Multiple Alternative Choice test items (MACs) deliver more comprehensive assessment information than traditional 4-option MC test items

Robert Michael Foster
University of Wolverhampton,
Wulfruna Street, Wolverhampton, WV1 1LY,
Research Institute in Information and Language Processing
R.M.Foster@wlv.ac.uk

Abstract

The experiment described in this paper investigates the effectiveness of a format of Multiple Choice test item that is new to the featured domain. The response data for candidates answering Multiple Alternative Choice (MAC) test items are compared to response data for 4-option Multiple Choice test items covering the same content. The results show how analysis of responses to the test items in the MAC test item format have identified candidate knowledge gaps more precisely than response analysis for the items in the traditional 4-option Multiple Choice format.

1. Introduction

Multiple Choice test items are used by our company to confirm knowledge of documents from the company's corpus of policy documents. The Multiple Choice test items are delivered in the form of pre and post tests associated with training courses and field audits. The stored responses from these tests allow us to demonstrate that training has been received by staff in accordance with requirements stated in UK Legislation.

The most important outcome in the featured domain is that candidates confirm that they have correctly assimilated the knowledge presented in their training. Therefore the format of test item used must produce response data that allows unambiguous identification of candidate mis understanding. The Multiple Alternative Choice (MAC) test item was identified during a review of the literature as an item format which might deliver this result more effectively than the traditional 4-option Multiple Choice test item format.

This hypothesis was tested using two parallel sets of test items incorporated into the Multiple Choice assessment routine delivered to new entrants to the company. The results clearly show that using MACs leads to response data which provide more complete information about candidate knowledge than the use of the more traditional 4-option Multiple Choice test item format.

2. Background

2.1. Assessments during apprentice induction

The Policy Library for Western Power Distribution consists of a small number of general policy statements (POLs) and a large number of Standard Techniques (STs). The STs are intended to give precise instructions for the correct methods that the staff must apply when they are carrying out work on behalf of the company. Several of the Standard Techniques contain requirements for staff to complete sequences of MCQ test items (called 'CBT tests'). For example: 'ST:OS7D – Relating to Audits of Operational field staff' states that

> *"3.1 All Senior Authorised and Authorised Persons who hold an authorisation for HV Operational Work (11SW, 33SW, 66SW, 132SW and restricted variations) shall complete an annual CBT test to the satisfaction of an Examining Officer qualified to examine for that authorisation."*

In 2007 a review was carried out of the costs of producing and maintaining an item bank of 130 Multiple Choice test items that were first created in 1991. The study demonstrated that item production and maintenance is particularly time consuming. A follow up research project has therefore been set up to analyse and improve the process for creating and maintaining the Multiple Choice assessment tests used by this company.

The most promising approach identified so far has been the use of a Multiple Choice test item generator [2],[3] to generate Multiple Choice test items and post edit them to form the item bank. It has been reported in [2]. [3] that generating Multiple Choice test items using the generator can speed up the process by 4 times without compromising the quality of the output. However preliminary experiments applying the system [2], [3] to the company policy library delivered no usable items, so significant improvements in performance are necessary before the system could be adopted.

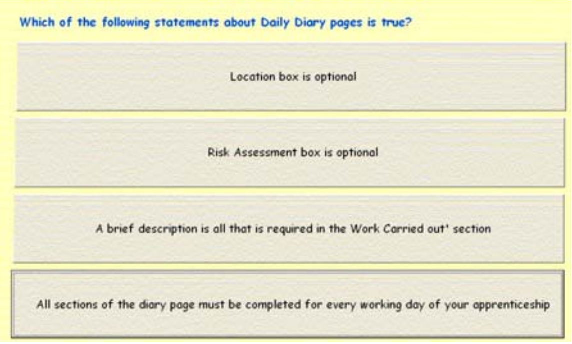
This paper describes important preparatory work that supplements improvements to the Multiple Choice test item generator software. We want to establish, with a combination of literary review and experimental evidence, the most appropriate format of Multiple Choice test item for use within the featured domain. The method for evaluating the decisions produced by this review must demonstrate that best practice has been adopted during design and delivery of the assessments.

