

Computerized long-menu questions as an alternative to open-ended questions in computerized assessment

L W T Schuwirth, C P M van der Vleuten, H E J H Stoffers¹ & A G W Peperkamp

Department of General Practice and 1 Department of Educational Research, University of Limburg, Maastricht, The Netherlands

SUMMARY

To optimally avoid cueing effects and computer scoring problems in computerized examinations a computerized long-menu question (CLM) was developed. This question type was compared to open-ended questions in one treatment group and to multiple-choice questions in another treatment group. Also, scores were compared to self-perceived computer anxiety of the participants. CLMs yield comparable scores to open-ended questions, but the scores differ significantly from those on multiple-choice tests. Correlations in the first comparison (CLMs with multiple-choice) were higher than those in the second comparison (CLMs with open-ended questions). The amount of positive and negative cueing was considerably higher in the first than in the second comparison. Response times of CLMs were higher than those of multiple-choice questions and open-ended questions, differing significantly from both. Computer anxiety did not influence the mean scores in either comparison. Therefore, in computerized testing CLMs seem to offer an acceptable replacement of open-ended questions.

Keywords

*Computers; *education, medical, undergraduate; *educational measurement; Netherlands

From their earliest uses, multiple-choice questions (MCQs) have been used extensively in all kinds of examinations (McGall 1920). The reasons for this may be that they yield more reliable scores and require low labour intensity in answering and scoring (Swanson *et al.* 1987). Two major disadvantages have also been reported; the relatively high possibility of a correct answer by sheer guessing and the cueing effect.

To discourage sheer guessing, several methods have been described. The effects of these different methods to oppose guessing and the guessing itself have been the object of many studies (West 1923; Ruch & DeGraff 1926; Votaw 1936; Van Naerssen 1961; Lord 1963; Mattson 1965; Harden *et al.* 1976). In summary, the

results show that correction for guessing may often be necessary, but it has negative psychometric effects, probably through the introduction of more random error.

The cueing effect, an effect suggesting that examinees on seeing the correct answer will recognize this as such, was first suggested in the 1950s in the domain of knowledge of word meaning (Hurlburt 1954). Twelve years later, a first study was conducted to assess the influence of cueing in medical examinations (McCarthy 1966). Especially in tests aimed at assessing higher-order cognitive skills, like problem-solving or clinical reasoning, cueing was perceived to be disadvantageous because recognition of a correct answer was not considered a higher-order cognitive skill (Newble *et al.* 1979). Many studies have been performed in which different item formats were compared by using parallel tests (McCarthy 1966; Hettiaratchi 1978; Newble *et al.* 1979; Norman *et al.* 1987; Page *et al.* 1990; Stalenhoef *et al.* 1990; Veloski *et al.* 1993; Schuwirth *et al.* 1994). In almost all of these studies, mean scores on the MCQs were higher than those on the open-ended questions (OEQs) (McCarthy 1966; Hettiaratchi 1978; Newble *et al.* 1979; Norman *et al.* 1987; Page *et al.* 1990; Veloski *et al.* 1993), but the opposite has also been found (Stalenhoef *et al.* 1990; Schuwirth *et al.* 1994).

Inter-test correlations, however, proved to be high in most comparative studies, especially when corrected for attenuation (by which an estimate of the correlation is given when both tests would have had an ideal reliability) (Maatsch & Huang 1986; Norman *et al.* 1987; Stalenhoef *et al.* 1990; Schuwirth *et al.* 1994). One study reports low to moderate correlations (Hettiaratchi 1978), but in view of the length of the tests used (the author does not report reliabilities) it may be assumed that disattenuated correlations would be much higher.

In one of our previous studies, the same test was administered to the examinees twice using a computer (Schuwirth *et al.* 1992). The design enabled cross-tabulations per item of the possible answer combinations: both answers correct, both answers incorrect and

two combinations of one correct and one incorrect answer. Apart from the (expected) finding that cueing in favour of the MCQ occurred in all items, an opposite effect was also found. In nearly all items, a substantial percentage of the examinees answered the OEQ correctly and the parallel MCQ incorrectly. We referred to the expected effect as 'positive cueing' and to the opposite as 'negative cueing'.

