many more CWAs are available in the original data set.

TABLE 3
AMENDED CWA REPORT FOR THE SAMPLE TASK

Item no.	*Key*	*1st most returned CWA*	*2nd most returned CWA*	*3rd most returned CWA*	*4th most returned CWA*	*5th most returned CWA*
1	ago	old	and	the	flavours	have
2	is	was	are	has	were	and
3	of	the	a	an	and	is
4	them	they	ice-cream	there	the	flavours
5	to	is	a	the	will	you

Step 3 – Distractor Set Creation

Distractor sets were generated by selecting the top two or three most commonly occurring CWAs for each open gap-fill task from the amended CWA reports. Two distractors were selected for tasks at A1 and A2 levels, and three distractors were selected for tasks at B1 level or higher. No item took four distractors. Table 4 shows the distractor sets that were created for the sample task in Figure 2, with the keys given in bold and percentages of selection given in parentheses.

TABLE 4
DISTRACTOR SETS FOR THE SAMPLE TASK

MC gap-fill A Item 1	*His shop opened over twenty years ……..* a. old (17.26%) b. **ago** (49.19%) c. and (1.63%)
MC gap-fill A Item 2	*His cream …… produced by local farmers.* a. **is** (33.88%) b. was (15.31%) c. are (5.54%)
MC gap-fill A Item 3	*David sells all the common flavours, but there are also a lot …… unusual flavours, including ginger and chilli.* a. a (1.63%) b. the (1.95%) c. **of** (61.56%)
MC gap-fill A Item 4	*David's son, Paul, also works in the shop and both of …… are famous in their city.* a. ice-cream (3.26%) b. **them** (28.01%) c. they (7.82%)
MC gap-fill A Item 5	*Next year they want …… open a new shop in another city.* a. a (1.95%) b. **to** (61.53%) c. is (3.26%)

A total of 135 items were produced and analysed in the study, although conclusions on outcomes also need to be drawn at task level. Table 5 shows the total number of distractors that were generated in this study for items at each CEFR level.

TABLE 5
TOTAL NUMBER OF DISTRACTORS GENERATED FOR EACH CEFR LEVEL

CEFR level	*A1*	*A2*	*B1*	*B2*	*C1*	*C2*	*Total*
Number of tasks generated	5	5	5	5	5	2	27
Number of items generated	25	25	25	25	25	10	135
Number of distractors generated	50	50	75	75	75	30	355

The resulting MC gap-fill tasks were pretested on 250 candidates with L1s from three distinct language families. Tasks were pretested at nominal Rasch difficulty values, as seen in Table 6.

TABLE 6
NOMINAL RASCH VALUES FOR TASKS AT EACH CEFR LEVEL

CEFR level	*A1*	*A2*	*B1*	*B2*	*C1*	*C2*
Nominal Rasch value	32	45	58	67	76	84

RESULTS

This section is divided into two parts. The first part reports on the performance of the distractors, which relates directly to RQ2 and to the evaluation of the effectiveness of distractors. The second part reports on the overall performance of the items and tasks as a whole, which must also be taken into account when making determinations about the effectiveness of their distractor sets.

Distractors

Percentages of Selection of Distractors

This section shows results based on the percentage of candidates who chose each distractor, known as the *percentage(s) of selection of distractors* or POSOD. Table 7 shows the number of distractors that achieved different percentages of selection. As mentioned in the section 'Effective distractors', a minimum POSOD of between 5% and 7% is regarded by some authors as a measure of the effectiveness of a distractor – effective distractors are presented in the shaded areas of Table 7. Please note that not all results sum exactly to 100 due to rounding.

TABLE 7
SUMMARY OF THE PERCENTAGES OF SELECTION OF DISTRACTORS

Percentage of selection	*Up to 1%*	*1.01% to 5%*	*5.01% to 7%*	*7.01% to 10%*	*10.01% to 25%*	*more than 25.01%*	*Total*
Number of distractors	1	41	30	43	186	54	355
Percentage of total	0.28%	11.55%	8.45%	12.11%	52.39%	15.21%	100%

Table 8 shows the descriptive statistics for the POSOD for each item grouped by CEFR level.

