

A closer look at cueing effects in multiple-choice questions

L W T Schuwirth,¹ C P M van der Vleuten¹ & H H L M Donkers²

Departments of 1 Educational Research and 2 Medical Informatics, University of Limburg, Maastricht, The Netherlands

SUMMARY

This study investigates the cueing effect occurring in multiple choice questions. Two parallel tests with matching contents were administered. By means of a computer program, examinees of different training levels and professional expertise were presented the same set of 35 cases (derived from patient problems in general practice) twice. The first time the cases were linked to open-ended questions; the second time they were linked to multiple choice questions.

The examinees consisted of 75 medical students from three different years of training, 25 residents in training for general practice and 25 experienced general practitioners. Across groups, total test scores reflected a difference in mean scores on both formats, and a high inter-test correlation. Within each level of expertise, differences in mean scores and high correlations were also found. The data were further explored per group of examinees. Two types of cueing effects were found: positive cueing (examinees were cued towards the correct answer) and negative cueing (examinees were cued towards an incorrect answer). These effects were found at all levels of expertise and in almost all items. However, both effects decline with increasing level of expertise. Positive cueing mainly occurs in difficult items, whereas negative cueing mainly occurs in easy items.

Keywords

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

INTRODUCTION

Pass-fail decisions in medical examinations have major implications both for the candidates and for society. This requires examinations to be highly reliable. The aim for high reliability has been pursued mainly by adding more structure to test items; but adding more structure to test items tends to diminish their reality similitude or fidelity. According to some authors, this is particularly the case if the objective is to assess higher-order cogni-

tive skills (Newble *et al.* 1979). One reason for this is that multiple-choice questions (MCQs) were mostly used to test simple factual knowledge. Nevertheless, researchers have now come to realize that it was possible to test 'understanding' by means of MCQs, and many of these questions have been developed and used since (Elstein 1993). Another reason, though, why MCQs were considered unsuitable to test higher-order cognitive skills is that recognition of the right answer suffices to answer the question correctly. In open-ended questions (OEQs), on the contrary, spontaneous generation of the right answer is a requisite for answering the question correctly. The recognition of the correct answer, on seeing it in a row of distractors, is called the cueing effect. It is also because of this cueing effect that MCQs are believed to test a lower level of cognitive skills than OEQs (Newble *et al.* 1979; Elstein 1993). Attempts to assess the amount of cueing occurring in MCQs are complicated by the fact that cueing may be confounded with guessing. In studies focussing on outcome rather than on an introspection of the examinees, it is virtually impossible to disentangle both effects. However, the problem of guessing also applies, in part, to OEQs; correct answers may be based on any problem-solving strategy ranging from sheer guessing through educated guessing to a fully competence-based deduction, the amount of which cannot be established. Other disadvantages of OEQs have been reported, mainly in terms of ambiguity and lower reliability (Case & Swanson 1993).

Many studies have been performed comparing scores on a multiple-choice test and an open-ended test. Most of these studies used parallel tests (Newble *et al.* 1979; Norman *et al.* 1987; Page *et al.* 1990; Stalenhoef *et al.* 1990), but identical subtests have also been used (Veloski *et al.* 1993). The results of practically all these studies indicated a higher mean score on the multiple-choice subtest (Newble *et al.* 1979; Case & Swanson 1993; Page *et al.* 1990; Veloski *et al.* 1993), although the opposite was found in one study (Stalenhoef *et al.* 1990). The latter may be explained by the scoring system used in that study; it included subtraction of incorrect responses from the total test scores.

There are two consistent findings that contrast with the assertion that the format of the question limits the

order of cognitive skills that can be measured. First, inter-test correlations are invariably high (Maatsch 1980; Maatsch & Huang 1986; Stalenhoef *et al.* 1990; Case & Swanson 1993). Second, the content of the question appears to influence the amount of cueing occurring: questions asking for diagnostic skills generate a more prominent cueing effect than questions asking about the laboratory or management skills (Swanson *et al.* 1987). The first finding raises the question of whether OEQs add unique information to that already obtained by MCQs. The second may indicate that apart from the format of the question, its content may influence cueing.

