Evaluating an application of the CAREGen MCQ Creation Methodology

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

In 2011 senior managers at a UK electricity distribution company faced a training and assessment challenge that would have been very difficult to address using conventional methods. This paper describes the challenge they faced and the reasons for using the Construed Antonym Realization Exercise (CARE) generation methodology to create Multiple Alternative Choice (MAC) test items as part of the solution. An innovative design of an evaluation exercise is then described which includes some novel methods for calculating the efficacy of a MCQ generation methodology. The success of this intervention provides further evidence to support planned developments of the methodology and continued use of the proposed measures of efficacy.

1. Introduction

At the start of 2011, a UK electricity distribution company who have helped this research team, employed approximately 2,200 staff to operate and maintain the Electricity Distribution Network in a clearly identifiable area of the UK. About 350 of these staff were authorized 'Sanction for test – Issue'. In April 2011 the company expanded to take responsibility for some new areas of the country. Over 3,500 staff were taken on and about 500 of these staff also required 'Sanction for Test – Issue' if they were to continue to operate and maintain the Network in their area of the country. While the rules for safe operation of the network were being revised, it was realised that a significant difference existed in the understanding between the two groups concerning 'Extended use of Sanction for Test'.

Under the UK Health and Safety at work, etc Act 1974 (HASWA)[1], participants in work activities in the UK have a Duty of Care for the health and safety of themselves and their colleagues. Sequences of MCQ test items [2] are regularly used within the company, as part of the process for helping safety and training managers to meet their Duty of Care under HASWA[1]. A recent study showed that changing the format of the MCQs to MAC (Multiple Alternative Choice) format, increases significantly the accuracy with which knowledge gaps can be identified [3]. Another study has provided a specification for the CARE generation methodology for creating MAC test items [4]. It was therefore decided to apply the CARE generation methodology to provide the new staff with some formative assessment [5] MCQs to allow them to develop a consistent understanding of 'Extended Use of Sanction for Test'. This project was also chosen as a context for designing a formalized evaluation procedure since the project would involve the use of new MCQs alongside existing MCQs. The CARE generation methodology is described elsewhere [4] so the focus of this paper is upon the new evaluation procedure design as it was applied to evaluating the MCQ generation methodology.

2. Background

Until 2011, evaluations of the performance of MCQ test item generation methodologies, reported opinion surveys from a small group of Subject Matter Experts [4],[7],[8]. However, in 2012 it was decided that these evaluation exercises needed to comply with the full requirements for scientific rigor [9]. This presented a problem because, as has already been mentioned, one of the items of legislation that the company is required to comply with (the Health and safety at Work act (1974) [1]) specifically requires companies to exercise a duty of care for employees which includes providing a safe place of work. No group of staff should be induced to forgo the benefits of the training purely for the purposes of an evaluation exercise.

3. Domain Specific definitions

The 'efficacy' this evaluation exercise seeks to measure is the efficacy of changes that have been made to the proposed MCQ generation methodology. However, it is the efficacy of the MCQ test items themselves that is available for measurement. Therefore this evaluation exercise will make a comparison between:

a) MCQ efficacy measurements for items generated using the **existing** MCQ generation methodology and

b) MCQ efficacy measurements for items generated using the **proposed** MCQ generation methodology.

Relevant staff, trainers and safety engineers were asked to present in a single sentence an accurate definition of their aims in using MCQ test items. The response was as follows:

> *“MCQs are effective when they demonstrate that staff have assimilated the intended knowledge by achieving a 100% score and when they persuade staff to repeat the formative assessment routine until they achieve a 100% score.”*

This suggests that some industrial users of MCQ test items might have different requirements from their MCQ test items when compared to other groups of users, from educational assessment for example, who are only interested in an MCQ test item’s ability to discriminate between those candidates who know the correct response and those who do not. A new domain specific definition of MCQ efficacy was therefore proposed and this has been described by adding some domain specific terms to the domain’s terminology. In the following table, the new terms are presented in accordance with recommendations for a domain specific terminology as defined by Le An Ha [12]:

Table 1: New Domain Specific Terminology items

MCQGen Method:	The proposed methodology for creating MCQs
KACE:	Knowledge Acquisition Confirmation Event – An instance when a user clicks the correct response button within a MCQ
MCQ generation Efficiency	The number of items that were included in the routine divided by the total number of ‘man-hours’ used generate all the MCQs that were considered for inclusion.
KACE Efficacy	The proportion of KACEs to the total number of responses made by those who eventually score 100%
MCQ Attractiveness Efficacy:	The proportion of a specified target audience who repeated the routine until they achieved a 100% score

There follows a description of an evaluation exercise which shows how the requirements for a Double Blind, Randomised, Controlled Trial (DB-RCT) [9] were observed and provides a hypothesis and a set of results that use a traditional definition of MCQ efficiency and the domain specific definitions of MCQ efficacy as defined above.