TABLE 8
DESCRIPTIVE STATISTICS FOR THE PERCENTAGES OF SELECTION OF DISTRACTORS

CEFR	*Mean*	*Median*	*Mode*	*S. Dev*	*Minimum*	*Maximum*	*Range*
A1	19.60	15.21	10.49	11.68	5.90	61.54	55.64
A2	17.33	17.00	21.43	10.41	2.22	49.84	47.62
B1	15.01	14.45	12.00	7.62	0.78	34.88	34.10
B2	13.97	10.93	16.12	9.89	2.64	45.42	42.78
C1	14.30	11.72	16.31	12.02	1.41	66.77	65.36
C2	16.64	13.46	N/A	10.00	2.73	38.00	35.27
All items	15.80	13.66	11.72	10.40	0.78	66.77	65.99

Non-Standard Distractors

One of the unexpected results of the study was the generation of distractors that did not fit conventional guidelines for producing distractor sets. Non-standard distractors were either:

- proper nouns, including names
- two words separated by an apostrophe or a hyphen
- homophones of the key.

Table 9 shows the non-standard distractors that emerged from this study, with the keys shown in bold and percentages of selection given in parentheses. All non-standard distractors had percentages of selection above 5%. Non-standard distractors are underlined.

TABLE 9
NON-STANDARD DISTRACTORS

MC gap-fill A Item 4	*David's son, Paul, also works in the shop and both of …… are famous in their city.* a. ice-cream (23.47%) b. **them** (48.87%) c. they (21.86%)
MC gap-fill C Item 2	*St Kilda is far enough out into the Atlantic to have …… own weather* a. his (2.12%) b. it's (13.43%) c. **its** (83.04%) d. their (1.41%)
MC gap-fill D Item 4	*…… sure the reward comes immediately after you have completed the task, and the size of the reward matches that of the task.* a. **Make** (64.37%) b. But (7.36%) c. For (10.45%) d. I'm (17.34%)
MC gap-fill E Item 4	*So the head teacher …… like a group of parents to tidy it up.* a. will (12%) b. **would** (57.45%) c. is (5.82%) d. doesn't (23.64%)
MC gap-fill F Item 1	*Immanuel Kant was an 18th century German philosopher …… work initiated dramatic changes in the fields of epistemology, metaphysics, ethics, aesthetics, and teleology.* a. who (11.6%) b. **whose** (74.74%) c. his (2.73%) d. who's (10.58%)
MC gap-fill G Item 3	*Naoko comes from …… same city.* a. to (10.8%) b. **the** (65.43%) c. Tokyo (20.68%)
MC gap-fill H Item 1	*Thank you …… your email.* a. Holly (13.41%) b. **for** (72.41%) c. Maria (12.64%)
MC gap-fill H Item 5	*Please tell …… if I make any mistakes.* a. you (14.94%) b. Maria (11.11%) c. **me** (66.67%)
MC gap-fill I Item 1	*My name's Antonio and I'm …… teacher.* a. the (10.33%) b. Spanish (28.78%) c. **a** (60.89%)
MC gap-fill J Item 3	*The researchers named this fossil the Altai dog after the mountains from …… it was recovered.* a. Siberia (13.19%) b. that (5.13%) c. there (6.96%) d. **where** (74.36%)

Repeated Word Distractors

In English, it is sometimes possible to repeat some words twice in a row. For example, 'I realise now that that is what I want …'. In this study, 'you' was the most commonly returned CWA for a gap immediately following the word 'you'. Table 10 shows an open gap-fill item and an MC gap-fill item, with the keys given in bold and percentages of return (above) and selection (below) given in parentheses. Percentages between 1% and 5% are highlighted in orange. Percentages above 5% are highlighted in turquoise. The word 'you' (underlined in Table 10) underperformed as a distractor in this item compared to the benchmark of 5–7% POSOD.