The consistent direction of the mean effect suggests that cueing predominantly triggers towards the correct answer in MCQs. However, the opposite could theoretically occur also. Participants giving the correct answer on an OEQ could be pointed to a wrong answer by the distractors in the MCQ. To make the distinction between both types of cueing, we will introduce the terms *positive* and *negative* cueing. The former means cueing in favour of the MCQ, the latter means cueing in favour of the OEQ.

The first aim of this study is to investigate the amount of positive and negative cueing occurring in MCQs. The second aim is to assess the influence of level of expertise and the difficulty of the question on the magnitude of the cueing effects. We expected that positive cueing would primarily occur in more difficult items and in test scores of examinees of a lower level of ability. Conversely, for easy items and at a higher level of expertise the (relative) amount of negative cueing was expected to be higher.

METHOD

In this study, two identical tests, each containing 35 patient cases, were used. Each case was followed by one question. Both tests were presented to the examinees by the use of an interactive computer interface. This interface presented the cases and questions on the screen. Questions could be answered by clicking a mouse button on an alternative in the MCQ or by typing in the answer in a dialogue box when an OEQ was asked. The computer prevented the examinee from going back to a previous case.

In the first test the examinees had to complete the OEQs linked to each case and were asked for the most probable diagnosis. These cases were in a fixed order. Immediately after that, the second test was presented, but now linked to MCQs prompting for the most probable diagnosis. These MCQs provided four to eight possible answers. The cases in this test were in a fixed, though different, order too. The computer program prevented the examinees from returning to a previous screen to alter answers already given.

All cases were derived from real patients from general practice. Their transcriptions to paper cases were based on concepts of the key-feature approach (Bordage 1987). This implied a description of all relevant signs, symptoms and findings of a patient, with questions prompting for the essential elements of a case only. Prior to the administration, the cases, answering keys and multiple choice alternatives were judged by three experienced general practitioners, and based on their comments some minor adjustments were made. Questions asking for diagnosis only were chosen because cueing appears to be most prominent in this area (Swanson *et al.* 1987).

The examinees were requested to give one best answer in the highest level of detail possible (e.g. pneumonia, but if possible bacterial pneumonia or even *Pneumococcus pneumoniae*). In addition to that, they were explicitly instructed to consider all the alternatives given in the MCQs, even when they recognized the case from the open-ended set and remembered the answer given there. The test was experimental only, and had no (educational) consequences for the participants. Examinees were encouraged, though, to use the same strategy that they would use in a real test. All participants were volunteers and received a financial compensation for their efforts. Figure 1 shows an example of one of the cases, including the two possible question formats.

One hundred and twenty-five examinees were used with five different levels of expertise, 25 randomly selected second-year, 25 fourth-year and 25 sixth-year medical students (from a 6-year medical programme). Also, a group of 25 residents in training for general practice (after 1 year of training in a 2-year training programme) and a group of 25 experienced general

Peter Johnson (3 years old) suddenly had a fever of about 38.7 °C and developed a sudden rash. It started, according to his mother, with little red spots that grew into little fluid-filled bullae. On physical examination you see, indeed, the little red spots and the little fluid-filled bullae which have a greyish colour. On the chest the exanthema is more severe than on the arms and legs.

OEQ: What is the most probable diagnosis?

MCQ: Which of the following diagnoses is the most probable one?

1. rubella	4. varicella
2. scarlatina	5. urtica
3. exanthema subitum	6. exanthema infectiosum

Figure 1 An example of a case with two question formats.

practitioners (mean duration of active practice 11.3 years) were used, all randomly selected. All students came from a problem-based medical school and were used to being confronted with patient cases for problem-solving exercises. Therefore, a higher ability to solve cases could be expected in students having completed more years of their studies.