Generalizations from these studies as to the size of the cueing effect or to one of the question formats being intrinsically superior must be made with extreme caution. Correlations are quite high in most studies. Furthermore, it is impossible to disentangle cueing from sheer guessing with the methodology used. Norman *et al.* (1996, under editorial review) have described several setbacks of the methodology used. First, and most important, differences in means scores do not indicate that a different trait is measured. Second, even when low correlations are found, they would probably be suppressed by the unreliability of the tests used. Finally, cueing appears to be an unpredictable effect that does not only act in two directions, but its size seems to vary with the level of expertise of the examinees (Newble *et al.* 1979), the item difficulty (Schuwirth *et al.* 1992) and the content of the question (Swanson *et al.* 1987). The easiest way to avoid cueing would therefore be the sole use of OEQs, but their labour intensity in answering and scoring has encouraged the development of alternative question types that avoid cueing (Case & Swanson 1993; Veloski *et al.* 1993). In one of these, a so-called long menu is used (Veloski *et al.* 1993). This consists of a booklet with an alphabetically ordered long list of possible answers (over 500) each with its specific code. Examinees are then supposed to generate the correct answer, look it up in the booklet and to transfer the code to the answer sheet. Although it appeared to diminish the cueing effect substantially, the method was time consuming, and mistakes could be made in the process of transferring the code from the booklet to the answer sheet.

Computer scoring of OEQs with algorithms or natural language understanding appears far too inefficient to be used yet in high-stakes examinations (Sabah 1993). Therefore, we developed a computerized version of a long-menu question (CLM) in which examinees may find the answer in a list by typing it into a dialogue box. The computer then searches a long (over 2500 alternatives) list for 'hits'. The alternatives found are reported back to the examinees immediately so they can check whether the retrieved option is the desired one.

In this present study, the CLM was compared to normal (6-8 options) MCQ in one condition and to normal OEQ in a second condition.

Furthermore, all participants were given evaluation forms to indicate their familiarity with or anxiety for computers. The following research questions were asked:

- 1 Do CLMs resemble (psychometrically) more the OEQs or the MCQs?
- 2 Does the amount of experience with handling computers influence proficiency scores and response-times?

As parameters for the first question, mean scores, correlations, cross-tabulations and mean response times were used. As some studies indicate that the level of expertise may influence the amount of cueing occurring (Newble *et al.* 1979; Schuwirth *et al.* 1992), we also investigated whether possible effects are influenced by the level of expertise of the examinee.

The second question has mainly emerged from the statements of some students indicating that they are relatively unfamiliar with the use of computers. Norcini *et al.* (1986), on the other hand, have indicated that the level of self-perceived 'computer-illiteracy' is not reflected in proficiency scores.

METHOD

Material

Thirty written cases were used in a computerized test. All cases were transcriptions of real patients seen in a general practice setting, using a so-called key-feature approach (Bordage 1987). In this approach, the case descriptions are kept short, only reporting relevant characteristics, signs and symptoms, and questions are directed towards essential decision. Because the literature indicates that questions prompting for diagnosis show the most prominent cueing effect (Swanson *et al.* 1987), each case was linked to one question asking for the most probable diagnosis. In order to assess their fidelity, all cases were presented to three physicians first.

The computer presentation uses a Windows environment in which students are presented with the case first. They then have to generate the most probable diagnosis, click an 'OK' button to activate the question and answer it. The time between this activation of the question and the confirmation of the answer is recorded by the computer.

Design

Two conditions were created. In one condition the examinees were presented the set of cases linked to CLMs first, and then the same set of cases linked to MCQs. In the second condition, the first set consisted of OEQs and the second of CLMs. Therefore, based on the question format and the condition, four subsets can be distin-

guished. These will be addressed further by a three-letter abbreviation and a number. CLM1 would thus be the subset containing the CLMs in condition 1. Examinees were randomly assigned to either condition. Also, all examinees received an evaluation form, presenting several items (in a five-point Likert format) about this kind of testing. Two items explicitly asked about computer-illiteracy or -anxiety.