TABLE 10
PERCENTAGE OF RETURN & SELECTION FOR A REPEATED-WORD CWA AND DISTRACTOR

Open gap-fill K Item 1	*I need to ask you …… some help.* a. **for** (20.21%) b. you (14.72%) c. about (6.78%)
MC gap-fill K Item 1	*I need to ask you …… some help.* a. **for** (68.52%) b. you (2.22%) c. about (29.26%)

Items

Pretesting Success Rate

135 items and their subsequent distractor sets were generated in this study and pretested. This produced a range of statistical measures, such as scaled standard error and Infit, which were used together with judgements from assessment experts to determine each item's suitability for use. Table 11 shows the percentage of items that came within prescribed limits on statistical measures.

TABLE 11
APPLICATION OF STATISTICAL MEASURES TO MC GAP-FILL ITEMS

Total number of items generated in the study	*Number of items that came within statistical measures*	*Percentage of items that came within statistical measures*
135	128	94.81%

Difficulty Drops

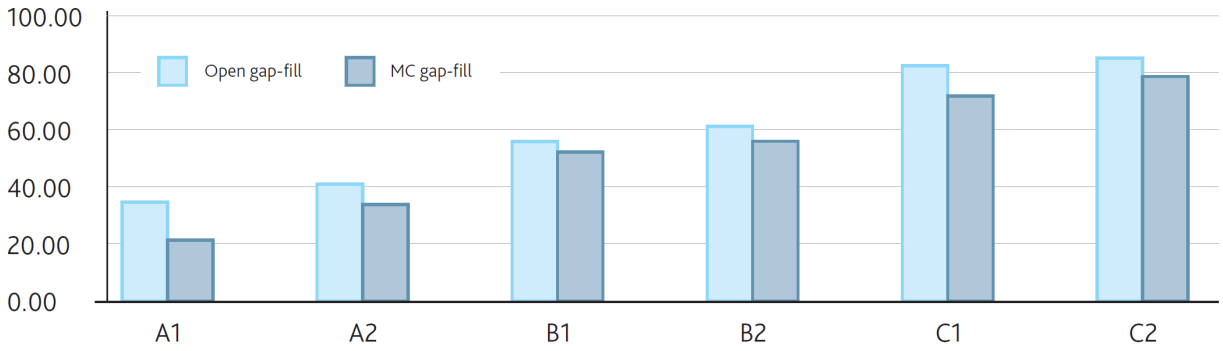
Pretesting also reported difficulties for each item, expressed as Rasch values. Table 12 shows descriptive statistics based on the observed difficulties for each item grouped by CEFR level. For comparison, the statistics for open gap-fill items are on the left, and the statistics for the corresponding MC gap-fill items are on the right.

TABLE 12
DESCRIPTOR STATISTICS FOR ITEM DIFFICULTIES BY CEFR LEVEL

	Open gap-fill items				MC gap-fill items			
CEFR level	*Mean*	*Median*	*S. Dev*	*Range*	*Mean*	*Median*	*S. Dev*	*Range*
A1	36.53	40.07	6.13	13.76	25.43	26.22	3.16	7.98
A2	44.65	43.51	3.94	10.02	34.06	32.77	3.65	8.64
B1	57.07	55.83	3.06	7.57	51.74	52.23	2.85	5.89
B2	68.78	67.67	4.09	10.47	60.65	61.60	3.04	7.80
C1	86.39	86.62	8.23	19.97	74.60	73.01	3.44	8.29
C2	93.20	93.20	2.56	3.62	82.15	82.15	7.30	10.33

Although not part of the RQs for this study, it is worth noting the changes in the mean observed difficulties between the open and MC gap-fill items. Figure 4 shows how the average Rasch difficulty for items in all CEFR levels decreased as a result of the change of task type.

FIGURE 4
MEAN OBSERVED RASH DIFFICULTIES BY CEFR LEVEL



A t-Test: Two-Sample Assuming Equal Variances comparing the average mean observed item difficulties gives a resulting P(T<=t) one-tail result of 0.043, which is significant at P=0.05 (see Table 13). This rejects the null hypothesis that there will be no difference in the difficulties of the items if the task type is changed.