All answers were recorded and filed by the computer. The OEQs were hand-scored afterwards according to a previously fixed answering key. All test scores were expressed as percentages of correct scores.

Potential bias effects that may influence the answers to the cases in the second test, like activation of prior knowledge, may be present, but their presence or magnitude could not be determined in this design.

RESULTS

In Table 1, descriptive statistics of both tests are provided. There is an approximately equal increase in mean scores on both the OEQ test and the MCQ test with increasing level of expertise, except for the general practitioners. A question format within subjects, by level of training between subjects, two-way ANOVA was used to establish the significance of both effects. Both main effects were statistically significant: level of training ($F = 134.09$, d.f. = 4, $P < 0.0001$), and question-format ($F = 110.46$, d.f. = 1, $P < 0.0001$). The interaction effect level of training by question format was also significant ($F = 3.26$, d.f. = 4, $P = 0.014$).

Reliabilities of the MCQs are lower than those of the OEQs in the student groups, whereas the opposite occurs in the residents and general practitioner groups. What is striking is the sudden drop in reliabilities in the last two groups.

Due to the unreliability of both tests, the observed correlations reported here reflect an underestimation of the real association. Therefore, in test-format comparisons

using correlations. Normal *et al.* (1996, under editorial review) suggest a correction for attenuation to estimate the true correlations. To apply this correction here is only partly correct; because both formats tested similar content, attenuation can only be caused by general error and not by content differences. Therefore, the true correlations will reflect an overestimation of the real correlation. To obtain an estimation of the magnitude of this overcorrection, a rough and ready method was used. Both tests were randomly divided into two subtests, yielding four separate tests. This made it possible to calculate three types of correlations: between subtests with similar content and different format (RR = 0.77 and 0.89; $P < 0.0001$); between subtests with different content and equal format (RR = 0.78 and 0.980; $P < 0.0001$); and finally between subtests of which both content and format are different (RR = 0.73 and 0.81; $P < 0.0001$). Of all correlations reported here, the true correlations were 1.0 except for one (RR = 0.73 $R_{xy} = 0.96$). Although this method only marginally reflects a true split-half design it seems to indicate that overcorrection of the correlations is not substantial.

To determine the percentage in which positive and negative cueing occurred, a cross-tabulation was made for each item of all four possible combinations of answers. Subsequently, these results were aggregated across all 35 cases and examinees per group. Table 2 provides these percentages, in addition to the total amount of cueing appearing (positive plus negative cueing) and the net cueing effect remaining (positive minus negative cueing).

As Table 2 shows, positive as well as negative cueing occurs at all levels of expertise. The amount of cueing covaries with expertise rather closely. With more expertise, cueing diminishes, except for the residents and the general practitioners (GPs). However, judged by their test-scores, the residents had more expertise than the GPs. The positive cueing is a more prominent effect than the negative cueing. Across groups, the total cueing effect

Table 1 Descriptive statistics of both tests and correlations

Level of training	OEQ		MCQ		Correlation	
	Mean (SD)	Alpha	Mean (SD)	Alpha	Observed	True
2	34.17 (10.4)	0.59	43.89 (10.2)	0.51	0.60	1.00
4	59.43 (10.4)	0.60	70.74 (8.5)	0.41	0.41	0.83
6	71.54 (10.5)	0.64	75.89 (9.0)	0.50	0.65	1.00
GP trainees	80.57 (6.2)	0.14	86.40 (6.1)	0.33	0.67	1.00
GPs	77.14 (6.6)	0.14	83.54 (6.9)	0.36	0.64	1.00
Total	64.57 (19.1)	0.88	72.09 (17.3)	0.87	0.90	1.00

