Evaluating a new MCQ Generation methodology (CAREGen) in the UK Electricity Distribution Industry

Robert Michael Foster
University of Wolverhampton, UK

Abstract

A recent survey demonstrated an increased usage of Multiple Choice Questions (MCQs) within a UK company. The report recommended a wider application of MCQs for knowledge assessments. Specifically, several contexts for Formative and Summative assessment were identified as suitable for MCQs, alongside the existing widespread use of MCQs for knowledge Refresher assessments. This result has led to increased interest in the process of MCQ-Creation, and in the evaluation measures used to compare the performance of MCQs created by different MCQ-Creation teams who each use their own MCQ-Creation methodology.

This article proposes some new context-specific efficacy measures for comparing the performance of MCQs and illustrates their application using a case study which combines features of both formative and refresher assessment contexts. The recommendation is that different assessment contexts should use different evaluation measures.

An additional observation is that the different evaluative requirements of summative and refresher assessments are more likely to be understood by MCQ designers if they were to attend MCQ-Creation workshops. These training events would allow them to better understand the different features of evaluation measures that can be applied to measure performance variations between alternative MCQ-Creation methodologies.

1. Introduction

UK legislation requires companies to provide training that covers their rules and procedures for Health and Safety [1]. In the company featured in this research, the evidence for achievement of the clearly defined objectives of their training [2].[3] comes from practical assessments carried out at the end of a course of training, and through on site field checks and interviews. However, an increasingly popular form of evidence for satisfactory achievement of knowledge acquisition objectives, is a high score in a test routine that covers the required knowledge. Such routines are often comprised of Multiple Choice Question test items (MCQs) [4].

Interest in robust evaluation methods [5] for alternative MCQ creation processes has increased alongside the increase in MCQ usage and so a survey of usage in different assessment contexts was carried out [6].

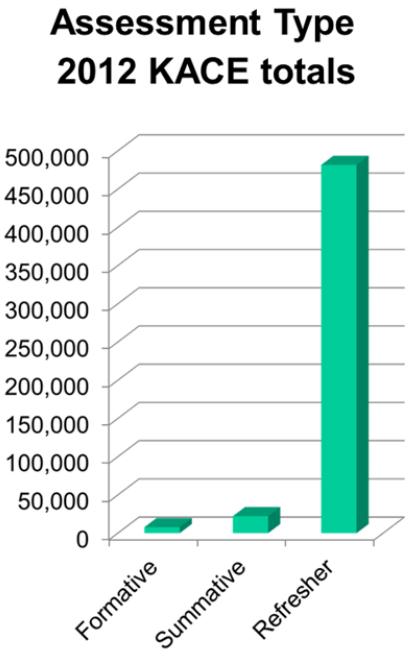

Figure 1. 2012 KACE counts by assessment type category

Figure 1 is taken from the results of this survey and it shows the proportions of MCQs used by the company in 2012 in each of three assessment categories: formative [7], summative [8] and refresher assessments. The measure used in the survey is a 'KACE' (Knowledge Acquisition Confirmation Event). The KACE count is simply a count of the number of times a user has clicked the correct response to a Multiple Choice Question during an assessment.

It is clear from Figure 1 that most of this company's MCQs are used in Refresher assessment contexts. However existing methods for evaluating MCQ efficacy are designed for Summative assessment contexts. For instance, the discriminative evaluation measures that are applied to MCQs in Medical examinations in the United States [9], seek to discover which MCQ test items are best at discriminating between those students who will pass the assessment and those who will fail. This is not

the focus of designers of MCQs that will be used in the Refresher assessment context.

Another illustration of the inadequacy of existing evaluation measures for the refresher assessment context is provided by considering the evaluations of MCQ test item generation methodologies that were applied in recent experiments that included tests of various possible methods for automatically generating MCQ test items. Evaluation during these experiments was based upon simple counts of the number of manually created and the number of generated MCQ test items that were selected by a small group of Subject Matter Experts [10],[11],[12].

Neither of these evaluation measures provides useful measurements or feedback for MCQ designers who create MCQs that will be used in refresher assessment contexts, because those designers have different aims from the designers of MCQs that will be used within summative assessments. New MCQ efficacy measures have therefore been proposed and tested to address this shortfall and they are presented in this article.