Subjects

Thirty medical students from each of the second, fourth and fifth year (of a 6-year, problem-based medical curriculum) participated in this study. Although the second and fourth year are preclinical years, students are regularly presented paper cases to solve. An increase in their ability to solve cases could therefore be expected. Fifth-year students were all in the last week of their clerkship in general practice (12 weeks). However, some of these students already had some clinical experience from previous clerkships, as the order of the rotations is not fixed. No academic credits could be gained by participation, but all examinees received a financial compensation for their efforts.

All examinees were instructed about the impossibility to return to a previous case. For the second set in each

condition, they were asked to consider each case as a new one, i.e. to reread the text and to reconsider the answer.

Scoring

Answers to the CLMs and MCQs were scored and filed by the computer; answers to the OEQs were filed and hand-scored afterwards. For this, a previously fixed answer key was used. All response times were filed. Scores were expressed as percentages of correct scores; response times were expressed in minutes.

Analysis

Descriptive statistics were calculated for scores and response times for every subset, including an estimate of the reliability of the scores using Cronbach's alpha. Within both conditions, a repeated-measures analysis of variance was used to test the effect of question format on mean scores and response times. Mean score differences between CLM1 and CLM2 were tested using an independent Student's *t*-test. Within both methods, correlations between the question formats were calculated. Cross-tabulations of all four possible answer combinations were made per item and were then summed for all 30 cases to

Table 1 Mean scores and reliabilities of all four subsets

Year of training	Comparison condition 1						Comparison condition 2					
	CLM 1			MCQ 1			OEQ 2			CLM 2		
	Mean	SD	Alpha	Mean	SD	Alpha	Mean	SD	Alpha	Mean	SD	Alpha
2	21.6	11.8	0.68	31.8	10.9	0.50	14.0	7.3	0.41	16.9	6.3	0.60
4	41.8	11.2	0.56	52.7	10.7	0.51	42.2	15.6	0.77	39.6	14.5	0.72
5	46.4	9.2	0.28	57.6	9.1	0.30	54.2	16.2	0.77	50.9	11.5	0.56
Total	36.6	15.2	0.76	47.3	15.1	0.74	36.2	21.2	0.89	35.5	18.4	0.84

Year of training	Comparison condition 1				Comparison condition 2			
	CLM 2		MCQ 1		OEQ 2		CLM 2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	29.91	9.04	9.63	2.83	15.17	8.87	17.93	4.34
4	36.04	8.31	10.71	3.10	20.36	12.83	22.80	6.72
5	30.77	12.86	10.69	5.47	19.49	12.00	19.0	5.90
Total	31.41	10.76	10.39	3.96	18.93	11.68	20.14	5.96

Table 2 Average response times needed to complete all items

Table 3 Correlations between the question formats

Year of training	Comparison condition 1 (CLM 1-MCQ 1)	Comparison condition 2 (CLM 2-OEQ 2)
2	0.72	0.74
4	0.46	0.88
5	0.57	0.95
Total	0.78	0.94

detect the magnitude of the positive and negative cueing effect. Positive cueing was defined here as cueing favouring the scores on the question format resembling the MCQ most strongly (MCQ1 and CLM2).

To test the effect of computer-illiteracy or -anxiety on mean scores and mean response times, a factor was calculated using the sum of the answers given to the items on the evaluation form. In both conditions, a repeated-measures analysis of variance was performed testing the mean score against this factor ('computer-anxiety'). This analysis was repeated for response times.

RESULTS

The scores and the response times are given in Tables 1 and 2.

In all subsets, an increase in mean scores is found with increasing level of training. Differences between the mean scores in condition 1 are larger than in condition 2. Analysis of variance shows a significant effect of question format in condition 1 ($F(1,44) = 52.21$; $P < 0.0001$). In condition 2, this effect is not significant ($F(1,44) = 0.46$; $P = 0.5$). The effect of level of training on mean scores is significant in all subsets ($P < 0.0001$).

