ADAPTING MULTIPLE-CHOICE ITEM-WRITING GUIDELINES TO AN INDUSTRIAL CONTEXT

Robert Michael Foster
Research Institute in Information and Language Processing, University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, United Kingdom
r.m.foster@wlv.ac.uk

Keywords: Electricity Distribution Industry, Apprentices, Knowledge Assessment, Test Routine Creation, Multiple Choice Question, MCQ, Multiple Alternative, MAC.

Abstract: This paper proposes a guideline for writing MCQ items for our domain which promotes the use of Multiple Alternative Choice (MAC) items. This guideline is derived from one of the guidelines from the Taxonomy of item-writing guidelines reviewed by Haladyna et al, 2002. The new guideline is tested by delivering two sets of MCQ test items to a representative sample of candidates from the domain. One set of items complies with the proposed guideline and the other set of items does not. Evaluation of the relative effectiveness of the items in the experiment is achieved using established methods of item response analysis. The experiment shows that the new guideline is more applicable to the featured domain than the original guideline.

1 INTRODUCTION

Multiple Choice Question (MCQ) test items (Haladyna, T.M., Downing, S.M., Rodriguez, M.C., 2002) are used by the UK company featured in this paper to confirm knowledge of documents from the company's corpus of policy documents. The MCQ 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 the company to demonstrate that staff have been trained in accordance with requirements stated in UK Legislation (UK Legislation Health and Safety at work, etc Act 1974).

The Revised Taxonomy of Multiple-Choice (MC) Item-Writing Guidelines (Haladyna et al, 2002) is published within a document entitled:

'A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment creation'

The emphasis in this title upon 'Classroom Assessment' highlights one of several differences between the aims of the Taxonomy and the requirement for guidance in the creation of assessments that can be used by UK company featured in this study. The current study therefore examines how one of the 31 guidelines contained within the Taxonomy can be adapted in order to provide more focussed guidance for those creating test items for use in this domain.

Taxonomy Guideline 9 about which Haladyna has highlighted disagreement in the literature, is tested using two sets of MCQ test items that have been built into the MCQ assessment routine delivered to a group of new entrants to the company. One set of items complies with the proposed guideline and the other set of items does not. The intention is to show through analysis of the responses to these items that the adapted version of the Multiple-Choice (MC) Item-Writing Guideline 9 should be applied when MCQ test items are being created for the featured domain.