The structure of the article is as follows. In section 2 there is a description of the case study that is used in the other sections to illustrate the new MCQ evaluation measures. Section 3 provides a review of existing MCQ evaluation measures and considers the issues surrounding their application to to the refresher assessment context. The presentation shows why MCQ evaluation measures for a refresher context require different measures from the evaluation measures that are usually applied to MCQs used for summative assessment. Section 4 gives a list of properties of new evaluation measures that would be more effective in measuring the features of MCQs in a refresher assessment context and Section 5 defines some suitable measures.

The rest of the article describes an experiment in which the proposed evaluation measures were applied, followed by the results, conclusions and recommendations arising from this work.

2. Case Study – Extended use of SFT

At the start of 2011, the featured UK electricity distribution company 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 [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 [4] are regularly used within the company, as part of the process for helping safety and training managers to meet their Duty of Care under the Act [1].

A recent study showed that changing the format of the MCQs from the traditional 4-option, single response 4MC format to the Multiple Alternative Choice (MAC) format, increases significantly the accuracy with which knowledge gaps can be identified [13]. Another study had provided a specification for the CARE generation methodology for creating MAC test items [14].

It was therefore decided to apply the CARE generation methodology to provide the new staff with some MCQs to allow them to develop a consistent understanding of 'Extended Use of Sanction for Test'. This project was 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 [14] so the focus of this paper is upon the new design for evaluation measures as it was applied to evaluating the MCQ generation methodology. Before introducing the new evaluation measures, we look briefly at the traditional methods for evaluating MCQs.

3. Traditional MCQ Evaluation methods

An extensive literature surrounds the production of Multiple Choice Question test items (MCQs) whose validity can be established [2].[3].[4]. The foundation for this framework is Bloom's Taxonomy of learning and within this framework, the validity of MCQs is established using well defined methods for creating instructional objectives [2].[3].

The works of Norman Gronlund,[2] and Robert Mager [3] provide MCQ designers with detailed guidance in the writing of instructional objectives. Objectives that have been written according to these standards provide the foundation upon which Thomas Haladyna [4] builds his case for establishing the validity of MCQ test items and these in turn provide the foundations for the report from the CRESST [5] group.

Researchers such as Swanson et al.[9] have then applied these evaluation measures to show how successful each of their MCQ test items has been in discriminating between those test candidates who will score above the pass mark and those who will not achieve the pass mark.

However, throughout Gronlund's text [2] the emphasis is upon writing objectives that will be tested in a summative assessment context:

> *"Formative assessment is a matter of monitoring the process and products of student learning, and providing feedback about the particular type of performance. Summative assessment is then a matter of using an assessment instrument that best measures that particular type of performance"*

By contrast the proposition of this article is that although it is important that a MCQ test item can discriminate between those who will pass and those who will fail the routine as a whole, this is only part of the function of a MCQ test item. Other important features of a MCQ test item in terms of its creation include:

(a) The item's ability to persuade candidates to repeat the routine until they have scored 100%.
(b) The ease with which critics of the item can establish the validity of the item against the accepted authoritative texts.
(c) The time taken to create the MCQ test item must be as small as possible.

The proposal in this article is that making a change to established measures and/or introducing new measures into the process for evaluating MCQ test item efficacy will result in improvements in manual MCQ-Creation. This in turn is also likely to result in improved results from software systems that are designed to generate MCQ test items automatically[15].[16].[17].

4. Requirements from New MCQ Evaluation Measures

The context of this research is MCQ-Creation and MCQ-Generation. Therefore the overall aim of the evaluation is to measure the effect of changes to the MCQ generation methodology through which the MCQs are created. However, it is measures of the efficacy of the MCQ test items themselves that are currently defined in the literature, and which can be measured in the experiments we design. Therefore the following definition of the evaluation task has been produced:

"An evaluation measure is required that will allow us to make comparisons 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 from the company including trainers and safety advisors, were asked to present an accurate definition of their aims in using MCQ test items in the context of the case study described above. 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 assessment contexts that use MCQ test items have different requirements from their MCQ test items compared to other assessment contexts. For example, summative assessment contexts require MCQ test items to discriminate between those candidates who know the correct response and those who do not. By contrast MCQs that are used in a formative assessment context are required to persuade candidates to repeat the MCQ until they achieve a 100% score.