TABLE 13
ANALYSIS OF FIGURE 4

t-Test: Two-Sample Assuming Equal Variances; P=0.05	0.043
Pearson's r	0.93

There is a strong correlation between the difficulties of the original and new tasks (Pearson r = .93) (see Table 13). When the open gap-fill tasks were transformed into MC gap-fill tasks, they became easier by an average of 9.5 Rasch points. The new tasks also exhibited better Infit values, which indicates that they may be discriminating better. However, Infit is a statistic that is designed for MC gap-fill tasks, and may therefore be expected to work better for this task type.

Difficulty Smoothing

The last result from the pretesting data was an effect called *difficulty smoothing*, where the item difficulty range in a task reduced in relation to outliers. The difficulty of an item that was noticeably more difficult than the other items in the task dropped by a large amount to become more closely aligned.

Tables 14 and 15 give the observed Rasch difficulties for the items in two tasks. The difficulty values for open gap-fill items are shown in column 2 and for MC gap-fill items in column 3. Column 4 shows that the difficulty values of outliers (shown in bold) dropped by much more than the difficulty values of non-outliers.

TABLE 14
DROP IN DIFFICULTY VALUES FOR ITEMS IN GAP-FILL F

Gap-fill F	*Open gap-fill difficulties*	*MC gap-fill difficulties*	*Difference*
Item 1	83.79	77.28	-6.51
Item 2	100.79	94	-6.79
Item 3	90.8	94.16	3.36
Item 4	**114.58**	**94**	**-20.58**
Item 5	85.07	77.09	-7.98
Range of values	30.79	17.07	-13.72

TABLE 15
DROP IN DIFFICULTY VALUES FOR ITEMS IN GAP-FILL L

Gap-fill L	*Open gap-fill difficulties*	*MC gap-fill difficulties*	*Difference*
Item 1	95.11	90.82	-4.29
Item 2	96.58	90.64	-5.94
Item 3	82.6	72.19	-10.41
Item 4	**129.43**	**86.88**	**-42.55**
Item 5	88.13	65.37	-22.76
Range of values	46.83	25.45	-21.38

DISCUSSION

Responding to RQ1: How Can We Use Common Wrong Answers to Create Distractor Sets?

This study outlines a process by which the data collected from the pretesting of open gap-fill items could be quickly and simply used to generate distractor sets for MC gap-fill items. This was achieved by amending raw CWA reports to remove unsuitable responses. The top two (for A1- and A2-level tasks) or three (for B1-, B2-, C1- and C2-level tasks) most commonly returned wrong answers became distractors. This process efficiently produced 27 MC gap-fill tasks, 135 items (including 135 distractor sets) and 355 distractors.

One of the most interesting findings of this study was that the process frequently produced non-standard distractors. These came in three forms.

- Proper nouns, either the name of a person (‘Maria’ and ‘Holly’) or a place (‘Tokyo’ and ‘Siberia’) or a proper adjective (‘Spanish’). It is notable that these CWAs were all words that appeared somewhere in the text for their task. Candidates found those words from other sentences and made the decision to use them as their response to repair the gap.
- Contracted words formed with an apostrophe, or compound words formed with a hyphen. Occasionally, these words were found in the text and used in the gap (‘ice-cream’). The rest were just commonly used contractions (‘I’m’ and ‘doesn’t’).
- Homophones of the correct answer (‘who’s’ instead of ‘whose’; ‘it’s’ instead of ‘its’).

These words would never have featured as distractors in MC gap-fill tasks that were written according to current guidelines for item writers. This is partly in regard to the face validity of the tests, and the presumption that such distractors will look out of place in a distractor set. However, the results of this trial indicate that the non-standard distractors were often more effective than more regular distractors, possibly providing evidence that there may be scope for including more variation in the writing of distractor sets.

Minimum Threshold for CWA Return

Evidence from our study suggests that, if CWAs are returned by at least 1% of candidates, they will produce effective distractors achieving POSOD of more than 5%. However, if CWAs are returned by less than 1% of candidates, they will produce relatively ineffective distractors achieving POSOD of only 1% to 5%.

Table 16 shows the pretesting data for task C, Item 1, as an open and MC gap-fill, with the keys given in bold and the percentages of return (above) and selection (below) given in parentheses. Percentages above 5% are highlighted in turquoise. Percentages between 1% and 5% are highlighted in orange. Percentages below 1% are highlighted in magenta.

