Modified Essay Question

PAUL STRATFORD
and HALYNA PIERCE-FENN

This article examines the reliability and validity of the modified essay question. An 18-item modified essay question was developed and pretested on a group of clinical experts. After the pretest, the modified essay question was administered to a group of 25 physical therapy students immediately before their second level orthopedic placement. A reliability index of .39 (coefficient alpha) was calculated for the test. In addition, measures of validity were obtained by correlating the modified essay question scores with multiple-choice question scores and the students' in-clinic clinical reasoning scores. The results suggest that the reliability coefficient found in this study is comparable to the medical literature. Also, the results showed that the modified essay question demonstrated a greater correlation with the in-clinic clinical reasoning score compared with the parallel content multiple-choice examination. Furthermore, preliminary evidence suggests that the modified essay question may be a useful paper and pencil tool in evaluating clinical reasoning. Finally, several shortcomings of this study are discussed in conjunction with a direction for further study.

Key Words: Educational measurement, Physical therapy.

The challenge to educators in the health care professions is to produce clinicians who are competent to practice in a variety of settings and who have the abilities to respond to fluctuating changes and demands placed on them within the health care system.

Aspects of clinical competence are considered to comprise a sound knowledge base and a repertoire of motor skills and clinical reasoning ability to gather data, to generate and refine hypotheses, and to plan and perform treatment. The achievement of these aspects of competence is usually expected within a given time, which students and faculty often consider as too short a duration. At Mohawk College, the physical therapy program extends over 33 months for achievement of clinical competencies.

The design of this program is different from traditional physical therapy education models in structure and process. Important, ongoing aspects of the curriculum are development of skills in efficient, appropriate accessing of learning resources; identification of personal learning needs; self-evaluation and peer evaluation; and critical reasoning for data gathering, analysis, synthesis, application, and problem solving. Major emphasis is placed on the process and quality of learning.

For learning to be relevant to the student, the university must provide the opportunity for factual knowledge and newly acquired motor skills to be applied and reinforced in the clinical environment. Thus, after the first year of the program, in which courses are geared primarily to introductory basic and behavioral sciences, the program design is based on an integrated systems model.

In this model, the curriculum design is laid out in order of complexity of body systems gauged by the complexity of learning that is thought to take place. A sequence of courses, consisting of three musculoskeletal courses, a cardiovascular-respiratory course, and a neurology course, is followed by junior and senior interning experience. The integrated systems model also requires that clinical experience be relevant to the academic learning, which is taking place concurrently with the clinical experience or immediately before it. Faculty members' concerns have continued with little abatement regarding program and course objectives, teaching methods, and evaluation procedures needed to assess student progress in achieving the clinical competencies of the program.

MODIFIED ESSAY QUESTION

A more immediate concern, given the growing problems of limited clinical resources and concerns regarding patient rights and safety, is finding an instrument that may predict student performance and success during a clinical affiliation. In integrated systems teaching blocks, in which the academic component was completed and examined before the same system's clinical affiliation, the question rose whether academic performance could be used to predict clinical affiliation performance. In June 1983, during the teaching of the musculoskeletal system (level II), we used the modified essay question (MEQ) as part of the academic evaluation procedure to answer this question.

The MEQ was initially designed by Hodgkin and Knox as an examination tool for the Royal College of General Practitioners. This testing approach tapped the students' ability in problem solving and in content-related aspects of basic and clinical sciences when real or simulated patients would be inappropriate. Subsequently, the MEQ has been used in undergraduate education.

The MEQ is based on a factual case history that is usually presented in stages (Appendix). In designing a question, the evaluator must decide on the nature and level of the competencies to be assessed. Feletti and Engel have suggested

Mr. Stratford is Teaching Master, Department of Physiotherapy, Health Sciences Education Centre, Chedoke Campus, Mohawk College, PO Box 2034, Hamilton, Ontario, Canada L8N 3T2, and Assistant Professor, Department of Medicine, McMaster University, Hamilton, Ontario.

Ms. Pierce-Fenn is Teaching Master, Department of Physiotherapy, Mohawk College.

This article was submitted May 7, 1984; was with the authors for revision 17 weeks; and was accepted December 28, 1984.
the best approach is to refer to the course objectives. When constructing a MEQ, care should be taken to avoid an abundance of questions that only require factual information. Both Feletti and Irwin and Bamber have demonstrated that items can be constructed to correspond with a spectrum of cognitive areas, such as knowledge, application, and synthesis. Another important aspect in designing a MEQ is whether cumulative error is desirable. Cumulative error can occur when the responses to several subsequent questions are related to a previous response. This error can be avoided by providing the students with the appropriate response before they answer a subsequent question. When this format is used, the MEQ can take on the appearance of the sequential management problem (SMP) as described by Berner and associates. The SMP is a structured patient management problem that provides feedback to the student after the student’s response to each subsection (eg, history, physical examination, investigation, and management). Finally, the questions or cases need not be the same length.