Some new definitions of MCQ efficacy have therefore been proposed and these have been described using some domain specific terms to the domain's terminology. The definitions of the new terms are provided in the next section.

5. New MCQ Evaluation measures

The terminology that is used to define the new MCQ evaluation measures are presented in accordance with recommendations for presenting a domain specific terminology using context as defined by Le An Ha [18]:

Table 1. New Domain Specific Terminology items

Term label	**Definition within the terminology**
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 MCQ items 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) [16] 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.

6. Experiment

6.1. Hypothesis

In accordance with training design principles as defined by Robert Mager, a Specific Objective [3] 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 MCQs within the refresher assessment routine would be fulfilling both formative and summative roles. The routine consisted of eight MCQ test items.

Four of the items within the routine were in the 4MC format [4]. They had existed before the routine was specified and had not been created using the MCQ-Creation methodology which was being evaluated. The other four items were created using the CAREGen methodology which produces MCQ test items in the MAC format [4].[13].

Application of the CAREGen methodology and the use of the MAC format of MCQ was in line with the recommendations of a recent study [13] which proved that MAC-formatted test items identify knowledge gaps more effectively, in accordance with the implied requirements of summative assessments and provide more immediately useful feedback which is an implied requirement of formative assessments. 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."

6.2. Method

The 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],[4],[9], 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 [19], 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:	• Attitudinal, Educational and Cultural background of the research subjects • Content of items included in the assessment
Randomising variables:	• 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 refresher assessment and fixing the content. Thus all significant variables are either measured, controlled or identified as randomising variables

7. Results

Given the above definitions, measures of efficacy were calculated as follows:

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

7.1 MCQ Generation Efficiency

The MCQ Generation efficiency calculations presented at the London International Conference on Education 2012 are independent of the number of responses to the routine that have been measured. Therefore these values remain the same as those presented at the conference. Table 5 shows the values when the CAREGen methodology was NOT used.

Table 5. MCQ Generation Efficiency for MCQ test items that were NOT created using the CAREGen methodology (ie before the MCQ routine was specified)

Description	**Measured Value**
Total Man hours	3.5 hours
Items selected	4
Items Considered	20
MCQ Generation Efficiency	**1.1**

Table 6 gives the values when the CAREGen methodology WAS used.

Table 6. MCQ Generation Efficiency for MCQ test items which WERE created using the MCQGen methodology

Description	**Measured Value**
Total Man hours	2 hours
Items selected	4
Items Considered	5
MCQ Generation Efficiency	**2**

7.2 KACE Efficacy

The values for KACE efficacy that were reported at the London International Conference on Education in 2012 were the most up to date at that time. However, since then there have been many more completions of the routine and so now the complete results and subsequent calculations can be presented.

First we see the data and results for the calculation of KACE Efficacy for the items within the routine which were produced WITHOUT the application of the CAREGen methodology:

Table 7. KACE Efficacy for MCQ test items that were NOT created using the CAREGen methodology

Description	**Measured Value**
Total Responses by 100% scorers	2,793
Correct Responses by 100% scorers	2,470
KACE Efficacy	**88.4%**

Then we see the data and results from the calculation of KACE Efficacy for the items within the routine which were produced using the CAREGen methodology:

Table 8. KACE Efficacy for MCQ test items that WERE created using the CAREGen methodology

Description	**Measured Value**
Total Responses by 100% scorers	11,604
Correct Responses by 100% scorers	8,660
KACE Efficacy	**74.6%**

7.3 MCQ attractiveness efficacy

MCQ attractiveness efficacy values quoted in 2012 have also changed as a result of more completions of the assessment routine. The full

results and subsequent calculations are presented below.

Table 9. MCQ Attractiveness Efficacy for MCQ test items that were NOT created using the CAREGen methodology

Description	**Measured Value**
Total Responses	4,960
Total Incorrect Responses by 100% scorers	323
MCQ Attractiveness Efficacy	**6.5%**

Then we see the data and results from the calculation of MCQ Attractiveness efficacy for the items within the routine which were produced using the CAREGen methodology:

Table 10. MCQ Attractiveness Efficacy for MCQ test items that WERE created using the CAREGen methodology

Description	**Measured Value**
Total Responses	20,592
Total Incorrect Responses by 100% scorers	2,944
MCQ Attractiveness Efficacy	**14.3%**

These results still show that the proposed methodology (CAREgen) has a higher percentage value of MCQ attractiveness efficacy and a lower value of KACE efficacy.