TABLE 16
RATES OF RETURN AND SELECTION FOR TASK C, ITEM 1

Open gap-fill C Item 1	*I knew the island was a volcanic stone outcrop in the middle of the ocean – but I had no idea exactly …… to expect.* a. when (0.78%) b. how (0.78%) c. **what** (9.61%) d. where (1.95%)
MC gap-fill C Item 1	*I knew the island was a volcanic stone outcrop in the middle of the ocean – but I had no idea exactly …… to expect.* a. when (2.12%) b. how (2.83%) c. **what** (89.75%) d. where (5.3%)

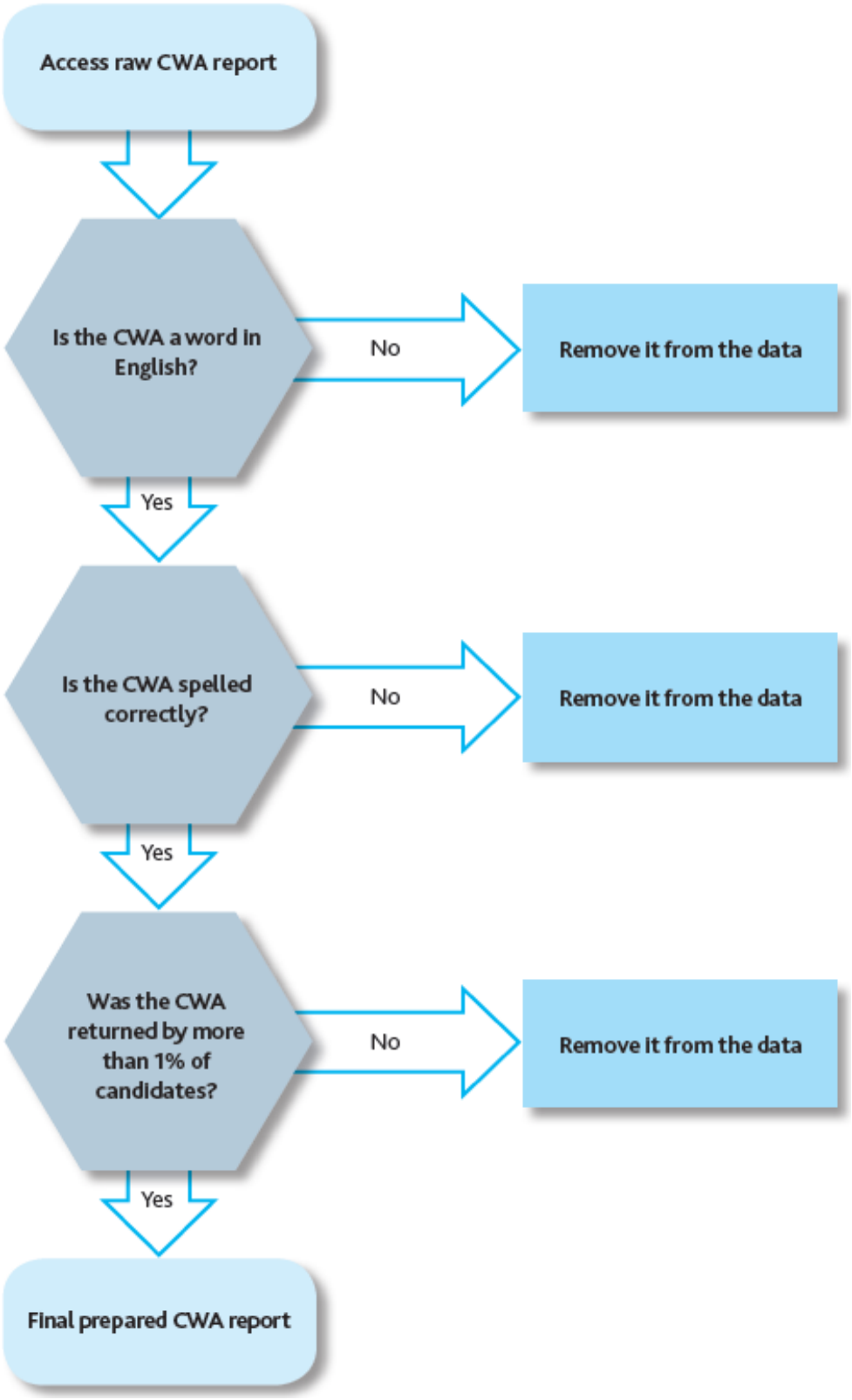
Table 17 gives the same information as Table 16, but for task C, item 2.