4. Evaluation process

In accordance with training design principles as defined by Robert Mager, a Specific Objective [6] for the routine was specified.

> *“All staff who joined the company after April 2011 and are authorised ‘Sanction for test – Issue’ will achieve a 100% score when answering a set of Multiple Choice Questions to demonstrate that they share with existing staff an understanding of how the company applies Approved Procedure 5.2 - Extended use of Sanction for Test.”*

The designers of the formative assessment routine that was to consist of MCQs, chose the MAC test item format to meet this objective in accordance with the recommendations of a recent study which proved that MAC-formatted test items identify knowledge gaps more effectively during summative assessments and provide more immediately useful feedback during formative assessments[3]. The hypothesis for this evaluation was therefore defined as follows

Table 2: Hypotheses for the MCQGen evaluation exercise

“When compared to traditional methods of MCQ creation...	
Hypothesis 1:	…. a greater or equivalent MCQ generation efficiency can be achieved with the MCQGen Method.”
Hypothesis 2:	….a greater or equivalent KACE efficacy can be achieved by applying the MCQGen Method.”
Hypothesis 3:	…. an equivalent MCQ attractiveness efficacy can be achieved by applying the MCQGen Method.”

These hypotheses draw from the intrinsic features of well-designed MAC formatted test items, in that they provide useful feedback when they are instantly marked incorrect and evidence of successful learning when marked correct. Therefore in contrast to other recommendations [2],[10],[11], there is no need to measure how many research subjects answered incorrectly and how many answered correctly since these numbers are irrelevant to both KACE efficacy and MCQ attractiveness as defined above.

In order to comply with the requirements for a DB-RCT as defined in the CONSORT 2010 statement [9], significant variables must be defined:

Table 3: Variables for the MCQGen evaluation exercise

Independent variable:	Method of MCQ creation
Dependant Variables:	MCQ generation Efficiency, KACE Efficacy, MCQ Attractiveness Efficacy
Controlled Variables:	Instructions for preparing for and completing the assessment, Content of items included in the assessment
Randomising variables:	Attitudinal, Educational and Cultural background of the research subjects Choice as to whether or not to repeat the formative assessment until 100% is achieved Technical and Physical characteristics of the Computer Based Assessment environment

The ‘double blind’ requirement is met since subjects are not aware that they are involved in a trial of MCQ creation methods, so there is no question of them knowing whether or not they are included in the control group. Randomised selection of members of the control group is achieved through a series of non-significant environmental conditions, such as interruptions by colleagues or customers, failures of the technology, lack of learner determination, lack of current knowledge in the learner of the content, lack of available lookup resources etc. Control is achieved by fixing the instructions for completion of the formative assessment and fixing the content. Thus all significant variables are either measured, controlled or identified as randomising variables.

5. Results

Given the above definitions, measures of efficacy were calculated as shown in this table:

Table 4: Calculation methods for MCQGen evaluation exercise

MCQ Generation Efficiency	Number of items selected / No of man-hours creating, selecting and testing
KACE efficacy	Number of correct responses made by 100% scorers / Total number of responses made by 100% scorers
MCQ attractiveness efficacy	Number of incorrect responses made by 100% scorers / Total number of responses

The results from this evaluation continue to be gathered but the results on 30th May 2012 are presented in this table:

Table 5: Measurements of Efficacy of the ExUSFT MCQ routine

	MCQGen methodology NOT used (ie MCQs existed before assessment was requested)	MCQGen methodology WAS used
Total Man Hours	3.5	2
Items selected (considered)	4 (20)	4 (5)
Total Responses	2532	10444
Total Responses by 100% scorers	1510	6240
Correct Responses by 100% scorers	1360	4928
MCQ Generation Efficiency	**1.1**	**2**
KACE Efficacy	**90.0%**	**78.9%**
MCQ Attractiveness Efficacy	**5.9%**	**12.6%**

The results table shows us that the proposed methodology has a higher percentage value of MCQ attractiveness efficacy and a lower value of KACE efficacy.

6. Discussion

The two new measures of efficacy of MAC formatted MCQ test items that have emerged as a consequence of this work demonstrate several benefits of MACs, in addition to the improved precision in the identification of learner knowledge gaps that was identified in previous research[3]. In addition to producing statistics that can be used to measure the efficacy of a MCQ generation methodology, the KACE efficacy and MCQ attractiveness measures can also be calculated for individual MCQ test items, thereby providing new response analysis possibilities. This is particularly beneficial for this company, for whom the MAC is now the preferred MCQ test item format[3].