Level of training	Both answers correct	Both answers wrong	Positive cueing	Negative cueing	Total cueing	Net cueing
2	25.7	47.7	18.2	8.5	26.7	9.7
4	54.4	24.2	16.3	5.0	21.3	11.3
6	64.7	17.3	11.2	6.9	18.1	4.3
GP Trainee's	76.0	9.0	10.4	4.6	15.0	5.8
GPs	71.4	10.7	12.1	5.7	17.8	6.4
Total	58.4	21.8	13.6	6.1	19.9	7.5

Table 2 Percentages of open-ended question (OEQ) and multiple-choice question (MCQ) answer combinations per level of expertise

is rather large (approximately 20%). However, the net effect of cueing (i.e. the remaining cueing when negative cueing is subtracted from positive cueing) is considerably smaller (approximately 7%). The net cueing effect will produce differences in the total scores on both tests.

At the item level, similar comparisons (not reported) revealed positive as well as negative cueing in nearly all items. In two items there was positive cueing only, there were no items in which only negative cueing occurred. Calculating the net cueing effect per item showed that, in 28 items a positive net cueing, and in 6 items a negative net cueing effect remained. In one item, the amount of positive cueing was equal to the amount of negative cueing.

As it was expected that item difficulty influences the amounts of cueing, items were divided into two categories, difficult and easy. In each expertise group, the mean *P*-value on the OEQs was used as the cut-off score. Results are shown in Table 3.

Several aspects may be of interest in this Table. Again, the difference in net and total cueing is striking, indicating that using parallel tests would underestimate the magnitude of the overall cueing effect considerably, this could be regarded as hidden cueing. To obtain an impression of the magnitude of this hidden cueing, the net cueing is divided by the total amount of cueing and recalculated to a percentage. Subsequently, this percentage is subtracted from 100%. The resulting percentage, reported in the table as ' $P_{\text{hidden cueing}}$ ', therefore represents the proportion of the total cueing that cannot be seen in parallel test comparisons. It is seen both in difficult and in easy items and in all expertise levels. The large difference between hidden cueing in difficult items and in easy items indicates that much more cueing remains hidden in easy items compared to difficult items. So parallel tests comparisons may well overestimate the influence item difficulty on the amount of cueing occurring. A more direct way of showing this was used by calculating the proportion of cueing that can be attributed to the easy items. This was calculated by dividing the percentage of cueing in easy items by the

sum of cueing in easy and difficult items. In the Table these figures are in the rows $P_{e/(e+d)}$ and indicated using italics. Even in this tighter representation, the amount of overestimation of the influence of item difficulty on cueing effect is clearly visible.

The net effect in the easy items of the GP group was negative. In the calculation of $P_{\text{hidden cueing}}$ and $P_{e/(e+d)}$ the negative sign was disregarded, because the outcome still represents a visible portion of the cueing effect.

Neither in the overestimation of the influence of item difficulty nor in the amount of hidden cueing a clear tendency could be found with regard to the level of training.

DISCUSSION

The results of previous studies investigating the cueing effect were reproduced in this study: differences in mean scores on both question formats have been found, and correlations between scores on both formats were relatively high.

A more detailed explanation of the nature of the cueing effect yielded a bidirectional effect, which sometimes inflates scores on the MCQs (positive cueing) and sometimes deflates the MCQ scores (negative cueing). The positive cueing effect is considered as a disadvantage of the MCQ, and has been documented consistently. However, MCQs apparently also cue in the opposite direction: they lead the examinee to choose the wrong answer.

The overall cueing effect is quite sizeable. In approximately 20% of all answers given, cueing plays a role, either positive or negative. So, in 20% of the cases, items disagree as to the ability of the examinees. This amount of disagreement is only marginally reflected in the total scores, where both effects are partly neutralized. Here, a net effect of approximately 7% remained.