2.2. Item format choices during item creation

A Review of the literature in relation to Multiple Choice test item formats revealed a helpful summary of guidelines in the 'Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment' [1].

> *"Although the number of guideline citations ranged considerably among textbooks, nearly all guidelines received unanimous endorsements when they were cited. These unanimously endorsed guidelines are 1–8, 11–16, 19–24, and 27–30."*

Application of these guidelines allows item designers to have confidence that best practice has been applied when their items comply with the guidelines as stated in Haladyna's revised Taxonomy. However the above statement deliberately excludes several guidelines from the list of those receiving unanimous endorsement. The guideline concerning the best item format is one of those under dispute. This led item designers for this domain to ask whether these guidelines could be adapted to ensure they met the particular requirements of their domain.



Which of the following statements about Daily Diary pages is true?

Location box is optional

Risk Assessment box is optional

A brief description is all that is required in the Work Carried out' section

All sections of the diary page must be completed for every working day of your apprenticeship

Figure 1 – 4MC01 appearance to candidates

Further investigation revealed that the lack of clarity over the most appropriate test item format often led to time being wasted debating the most appropriate format. The follow up literature review identified several studies which cited problems with the 4-option Multiple Choice item format [4],[5] which was the format most often chosen by default.

In particular, as has been stated, certainty about a candidate's state of knowledge during response data analysis [5] is the most important outcome in this domain from asking candidates to take Multiple Choice test items. This problem is well illustrated by considering an example. The appearance of test item 4MC01 is shown in Figure 1. The corresponding Response data are presented in Table 1.

Table 1 – 4MC01 Response data analysis

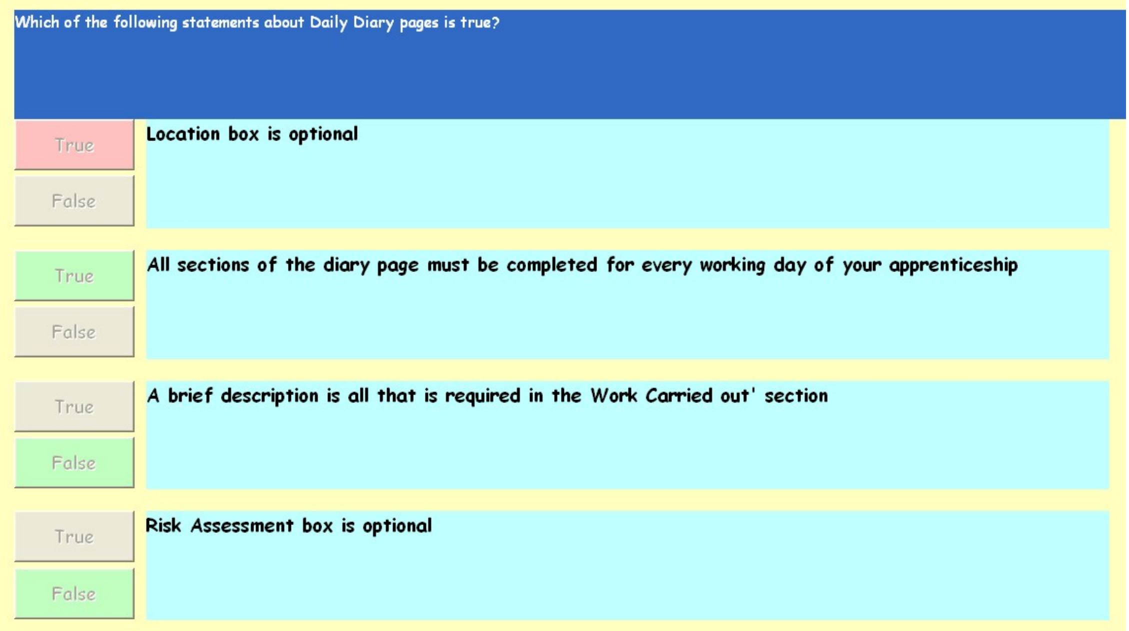
Item ID	**Number of Responses to each option in a 4-option MC test item**	
	Knowledge Confirmed as Present	***Knowledge Confirmed as Absent***
4MC01a	13	1
4MC01b	Unknown N1 = 1 to 14	14 – N1
4MC01c	Unknown N2 = 1 to 14	14 – N2
4MC01d	Unknown N3 = 1 to 14	14 – N3