The reason for the restriction of these efficacy measures to comparing MCQ generation methodologies that produce MAC formatted test items, is that these measures rely upon two intrinsic features of MAC test items:

a) Inherent feedback is delivered by a MAC-formatted MCQs when an incorrect response is highlighted in red.

b) Inherent confirmation of knowledge acquisition has been delivered when a subject achieves a 100% score in a routine consisting of MAC test items.

The justification for accepting the limitation of this design of evaluation process that can only evaluate MCQ generation methodologies that produce MAC-formatted MCQ test items in this company is based upon some recent experiments which established MAC-formatted MCQs as the preferred MCQ format in this company[3]. These experiments showed that changing to the MAC (Multiple Alternative Choice) format improves the chance of identifying knowledge gaps and improves the quality of feedback during formative assessments. This resulted in the company specifying the MAC formatted test item as the preferred MCQ test item format.

In spite of the restriction that the KACE efficacy measure and the MCQ attractiveness measure can only be used to compare the performance of MCQ generation methodologies that generate MAC formatted MCQs, it can be seen from the above that these two new efficacy measures provide many benefits to the MCQ routine designer.

7. Conclusions

The requirement from this evaluation process was that it must measure the efficiency and efficacy of the a proposed MCQ generation methodology. The efficacy measures we have devised are more convenient to calculate in the featured domain than the efficacy measures used in other evaluation exercises [10][11]. This is possible as a consequence of certain properties of the Multiple Alternative Choice item type that is generated by the methodology. This convenience is also desirable so as to avoid any suggestion of concerns about members of the control (placebo) group.

When software has been written which is capable of automatically generating MCQ test items of this format from source documents, then it will be possible for the same evaluation process and the same measures of efficiency and efficacy to be applied in order to evaluate them.

It would appear that in addition to the established benefits of MAC formatted MCQs identifying knowledge gaps more accurately [3], initial indications from this study are that MAC formatted items also have a higher MCQ attractiveness efficacy, while 4-option Multiple Choice formatted MCQs have a higher KACE efficacy. However more extensive studies are required before such a claim could be made with any degree of certainty.

8. Recommendations

A general recommendation arising from this study is that designers of evaluation processes might benefit from defining efficacy measures in their hypotheses that accurately reflect the requirements of interested parties, as opposed to automatically using the measures that are defined in the academic literature.

A more specific recommendation is that when a new MCQ generation methodology is being evaluated it is important for evaluators to decide which features of the MCQs are more important in their domain before deciding whether one methodology is to be preferred over another, because the measurement of two efficacy measures in this evaluation process design has highlighted the possibility of either methodology being judged to have a higher efficacy, depending upon which efficacy measure (MCQ attractiveness efficacy or KACE efficacy) is judged to be more significant for a particular application.

9. References

[1] UK Legislation – Health and Safety at Work, etc Act (http://www.hse.gov.uk/legislation/hswa.htm) 1974

[2] 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

[3] Foster, R.M., "Multiple Alternative Choice test items (MACs) deliver more comprehensive assessment information than traditional 4-option MC test items" – London International Conference on Education 2010

[4] Foster, R.M., 'Creating a High Voltage Cable-Jointing knowledge check using the CARE generation methodology ' – London International Conference on Education 2011

[5] Crooks, T., "The Validity of Formative Assessments". British Educational Research Association Annual Conference, University of Leeds, September 13-15 2001

[6] Mager, R., "Preparing Instructional Objectives (2nd Edition)". Belmont, CA: Lake Publishing Co. 1975

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

[8] Foster, R.M., "Automatic generation of Multiple Alternate Choice (MAC) test item stems by applying Causal Relation Explication, Addition and Manipulation (CREAM) to pre-processed source documents" – London International Conference on Education 2010

[9] Schultz K.F., Altman D.G., Moher D.; for the CONSORT Group (2010). "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials" Br Med J 340:c332. DOI:10.1136/bmj.c332

[10] Gronlund, N., Constructing achievement tests. New York: Prentice-Hall Inc. 1982

[11] Swanson D.B., Holtzman, K.Z.,Allbee K.,Clauser, B.E., "Psychometric Characteristics and Response Times for Content-Parallel Extended Matching and One-Best-Answer Items in Relation to Number of Options." 2006

[12] L.A. Ha, L.A., "Advances in automatic terminology processing: Methodology and application in Focus" – PhD Thesis (http://clg.wlv.ac.uk/papers/ha-thesis.pdf) 2007