All interaction effects are not significant in both conditions. Some differences in mean scores are seen

between the CLMs in condition 1 and the CLMs in condition 2, but they are all not significant.

Reliability estimates within the year-groups are moderate, but over the total of the examinees they are acceptable. In combination with the mean response times, an estimate of the reliabilities per hour of testing time can be obtained using the Spearman-Brown prophecy formula. Reliabilities would then be 0.86 and 0.94 for the CLM1 and MCQ1, and 0.96 and 0.94 for the OEQ2 and the CLM2.

The shortest response times are needed for the multiple choice. The differences in condition 2 appear to be smaller than in condition 1, but in both conditions the effect of question format on response time is significant ($P < 0.0001$). The differences between the mean response times on the CLMs in condition 1 and 2 are remarkable: they are all statistically significant (independent t -test, $P < 0.0001$).

In Table 3, the intertest correlations are shown. The correlations shown in this Table are observed correlations. In condition 1, these are lower than in condition 2. The correlation in year-group 4 in the first condition is considerably lower than all other correlations. This correlation estimate, however, is not statistically significant ($P = 0.232$), indicating that the number of observations is too small in this sample to determine such low correlations. All other correlations can be considered significant ($P < 0.05$).

The results of the cross-tabulations of the possible answer combinations are shown in Table 4. The net cueing (being the positive cueing minus the negative cueing) indicates the size of score differences resulting from the cueing (which has, in fact, been assessed by the analysis of variance). However, the total amount of cueing indicates the percentages in which both question formats did not agree about the competence of the examinee (based on the same case). In the comparison of CLMs with MCQs, more positive cueing occurs than in the comparison CLMs and OEQs. However, negative cueing occurs

Cueing type	Comparison condition	Year 2	Year 4	Year 5	Total
Positive cueing	1	17.1%	16.7%	15.8%	16.5%
	2	7.8%	6.7%	8.7%	7.7%
Negative cueing	1	6.9%	5.8%	4.7%	5.8%
	2	5.8%	9.3%	10.2%	8.4%
Total cueing	1	24.0%	22.5%	20.5%	22.3%
	2	13.6%	16.0%	18.9%	16.1%
Net cueing	1	10.2%	10.9%	11.1%	11.7%
	2	2.0%	-2.6%	-1.5%	-0.7%

Table 4 Percentages in which the four possible answer combinations occurred

slightly more often in the second comparison. A reasonable amount of net cueing remains in the first comparison; in the second comparison the net cueing deviates only marginally from 0%.

The effect of 'computer-anxiety' on the mean score is not significant in either condition ($P = 0.133$ in condition 1; $P = 0.678$ in condition 2), nor is the effect on response time ($P = 0.991$ and 0.858 , respectively). All interaction effects are not statistically significant.

DISCUSSION

The aim of the study described here was not to establish whether CLMs tap into different kinds of cognitive skills than OEQs or MCQs. The design of this study, the number of questions used, and the number of examinees involved would probably not allow such conclusions.

Because the CLMs are meant to be an acceptable replacement of OEQs, it is more practical to assess whether scores obtained using CLMs are more comparable to OEQs than to MCQs.

An issue that strongly influences the comparability of CLMs to either of the other question formats is the kind of standard setting used. The use of an absolute standard requires the scores on the question formats to be identical in absolute terms, whereas a relative standard setting method would only require the question formats to rank-order students similarly. Both parameters have been studied here.

It appears that an effect of question format on mean scores could only be detected when comparing the CLMs with MCQs. In view of the fact that no interaction effects of level of training could be detected, it seems that the mean effect is fairly stable.

The mean scores on the CLMs in condition 2 are somewhat lower than those on the CLMs in condition 1. As this difference is not statistically significant to any acceptable level, the most appropriate explanation is that it is an artefact; still, it would not have been surprising to have found a difference favouring the CLM2 because they were presented as the second set in a condition. Mobilization of prior knowledge could have been expected to result in a higher mean score on the CLM2, but an effect on proficiency seems to occur only when time allocation to the cases is restricted (Machiels-Bongaerts *et al.* 1993), which has not been the case here. Nevertheless, some inequality in the mean level of competence could be present, although 45 examinees were fully randomly assigned to both conditions. This has to be borne in mind in all of the comparisons between conditions described.