Although this analysis shows that 13 of the candidates confirmed they know the correct response and 1 candidate confirmed that he / she does not know the correct response. The table also highlights the lack of information we have about the 14 candidates' state of knowledge about the other three 'distractor' options that were available to them as they selected their response.

2.3. Multiple True False (MTF) / Multiple Alternate Choice (MAC) item format

The Multiple Alternate Choice (MAC) test item format is described in Haladyna's Review of item-writing guidelines [1] as a more general version of the Multiple True False (MTF) test item format in that the two responses available are not restricted to 'True' / 'False'. They could be 'Agree' / 'Disagree' or 'Yes'/'No' etc.

Figure 2 displays a screen print showing the format of a MAC test item that presents the same content as is presented in 4MC01. The response data for the second group of 14 candidates who were presented with the same content but in this alternative format, is shown in Table 2. The response data analysis presented in table 2 shows that there are three separate instances of error among these candidates. These gaps in knowledge would not have been identified if these candidates had given their response to 4MC01 because that format does not require them to give a response in relation to each of the individual test item stems.



Which of the following statements about Daily Diary pages is true?

Location box is optional

All sections of the diary page must be completed for every working day of your apprenticeship

A brief description is all that is required in the Work Carried out' section

Risk Assessment box is optional

Figure 2 – MAC appearance to candidates

In some cases it might be vitally important for the summative assessment to identify such knowledge gaps. An unidentified lack of knowledge can lead to candidates causing a serious incident.

Table 2 – MAC01 Response data analysis

Item ID	**Number of Responses to each option in a MAC**	
	Knowledge about 'correctness' confirmed as Present	***Knowledge about 'correctness' confirmed to be absent***
MAC01a	13	1
MAC01b	13	1
MAC01c	13	1
MAC01d	14	0

Some quantitative research [6] in the Western Power Distribution domain that compares the performance of MAC formatted test items to 4-option MC formatted test items has recently been conducted and published. The mean response time when comparing 4-option MC formatted test items with MAC formatted test items shows a significant reduction (over 39 seconds) and the mean change in Item Difficulty (0.07) was small. These results indicate that the effectiveness of the test routine was not significantly affected by the use of MAC formatted test items, thanks to the small change in mean item difficulty value. Also the time taken to work through the test routine was significantly reduced by the use of MAC formatted test items. This was interpreted as providing evidence in support of the adoption of MAC formatted test items into the item-writing guidelines for Western Power Distribution.

3. Experiment

3.1. Hypothesis

The hypothesis is that the use of the MAC test item format will allow clearer identification of misunderstandings of training course content than the use of the 4MC test item format.

3.2. Method

The hypothesis was tested by delivering both 4-option Multiple Choice test items and Multiple Alternate Choice test items to new entrants to the featured UK company. Two parallel experiments were conducted:

Group A (14 candidates) took the assessment routine containing 2 4MC format items (4MC01, 4MC03) each with four options (these are listed as 4MC01a, 4MC01b and 4MC03a, 4MC03b.. etc) and 2 MAC format items (MAC02, MAC04) each containing four Alternative Choice options (listed MAC02a, MAC02b.. etc and MAC04a, MAC04b… etc).