TABLE 17
RATES OF RETURN AND SELECTION FOR TASK M, ITEM 2

Open gap-fill M Item 2	*There are lots of ways …… enjoy life socially too.* a. for (2.86%) b. of (1.9%) c. **to** (83.49%) d. and (0.95%)
MC gap-fill M Item 2	*There are lots of ways …… enjoy life socially too.* a. for (10.09%) b. of (5.87%) c. **to** (79.81%) d. and (3.29%)

As a result of these findings, an extra step can be added to the process described in the section 'Step 2 – CWA report preparation', where any CWA returned by less than 1% of candidates is automatically removed from the data set. This is shown in Figure 5. This will have the effect of making it impossible to apply this process to some items and subsequently some tasks.

FIGURE 5
REVISED PROCESS FOR AMENDING CWA DATA SETS



Responding to RQ2: Are the Common Wrong Answer Distractor Sets Effective for Use in Language Testing?

Based on our review of the literature (see Green, 2013; Kilgour & Tayyaba, 2015), it was decided that a POSOD of at least 5% would mean that a distractor was performing effectively. A POSOD between 1% and 5% would mean that a distractor was performing sub-optimally. A POSOD below 1% would mean that a distractor was performing poorly.

Given these values, we can say that as much as 88.16% of the distractors produced in this study performed effectively. A further 11.55% of the distractors produced in this study could be said to be performing marginally effectively. Only 0.28% (1 out of 355) of the distractors performed too ineffectively to be considered usable.

Common Lexico-Grammatical Sets

67.40% of the distractor sets generated in this study did not form a single lexico-grammatical set with the key, such as nouns, verbs, prepositions, etc. An example is given in Table 18, with the key given in bold and percentages of selection given in parentheses. The underlined distractor does not share a lexico-grammatical set with the key.

TABLE 18
DISTRACTOR SET WITHOUT A COMMON LEXICO-GRAMMATICAL CATEGORY

MC gap-fill M Item 3	…… *part of a sports team or just playing for pleasure is another good way of taking part in student life.* a. Taking (22.54%) b. **Being** (61.03%) c. Be (5.4%) d. <u>The</u> (10.09%)

Fulcher (2010) argues that a common lexico-grammatical category between a key and its distractors is needed for an MC item to be effective. The evidence in this study does not support this. 92.31% of the items generated in this study that featured one or more distractors with a different lexico-grammatical category to the key passed pretesting within statistical measures, including scaled standard error and Infit.

Capkova et. al. (2015, p.552) found similar results in a study of banked gap-fill tasks. Banked gap-fill is a type of gap-fill task similar to MC gap-fill, except candidates have a bank of possible answers from which to answer a number of questions rather than three or four options with which to answer each question. Their study showed that candidates were frequently choosing words that had not been identified by item writers as the most likely distractors. This is in line with the findings in this study that the CWAs produced by candidates for the open gap-fill tasks were frequently outside of coherent lexico-grammatical sets.

Item Difficulties

Another noteworthy finding in this trial was the drop in difficulty that all of the tasks exhibited as a result of transforming from an open gap-fill task type to an MC gap-fill task type. Based on the review of the literature regarding the differences between partly production (open) and wholly comprehension (MC) task types, we can speculate that at least part of the reason for this observed difference in difficulty between the two task types may be due to the removal of the productive component. Tasks become easier when candidates do not have to recall words from their long-term memory – they only need to recognise one of the options given as the key.

One implication of the drop in difficulty may be that it will not be possible to apply this approach to open gap-fill items at pre-A1 or low A1 difficulty levels, as their difficulty may drop below the minimum meaningful level for testing.