Answering CLMs appears to be more time-consuming than answering MCQs, and even than OEQs. Reasons for

this could be unfamiliarity with this question type, or incompleteness of the list used. Although over 2500 options have been included in the list, it is certainly possible that some of the alternatives considered by the candidates have not been included yet. This can, of course, be regarded as a (temporary) setback of this method. A clear difference exists between the mean response times for the CLMs in both conditions. The most probable explanation lies in the fact that, in condition 1, the CLMs were the first subset, and, in condition 2, the second. So, in the second condition students already had to formulate their answer, thus allowing them to find the answer in the CLM2 sooner than in CLM1. Another effect that could have attributed to this difference may be the mobilization of prior (related) knowledge (Machiels-Bongaerts *et al.* 1993). This mobilization appears to shorten the response times needed. This, then, could account for a decrease in response times on CLM2, even when the answer on the OEQ2 had been incorrect. The difference in response times did not influence the reliability estimate per hour of testing time. In these reliabilities, only minor differences were found. The observed correlations reported are suppressed by the unreliability of the subsets. Norman *et al.* (1996, under editorial review) suggest, therefore, that a disattenuation of the observed correlations should be performed. However, this procedure also corrects for content differences of both subsets. Because all subsets had an identical content, this would have resulted in an overestimation of the true correlation. However, it may be expected that some of the differences in correlations may be attributable to the difference in reliabilities, as these are slightly lower in condition 1.

It appears that the net cueing effect of CLMs can be neglected, but even in condition 2 disagreement in 13-19% of both question formats about the ability of the examinee (total cueing) exists. Negative cueing seems to appear more in condition 2 than in condition 1. A possible explanation for this could be that the number of synonyms and alternatives in the list used for the CLM is still too low. An examinee would then be able to answer the OEQ correctly, but not to find the answer in the CLM. In view of the already large numbers of alternatives, this may be expected to have accounted only for a small portion. Extension of this list would eventually lead to an elimination of this problem. In all year-groups, MCQs cue more than CLMs. In MCQs, the total amount of cueing averages at about 20%, which is congruent with one of our earlier studies (Schuwirth *et al.* 1992).

The level of perceived computer-anxiety seems to influence neither the mean scores nor the mean response times. The former is congruent with the findings of

Norcini *et al.* (1986). The latter would indicate that either the self-perception of computer-anxiety of the students is inadequate or that it is adequate but still does not interfere with performance. Our favourite explanation would be that the interface is so user-friendly that everybody can work with it. But whether this explanation is correct should certainly be studied further.

In summary, it seems that CLMs resemble an OEQ more than an MCQ and could therefore be an acceptable replacement for the OEQ when using a computer. However, because the disadvantage of response times needed is also present in CLMs, it is still advisable to let the content of the question decide what the most appropriate format for the question must be (McGuire 1987; McGuire 1994).