Meanwhile, group B (the other 14 candidates) were presented with 2 4MC format items 4MC02, 4MC04) which tested equivalent content to MAC02, MAC04) and 2 MAC format items (MAC01, MAC03) which tested equivalent content to items (4MC01, 4MC03).

3.3. Evaluation

Evidence in support of the hypothesis will have been produced if the response analysis of MAC test items provides a more precise indication of the candidate's knowledge gap than the response analysis of 4-option MC test items.

4. Results

For each test item a record was made of the option selected by each candidate. A total of 28 sets of responses for the featured test items for each experiment was retained for analysis consisting of 14 sets of responses for 4MC01, 4MC03, MAC02 and MAC04 and 14 sets of responses for 4MC02, 4MC04 and MAC01 and MAC03.

There was no perceived need to exclude any response data since all candidates gave responses to all the test items presented to them and all tests were conducted under controlled conditions. A summary of the results is provided in Table 3.

Table 3 - Results summary from response data analysis of MAC test items vs 4MC test items

Item Category	Number of Responses in each category		
	Knowledge about 'correctness' confirmed as Present	Knowledge about 'correctness' confirmed to be absent	Uncertain
4MC test items	53	3	168
MAC test items	205	19	0

5. Conclusions and Recommendations

This study has conclusively demonstrated that the Multiple Alternate Choice (MAC) format of Multiple Choice assessment test item provides a much clearer picture of candidate knowledge than the much more widely used 4-option Multiple Choice item type (4MC). The full criteria for acceptance of the hypothesis have been met in this experiment, and it is very hard to see how future experiments could demonstrate any other conclusion. It is therefore the recommendation of this study that the MAC type Multiple Choice test item format is to be preferred over the more traditional 4-option Multiple Choice test item format in the design of assessment items in this domain.

6. Future Work

Further experiments are planned to investigate comparisons of more sophisticated characteristics of these item types. The characteristics that will be investigated include Item Difficulty and Item-Total Bi-serial correlation coefficient [4],[5]. Nevertheless, the results from this study provide a very strong case for the adoption of the MAC format of test item into this domain. The study also provides strong support for the suitability of the MAC type item for future work in the development of the Multiple Choice test item generator software [2],[3]. The first steps in this work have already been published at recent conferences [7],[8].

7. Acknowledgements

This work has only been possible thanks to the support of staff from Western Power Distribution. In particular, I wish to thank my Manager Steve Loveridge and Trainers Nick Virgin and Jake Ramsden.

I also want to record a general 'Thank you' to the members of staff at RIILP, Wolverhampton University, UK and in particular I want to thank my supervisors Dr Le An Ha and Professor Ruslan Mitkov for their continued guidance and support.

8. References

[1] 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(3), 309–334

[2] Mitkov, R., and L. A. Ha. (2003). *"Computer-Aided Generation of Multiple-Choice Tests."* In Proceedings of HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing, pp. 17-22. Edmonton, Canada.

[3] Mitkov, R., L. A. Ha, and N. Karamanis. (2006). *"A computer-aided environment for generating multiple-choice test items."* Natural Language Engineering 12(2): 177-194.

[4] Gronlund, N. (1982). *"Constructing achievement tests."* New York: Prentice-Hall Inc.

[5] Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). *"Psychometric Characteristics and Response Times for Content-Parallel Extended-Matching and One-Best-Answer Items in Relation to Number of Options"*: Academic Medicine Vol 81(10,Suppl) Oct 2006, S52-S55.

[6] Foster, R.M. (2010) *"Adapting multiple-choice item-writing guidelines to an industrial context."* In proceedings from ICEIS 2010, Funchal, Madeira.

[7] Foster, R.M. 2010 *"Improve the output from a MCQ test item generator using Statistical NLP"* ICALT 2010, Tunisia

[8] Foster, R.M. (2009) *"Improving the Output from Software that Generates Multiple Choice Question (MCQ) Test Items Automatically using Controlled Rhetorical Structure Theory"* RANLP 2009, Borovets – Student Conference