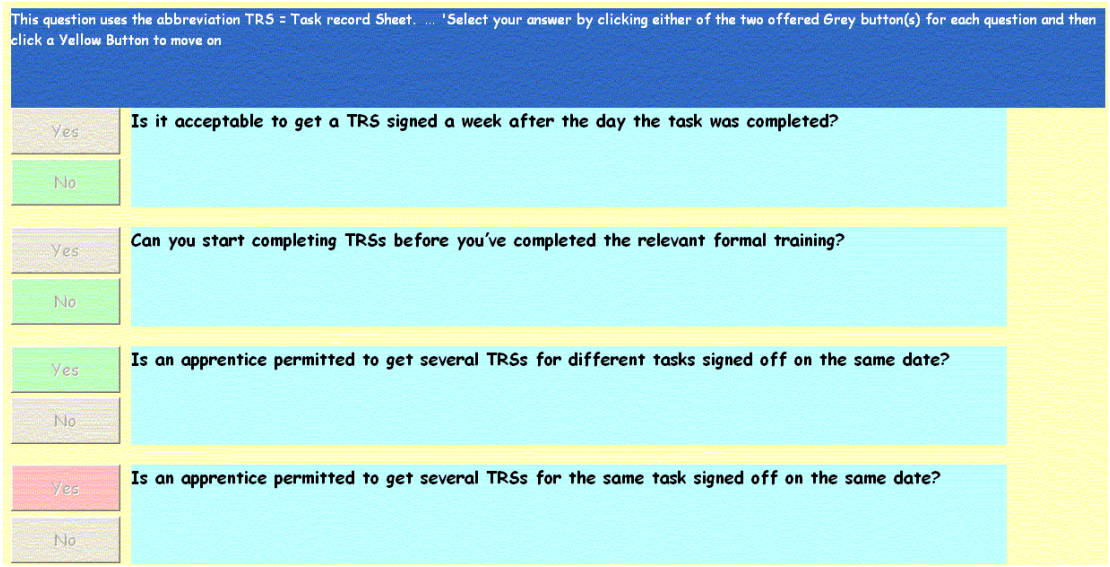

Figure 1 – An example screen print showing an example of a Multiple Alternative Choice test item containing a randomised sequence of four of the items from the experiment (MAC02, MAC04, MAC01, MAC03)

2 CONTEXT

This paper presents supporting evidence for a wider ranging study (Foster 2009) which seeks to improve the output from software that generates MCQ test items automatically (Mitkov and Ha 2003, 2006). I seek to establish, with a combination of literary review and experimental evidence, the most appropriate format of MCQ test item for use within the featured domain. The method for evaluating the decisions produced by this review must demonstrate best practice in the design and delivery of the assessments. The results will assist the progress of the main study by establishing the suitability of the Multiple True False (MTF) / Multiple Alternative Choice (MAC) format in the featured domain.

2.1 The Revised Taxonomy of Multiple-Choice (MC) Item-Writing Guidelines

The Review of the updated Taxonomy (Haladyna, et al 2002) identifies Guidelines, which attract unanimous agreement from all available studies:

"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."

This leaves several guidelines, which the review accepts are open to debate. These Guidelines have been satisfied as stated in the MCQ routine featured in this paper however Guideline 9 (G9) is insufficiently precise for it to be useful in the featured domain. In its original form Guideline 9 states:

"Use the question, completion, and best answer versions of the conventional MC, the alternative choice, true-false (TF), multiple true-false (MTF), matching, and the context-dependent item and item set formats, but AVOID the complex MC (Type K) format".

2.2 Multiple True False (MTF) / Multiple Alternative Choice (MAC) item format

The Multiple Alternative Choice (MAC) test item format is a generalised 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. An example screen print showing a randomised sequence of four of the items from the experiment (MAC01, MAC02, MAC03, MAC04) is provided in Figure 1.

The experiment involves the comparison of responses to the item shown in Figure 1 with responses to the following four equivalent individual Multiple Choice test items.

Table 1– Four non-G9a-compliant MC items that cover the same content as the G9-compliant MAC item displayed in Figure 1.

MCQ01	**TRS = Task record Sheet** Can you start completing TRSs before you've completed the relevant training? A) Yes B) No
MCQ02	**TRS = Task record Sheet** Is it acceptable to get a TRS signed a week after the task was completed? A) Yes B) No
MCQ03	**TRS = Task record Sheet** Is an apprentice permitted to get several TRSs for the same task signed off on the same date? A) Yes B) No
MCQ04	**TRS = Task record Sheet** Is an apprentice permitted to get several TRSs for different tasks signed off on the same date? A) Yes B) No

3 EXPERIMENT

3.1 Hypothesis

In order to be useful, the guideline instructing item designers in the featured domain needs to be more specific by recommending a specific MCQ item format that is most suitable for the domain. For the purposes of this study, an adapted version of Guideline 9 (G9a) has been prepared.

"Use the multiple true-false (MTF) or multiple-alternative-choice (MAC) item format, but AVOID the matching, context-dependent item, item set, question, completion, and best answer versions of the conventional MC and single alternate choice, true-false (TF), formats"

3.2 Method

The hypothesis has been tested by delivering both G9a-NON-compliant test items and G9a-compliant test items to 28 new entrants to the featured UK company. Two parallel experiments have been conducted to test G9a in the featured domain. Group A (14 members) took the assessment routine containing 8 G9a-non compliant items MCQ01, MCQ02, MCQ03 etc. and 8 G9a-compliant items (MAC09, MAC10, MAC11 etc).