There was a clear relation between cueing and difficulty of items. Easy items tend to show more negative cueing; difficult items elicit more positive cueing. This has led to a striking difference in total and net cueing

Table 3 Percentages of different types of cueing in easy and difficult items

Level training	Difficulty	No. of items	Positive cueing	Negative cueing	Total cueing	Net cueing	$P_{\text{hidden cueing}}$
2	Difficult	19	20.8	4.6	24.4	16.2	36.3
	Easy	16	15.0	13.0	28.0	2.0	92.9
	$P_{e/(e+d)}$				52.4	11.0	
4	Difficult	18	25.3	4.2	29.5	21.1	28.5
	Easy	17	6.8	5.9	12.7	0.9	92.2
	$P_{e/(e+d)}$				30.1	4.1	
6	Difficult	14	18.3	7.7	26.0	10.6	59.2
	Easy	21	6.5	6.3	12.8	0.2	98.4
	$P_{e/(e+d)}$				33.0	1.9	
GP trainees	Difficult	15	18.9	5.6	24.5	13.3	45.7
	Easy	20	4.0	3.8	7.8	0.2	97.4
	$P_{e/(e+d)}$				24.1	1.5	
GPs	Difficult	13	23.4	3.1	26.5	20.3	23.4
	Easy	22	5.5	7.3	12.8	-1.8	14.1*
	$P_{e/(e+d)}$				32.6	8.1*	

*These values were calculated disregarding the negative sign.

appearing in easy and difficult items. The result of this would be that using parallel tests severely underestimates the amount of cueing occurring, and overestimates the influence of the item difficulty has on the magnitude of the cueing effect.

The negative cueing effect as such is rather peculiar as it is normally expected that it would lead to higher scores on the multiple-choice part than the open-ended part. We assume that, sometimes, with an OEQ, the cases described may be perceived as rather straightforward, while with the MCQ, the distractors may lead the examinee to believe that the case is more difficult, therefore tempting him or her to select a more unlikely answer.

The fact that the overall cueing effect tends to diminish as the level of expertise increases supports this interpretation. This effect may explain the loss in reliability which was found for the MCQs in relation to the OEQs. The cueing effect introduces more 'noise' in the measurement. However, the fact that in the resident and GP groups the MCQs proved more reliable than the OEQs is not in accordance with this explanation.

Although the validity of both formats was not the prime interest of this study, a few remarks can be made. Both formats seemed to discriminate the levels of expertise equally effectively. Nevertheless, a few subtleties can be noted. In both tests the average score of the residents exceeded the GP scores. This has been seen and documented before (Day *et al.* 1988; Grant & Marsden 1988;

Van Leeuwen *et al.* 1993). The phenomenon may be explained by findings in cognitive psychological research of clinical expertise. A consistent finding here is the 'intermediate effect' (Schmidt *et al.* 1990; Schmidt & Boshuizen 1993). This intermediate effect supposes a shift in the storage and retrieval of knowledge in the process of becoming an expert. This shift implies a change from a more fragmented to more encapsulated or compiled pathway in storing and retrieving knowledge, thus leading to a different approach to solving patient problems. Accepting this intermediate effect as a logical and valid phase of expertise, the OEQs in this experiment were slightly better at differentiating the GPs from the residents. Again, the noise introduced by the cueing effect may be responsible for this.

An equally interesting phenomenon was the sudden decrease in reliability, paralleled in both tests, in the high expertise groups. A ceiling effect as a possible explanation for this, could not be demonstrated. However, cognitive psychological research indicates an increasingly individually differentiated knowledge base, usually as a result of an accumulation of idiosyncrasies due to clinical exposure. The phenomenon of 'content specificity', i.e. the variability of examinee performance across cases, is well documented in the medical assessment literature (Case & Swanson 1993; Elstein 1993; Van der Vleuten & Swanson 1990; Van der Vleuten *et al.* 1994). It would be logical for content specificity to increase with

level of expertise. The decrease of reliabilities could be a reflection of this. Unfortunately, the design used here does not allow separation of examinee by case interaction from general error in order to check this explanation.