REFERENCES

- Bordage G (1987) An alternative approach to PMPs: the 'key features' concept. In: *Further Developments in Assessing Clinical Competence* (ed. by I R Hart & R M Harden), pp. 59-75. Can-Heal Publications, Montreal.
- Case S M & Swanson D B (1993) Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine* 5, 107-15.
- Harden R M, Brown R A, Biran L A, Dallas Ross W P & Wakeford R E (1976) Multiple choice questions: to guess or not to guess. *Medical Education* 10, 27-32.
- Hettiaratchi E S G (1978) A comparison of student performance in two parallel physiology tests in multiple choice and short answer forms. *Medical Education* 12, 290-6.
- Hurlburt D (1954) The relative value of recall and recognition techniques for measuring precise knowledge of word meaning. *Journal of Educational Research* 47, 561-76.
- Lord F (1963) Formula scoring and validity. *Educational and Psychological Measurement* 23, 663-72.
- Maatsch J L & Huang R H (1986) An evaluation of the construct validity of four alternative theories of clinical competence. In: *Proceedings of the Twenty-Fifth Annual Conference on Research in Medical Education*, pp. 69-74. AAMC, Washington DC.
- Machiels-Bongaerts M, Schmidt H G & Boshuizen H P A (1993) Effects of mobilizing prior knowledge on information processing: studies of free recall and allocation of study time. *British Journal of Psychology* 84, 481-98.
- Mattson D (1965) The effects of guessing on the standard error of measurement and the reliability of test scores. *Educational and Psychological Measurement* 15, 727-30.
- McCarthy W H (1966) An assessment of the influence of cueing items in objective examinations. *Journal of Medical Education* 41, 263-6.
- McGall W (1920) A new kind of school examination. *Journal of Educational Research* 1, 33-46.
- McGuire C (1987) Written methods for assessing clinical competence. In: *Further Developments in Assessing Clinical Competence* (ed. by I R Hart, & R M Harden), pp. 46-58. Can-Heal Publications, Montreal.
- McGuire C (1994) Letter to the editor. *Teaching and Learning in Medicine* 6, 74.
- Newble D I, Baxter A & Elmslie R G (1979) A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education* 13, 263-8.
- Norcini J J, Meskauskas J A, Langdon L O & Webster G D (1986) An evaluation of a computer simulation in the assessment of physician competence. *Evaluations and the Health Professions* 9, 286-304.
- Norman G R, Smith E K M, Powles A C, Rooney P J, Henry N L & Dodd P E (1987) Factors underlying performance on written tests of knowledge. *Medical Education* 21, 297-304.
- Page G, Bordage G, Harasym P, Bowmer I & Swanson D (1990) A revision of the Medical Council of Canada's qualifying examination: pilot test results. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, R J Hiemstra, A J J A Scherpbier & R P Zwierstra), pp. 403-7. Boekwerk Publications, Groningen.
- Ruch G M & DeGraff M H (1926) Corrections for chance and 'guess' vs. 'do not guess' instructions in multiple-response tests. *Journal of Educational Psychology* 17, 368-75.
- Sabah G (1993) Knowledge representation and natural language understanding. *AI-COM* 6, 155-86.
- Schuwirth L W T, van der Vleuten C P M & Donkes H H L M (1992) Open-ended questions versus multiple choice questions: an analysis of cueing effects. In: *Approaches to the Assessment of Clinical Competence* (ed. by R M Harden, I R Hart & H Mulholland), pp. 486-91. Page Brothers, Norwich, UK.
- Schuwirth L W T, Jean P, van der Vleuten C P M & van Santen M (1994) Problem analysis questions, een korte casusvorm voor het pre-klinische domein. [Problem analysis questions, a short case-format for the pre-clinical domain.] In: *Gezond Onderwijs 3* (ed. by E Houtkoop, J Pols & M Verwijnen), pp. 104-11. Gravenhage, The Netherlands.
- Stalenhoef B F, van der Vleuten C P M, Jaspers T A M & Fiolet J B F M (1990) The feasibility, acceptability and reliability of open-ended questions. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, R J Hiemstra, A J J A Scherpbier & R P Zwierstra), pp. 552-7. Boekwerk Publications, Groningen.
- Swanson D B, Norcini J J & Grosso L J (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 12, 220-46.
- Van Naerssen R A (1961) A scale for the measurement of subjective probability. *Acta Psychologica* 20, 159-66.
- Veloski J J, Rabinowitz H K & Robeson M R (1993) A solution to the cueing effects of multiple choice questions: the Un-Q format. *Medical Education* 27, 371-5.
- Votaw D F (1936) The effect of do-not-guess directions upon the validity of true-false or multiple choice tests. *Journal of Educational Psychology* 27, 698-703.
- West P V (1923) A critical study of the right minus wrong method. *Journal of Educational Research* 8, 1-9.

Received 5 June 1995; accepted for publication 31 August 1995