Meanwhile, group B (14 members) were presented with 8 G9a-non compliant items MCQ09, MCQ10, MCQ11 etc. which tested equivalent content to MAC09, MAC10, MAC11 etc) and 8 compliant items (MAC01, MAC02, MAC03 etc.) which tested equivalent content to items (MCQ01, MCQ02, MCQ03 etc).

3.3 Evaluation

The adapted Guideline will be assessed as acceptable if the change in item difficulty between a G9a-compliant item and a G9a-non-compliant item is not significant and the response time for the G9a-non-compliant item is the same or less than the response time for a G9a-compliant item.

4 RESULTS

For each test item a record was made of the option selected by each apprentice along with the time taken to respond. 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 G9a-compliant items and 14 sets for responses for G9a-non-compliant items. All tests were conducted under controlled conditions however some candidates completed the test without recording a response to some components of some of the MAC items. All responses from any candidate whose response record included one or more 'no response' record(s) were excluded from the analysis of results in order to facilitate comparisons.

Table 2 summarizes the comparison between G9a-compliant and G9a-non-compliant test item response data where ID is the Item Difficulty calculated using the technique described in Swanson D.B., Holtzman, K.Z.,Allbee K.,Clauser, B.E., 2006.

Table 2: Results from the comparison of response data between G9a-compliant and G9a-non compliant items.

Item	IDn-IDc	RTn – RTc
1	0.05	-86
2	0.10	-315
3	-0.11	-88
4	0.03	-103
5	-0.01	18
6	0.31	101
7	-0.14	-95
8	0.15	45
9	0.09	103
10	0.31	45
11	0.06	-15
12	0.09	5
13	-0.08	-16
14	-0.24	-30
15	0.32	-102
16	0.20	-96

IDc indicates that the measurement has been calculated from item response data for G9a-compliant items and IDn indicates that the measurement has been calculated for G9a-non compliant items. The IDn-IDc column therefore contains the difference between these two Item Difficulty values and the RTn – RTc column contains the difference in seconds between the total recorded response times for compliant and non-compliant items.

The method chosen to calculate Item Difficulty (ID) for this experiment is the one used by (Swanson et al. 2006). The psychometric characteristics used in (Swanson et al 2006) also include the Logit Transform of the ID. This is an attempt to compensate for the non-linear nature of the difficulty curve as ID values approach ID=1.0. There is no need for this adjustment in the current study since the purpose is comparison of equivalent measures. The Item-Total bi-serial correlation coefficient for each of the altered MCQ test items is also referred to in (Swanson et al. 2006). This measure is not usable because the calculation leads to 'divide by zero' errors when ID=1.0.

5 CONCLUSIONS

Both the wide variation in Item Difficulty values and the wide variation in response times for each of the 16 item pairs featured in this experiment prevent any definite conclusions from this experiment. However the mean response time when comparing G9a-non-compliant items with G9a-compliant items shows a significant reduction (over 39 seconds), and the mean change in Item Difficulty (0.07) is small, and so this provides some evidence in support of the adoption of G9a into the item-writing guidelines for the featured company.

Future experiments will include further investigations of the performance of the MAC item format in this domain. These experiments will also include the application of some new theories in source document pre-processing (Foster, 2009) in on-going work that seeks to improve the output from the MCQ test item generator software (Mitkov, R., and L. A. Ha. 2003, 2006).

ACKNOWLEDGEMENTS

I 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.

REFERENCES

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, Bulgaria (Student Conference)

Haladyna, T.M., Downing, S.M., Rodriguez, M.C., 2002 *A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment* In Applied Measurement in Education, 15(3), 309-334

Mitkov, R., and Ha, L. A., 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.

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

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: October 2006 - Volume 81 - Issue 10 - pp S52-S55 doi: 10.1097/01.ACM.0000236518.87708.9d