8. Discussion

The two new measures of efficacy of MCQ test items intended for use within a refresher assessment 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 [13].

The approach also introduces statistics that can be used to measure and compare the efficacy of a MCQ generation methodology for different assessment contexts. 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 [13].

The reason for the suggestion that these efficacy measures be used to compare MCQ generation methodologies that produce MAC formatted test items is that these measures focus 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 after the item has been marked.

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 the 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 recommendation that the KACE efficacy measure and the MCQ attractiveness measure should be used to compare the performance of MCQ generation methodologies that generate MAC formatted MCQs, it can be seen from the results from the experiment that useful comparisons can still be derived from applying the measures to items that do not have the MAC format.

9. Conclusions

The requirement from this evaluation process was that it must measure the efficiency and efficacy of a proposed MCQ generation methodology. A suitable MCQ efficacy measure has been proposed for the Refresher assessment category, which takes into account both formative and summative features of MCQs. The efficacy measure is more convenient to calculate in the featured domain than the efficacy measures used in other evaluation exercises [10][11]. The design was inspired by certain properties of the Multiple Alternative Choice item type that is generated by the methodology.

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 [13], MAC formatted items can also yield a higher MCQ attractiveness efficacy which is in line with the one of the primary objectives of designers of MCQs that will be used in a formative assessment context, while

4-option Multiple Choice formatted MCQs have a higher KACE efficacy. More extensive studies are required before such a claim could be made with any degree of certainty.

10. Recommendations

10.1. Recommendations from the experiment

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 literature or are included in the MCQ test item delivery software they have been offered.

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 chosen assessment context 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.

The recommended next steps are that interested researchers could:

a) Define distinct evaluation measures for MCQ efficacy for different categories of assessment (eg formative, summative and refresher)
b) Calculate values for these measures using existing MCQ response data
c) Use appropriate regression techniques to combine the calculated values in appropriate proportions to produce an evaluation measure that was therefore adapted to the specific content-sub-domain and assessment context.

10.2. Recommendations for MCQ-Creation

Another recommendation is that designers of MCQ test items that are intended for use in assessment contexts which do not fall neatly into the category of summative or formative assessments, might benefit from attending training or workshops that explore the issues surrounding the creation of MCQ test items. One example is the MCQ-Creation workshop that was delivered at the London International Conference on Education 2012.

Figure 2. Logo for the MCQ-Creation workshop

A customised version of the workshop might be appropriate which focusses upon the issues of MCQ test item evaluation.

11. References

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

[2] Gronlund, N., Brookhart, S. – Writing Instructional Objectives' – Pearson 1991

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

[4] 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

[5] Behrens, J.T., Mislevy, R.J., DiCerbo K.E., Levy R. (2010) "CRESST REPORT 778 - An Evidence Centred Design for Learning and Assessment in the Digital World" - The National Center for Research on Evaluation, Standards, and Student Testing Graduate School of Education & Information Sciences

[6] Foster, R.M. (2013) – 'Improve MCQ-Creation Skills To Support Corporate Learning' – Ireland International Conference on Education 2013

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

[8] Glickman, C.D., Gordon, S.P. & Ross-Gordon, J.M. (2009) – Summative Assessment

[9] 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

[10] 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

[11] 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

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

[13] 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

[14] Foster, R.M. (2012a) "Using a new MCQ Generation methodology (CAREGen) in the UK Electricity Distribution Industry" In 'International Journal of Digital Society (IJDS), Volume 3, Issues 1 and 2, March/June 2012 643 - 651

[15] 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.

[16] 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.

[17] Brown J.C., Frishkoff G.A. Eskenazi M., 2005 "Automatic Question Generation for Vocabulary Assessment" Processing (HLT/EMNLP), pages 819–826, Vancouver, October 2005. © 2005 Association for Computational Linguistics

[18] 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

[19] 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