Some practical implications may come from this study. As cueing was found to be a bidirectional effect, dependent not only on the level of expertise but also on the difficulty of the question, and, as the literature shows, the content of the question (Swanson *et al.* 1987), one cannot simply assume that all MCQs are easier than (parallel) OEQs. Yet, with regard to accuracy of measurement, these data could indicate that OEQs are to be preferred over MCQs. MCQs provide, to some extent, a biased estimate of the ability due to cueing effects. Yet the net effect is not dramatic and the amount of common information, i.e. their intercorrelation, is considerable. This should be weighed against the increase in resource requirements when OEQs are used (correction time, testing time needed to cover domain). The decision about which question format to use may vary from situation to situation.

REFERENCES

- Bordage G (1987) An alternative approach to PMP's: The 'Key features' concept. In: *Further Developments in Assessing Clinical Competence*. (ed. by I R Hart & R M Harden). 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.
- Day S C, Norcini J J, Webster G D, Viner E D & Chirico A M (1988) The effect of changes in medical knowledge on examination performance at the time of recertification. In: *Proceedings of the 27th Annual Conference on Research in Medical Education* (pp.139-144). Association of American Medical Colleges, Chicago, Illinois.
- Elstein A S (1993) Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Academic Medicine* 68, 244-9.
- Grant J & Marsden P (1988) Primary knowledge, medical education and consultant expertise. *Medical Education* 22, 173-9.
- Maatsch J L (1980) *Model for a criterion-referenced medical specialty test*. Final report, Grant No. HHS-02038-02. Office of Medical Education Research and Development, Michigan State University.
- Maatsch J L & Huang R H (1986) An evaluation of the construct validity of four alternative theories of clinical competence. *Proceedings of the Twenty-fifth Annual Conference on Research in Medical Education*. Association of American Medical Colleges, Washington DC.
- Newble D I, Baxter A & Elmslie R G (1979) A comparison of multiple choice and free response tests in examinations of clinical competence. *Medical Education* 13, 263-8.
- Norman G, Smith E K G, Powles A C P, 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 B (1990) A new approach to assessing clinical problem solving skills by written examination: conceptual basis and initial pilot test results. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, R J Hiemstra, A J J A Scherpbier & R P Zwierstra). Boekwerk Publications, Groningen.
- Schmidt H G & Boshuizen H P A (1993) On acquiring expertise in medicine. *Educational Psychology Review* 5, 205-21.
- Schmidt H G, Norman G R & Boshuizen H P A (1990) A cognitive perspective on medical expertise: theory and implications. *Academic Medicine* 65, 611-21.
- Stalenhoef-Halling B F, van der Vleuten C P M, Jaspers T A M & Fiolewt J F B M (1990) The feasibility, acceptability and reliability of open-ended questions in problem-based learning curriculum. In: *Teaching and Assessing Clinical Competence* (ed. by W Bender, R J Hiemstra, A J J A Scherpbier & R P Zwierstra). Boekwerk Publications, Groningen.
- Swanson D, Norcini J & Grosso L (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 12, 220-46.
- van der Vleuten C P M & Swanson D B (1990) Assessment of clinical skills with standardized patients: the state of the art. *Teaching and Learning in Medicine* 2, 58-76.
- van der Vleuten C P M, Newble D I, Case S *et al.* (1994) Methods of assessment in certification. In: *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence* (eds D I Newble, B Jolly & R Wakeford). Cambridge University Press, Cambridge.
- van Leeuwen Y D, Pollemans M C, Düsman H, van der Vleuten C P M & Grol R P T M (1993) De huisartsgeneeskundige kennistoets constructvaliditeit en betrouwbaarheid welke maten meten wat? [A knowledge based test for general practitioners, construct validity and reliability, what is measured by which measures?] In: *Gezond Onderwijs 2, proceedings van het tweede Gezond Onderwijs Congres [Proceedings of the second National Conference on Medical Education]*. (ed. by J C M Metz, A J J A Scherpbier & E Houtkoop). Universitair Publikatiebureau KUN, Nijmegen, The Netherlands.
- 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.

Received 20 July 1995; accepted for publication 31 August 1995