Difficulty Smoothing

The process of transforming open gap-fill items to MC gap-fill items may also have the effect of reducing the range of item difficulties with the effect of removing outliers. Items that are noticeably harder than others in a task may become more in line with the other items.

- In Task F (see Table 14), non-outlier items dropped or raised between 3 and 8 points, while an outlier dropped by 20.58 points. The difficulty range for the original task was 30.79, but dropped by almost half to 17.07, meaning that the items were more aligned in difficulty.
- In Task L (see Table 15), non-outlier items dropped between 4 and 23 points, while an outlier dropped 42.55 points. The difficulty range for the original task was 46.83, but dropped by almost half to 25.45, meaning that the items were more aligned in difficulty.

These results suggest that the process of changing the task types from open to MC gap-fill by generating distractors from CWA reports will, in general, make the tasks easier and reduce the range of difficulties between items.

LIMITATIONS AND FURTHER RESEARCH

Larger Studies

This trial study consisted of 27 tasks, including five from each CEFR level from A1 to C1, and two from C2. A larger study comprised of many more tasks would be needed to see if the results presented here are consistent with an expanded data set. Larger studies will help to determine the true viability of using CWAs to generate distractor sets in an operational context, not only for the MC gap-fill tasks produced in this study, but for other task types and skills, such as listening.

Comparability Studies

The next step for the current study is to expand on the data available to evaluate the effectiveness of the process outlined here by engaging in comparability studies. These are necessary in order to determine which of two variables is responsible for the change in task difficulty.

The first variable under consideration is the non-standard distractors introduced by using CWAs as distractors. Normally, items written following standard guidelines are required to only include distractors which all belong to the same lexico-grammatical sets (in most cases) as the key, are only one word (no apostrophes or hyphens), are not proper nouns or adjectives, and are not homophones of the key. These guidelines were not followed in this process, and the resulting non-standard distractors could have impacted the item difficulties. The other variable being reviewed is the change in the item type from open gap-fill, which is partly a production task, to MC gap-fill, which is entirely a comprehension task. Theoretically, receptive tasks should be easier than productive tasks regardless of standard or non-standard distractor sets.

It is necessary to determine whether the use of non-standard distractors or the change of task type is responsible for the change in item difficulties that we observed in this trial. One way to do this would be to have trained item writers create new distractors, based on standard guidelines, for the tasks produced in this study, without showing them the distractors generated from CWAs. The distractors they produce could then be pretested in the same way as the distractors from this trial. If the difficulties for the items with the new, standard distractors are the same as the difficulties for the items with non-standard distractors, then we can determine that it is the change of task type, rather than the non-standard distractors, that is the cause of the change in task difficulties.

Validating Non-Standard Distractors

Our review of the literature, including the work instructions used by Cambridge English, recommends that all distractors in a set belong to the same lexico-grammatical category as the key (Cambridge Assessment English, 2018a; Fulcher, 2010). However, the results from this trial suggest that distractors which violate this guideline can function as distractors. In fact, some of the most effective distractors generated from this study deviated from what would be produced according to current guidelines. Clearly, more research is needed in order to potentially validate the use of non-standard distractors, such as we have seen here, for use in high-stakes language assessment.

CONCLUSION

This study explored a new process for using data from pretesting to generate distractor sets for new language test items. This process took the CWAs returned by candidates as they attempted to repair the gaps in open gap-fill tasks, and used the most commonly returned wrong answers as distractors in newly formed MC gap-fill tasks. This process was both simple and quick to implement. The results of this study included a new distractor set for each of the items, which were then pretested.

Overall, the results were very positive. 88.16% of the distractors generated in this study were selected by at least 5% of pretest candidates. This number increases to 99.71% when those distractors that were selected by between 1% and 5% of pretest candidates are added. 94.81% of all items generated in this study passed pretesting. However, some unexpected results were produced, including a small number of non-standard distractors, and a significant drop in the difficulties of the items with some evidence of difficulty smoothing.

We believe there is a case to be made for the continued exploration of evidence-based item generation, including automated evidence-based item generation, to increase the capacity of awarding bodies to generate high-quality test materials with reduced cost and time requirements. Such processes also take advantage of the resources being increasingly made available through the collection of pretest data. More research is needed to replicate and validate the results of this trial, including larger studies and comparability experiments.

ACKNOWLEDGEMENT

This article was originally published in issue 72 of Cambridge Assessment English *Research Notes* (2019), cambridgeenglish.org/researchnotes.

REFERENCES

Cambridge Assessment. (2017). *The Cambridge Approach to Assessment.* Cambridge.

Cambridge Assessment English. (2018a). *Item Writer Guidelines.* Cambridge.

Cambridge Assessment English. (2018b). *Linguaskill Reading and Listening sample test.* Retrieved from www.cambridgeenglish.org/exams-and-tests/linguaskill/preparation/

Capkova, H., Kroupova, J., & Young, K. (2015). An analysis of gap-fill items in achievement tests. *Procedia – Social and Behavioural Sciences*, *192*, 547–553.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Green, R. (2013). *Statistical Analyses for Language Testers.* Basingstoke: Palgrave Macmillan.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 209–334.

Kane, M.T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*(2), 125–160.

Khalifa, H., & Weir, C.J. (2009). Examining Reading: Research and Practice in Assessing Second Language Reading. *Studies in Language Testing*, *29.* Cambridge: UCLES/Cambridge University Press.

Kilgour, J.M., & Tayyaba, S. (2015). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education*, *12*, 571–585.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-based Approach.* Basingstoke: Palgrave Macmillan.

APPENDIX

Examples of an open gap-fill task and an MC gap-fill task (Cambridge Assessment English, 2018b).

FIGURE 6
AN EXAMPLE OF AN OPEN GAP-FILL TASK

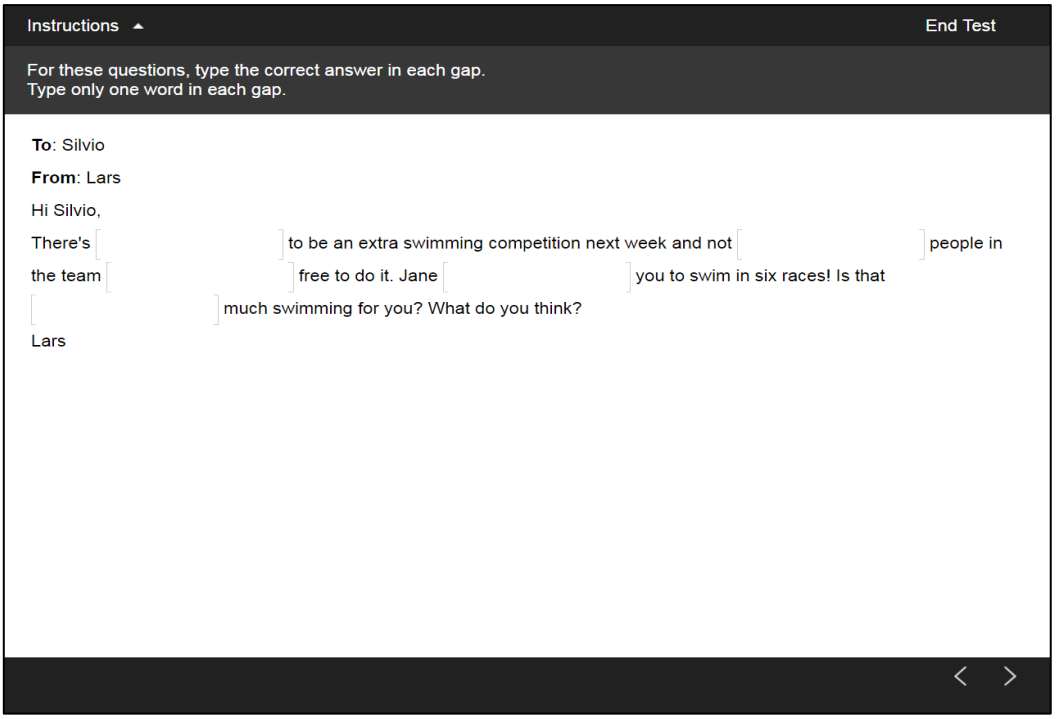


FIGURE 7
AN EXAMPLE OF AN MC GAP-FILL TASK