The questions in a MEQ are usually presented to the students in a booklet with one question a page. The examinations are supervised, the students are not allowed to read ahead, and they are not allowed to go back and correct previous answers. Other presentation formats, such as the presentation of problems on an overhead, have been used in the past but have been reported to be too stressful for the students. The correct responses for MEQ questions are usually generated by a small number of clinical experts. To complement the response list, the Newcastle group has used an open form approach. After the completion of an examination, the students and faculty meet to review the correct responses. Additional responses gathered by the students may be added to the correct response list.

The scoring of the MEQ is usually conducted in one of two ways depending on whether the results are aimed at differentiating among students (norm referenced) or whether the student met some criterion standard (criterion referenced). If the results are to be norm referenced, an extended scoring approach (0 to 10) for each item has been suggested. A criterion referenced format only requires a 2-point scale (0 and 1) for each item. Few formal studies have been performed to evaluate the reliability and validity of the MEQ. Initial work reported by the Royal College of General Practitioners suggested that the MEQ has several advantages over the traditional essay question (TEQ). First, interrater reliability was much higher for the MEQ. This finding agreed with that of Newble and colleagues who reported an interrater reliability coefficient of .95 for the free response question. Second, the MEQ can cover a greater spectrum of content than the TEQ. The MEQ also is able to test more than factual information. When candidates’ scores for the MEQ were compared with those scores for the TEQ, the MEQ tended to have fewer equivocal grades. These results were not statistically significant.

The multiple-choice questionnaire (MCQ) also can sample a broad spectrum of content within a topic area. The major limitation to the MCQ is that cueing occurs. The student is only required to recognize the correct response rather than recall the response. The MEQ does not appear to have this limitation. Newble et al compared the MEQ with the MCQ and demonstrated that the scores on the MEQ were significantly lower than the scores on the MCQ.

Reliability

Initial reports on the reliability of the MEQ are encouraging. Feletti has reported reliability coefficients (alpha-60) across various body systems ranging from a low of .43 for psychiatric disorders to a high of .90 for hematopoietic and endocrine disorders. The median value was .74 for gastrointestinal disorders. Newble and colleagues also reported statistically significant test reliability coefficients. The magnitude of these coefficients, however, was not reported. Subsequent work by Feletti and Gillies supports the magnitude of previously reported coefficients.

Validity

Attempts to establish the extent to which the MEQ is a valid instrument to evaluate problem solving have taken several directions. Educators are in general agreement that the MEQ possesses content validity in requiring the student to integrate data from the history and physical examination to generate hypotheses related to the patient’s problem and to suggest the most appropriate forms of treatment. Construct validity has been examined in several ways. Consistent results suggest that the scores obtained on MEQs improve with the level of medical education (eg, residents perform better than senior students who, in turn, perform better than junior students). The Newcastle group has advanced a second construct demonstrating that the ratio of problem solving to knowledge questions is greater on second-year assessments compared with first-year assessments. Other validity studies have attempted to compare the students’ performance on the MEQ with other concurrent measures of clinical competence. Irwin and Bamber found a moderate correlation of .41 between the MEQ and the final physician clinical examinations. Newble and colleagues were unable to correlate the students’ results on a free-response test to that of their clinical performance because of the low reliability of the clinical evaluation. The purposes of our study were 1) to examine the reliability of MEQ when the content consists of physical therapy musculoskeletal questions, 2) to determine the extent to which the students’ scores on the MEQ examination correlated with their scores on a parallel MCQ examination, and 3) to determine the extent to which the students’ scores on the MEQ could predict the students’ clinical reasoning ability when evaluated in the clinical setting.

METHOD

Six MEQs were developed and pretested for face validity and construct validity. We based face validity on the assessment by a group of clinical experts of the items on the MEQ as representative of the content required to work in the clinical setting. We based our construct validity on the relative performance of clinical experts and students. Clinical experts performed significantly better than second-year students who, in turn, performed significantly better than first-year students on the MEQ. Linear contrasts had demonstrated that these results were statistically significant (p < .05).

The study sample consisted of all second-year students (N = 25) enrolled in the musculoskeletal system block. During the final week of the academic por-
tion of the musculoskeletal system II block, we formally evaluated the students. The evaluation consisted of the students completing six MEQs. Each MEQ consisted of three items: hypotheses generation, hypotheses refinement, and patient management. Four of the six problems presented different problems at different anatomical sites; the two remaining problems represented a different problem at the same joint and the same problem at a different joint. The six MEQs and a 50-item MCQ on parallel content were administered at the same sitting. Each item of the MEQ was presented on an overhead projector. The time that each item was displayed was predetermined and ranged from a period of one minute to a period of four minutes. The 50-item MCQ was administered immediately after the students completed the six MEQs. The students were then divided into groups of three to five and placed in six teaching hospitals. At the end of 2.5 weeks of clinical practice, the students were evaluated for their clinical reasoning skills by their supervising therapist and clinical instructor. The in-hospital clinical reasoning skills were scored on a 10-point scale; 6 was the minimal acceptable level (criterion). Behavioral objectives were written to define the minimal acceptable standard.

Data Analysis

An index of reliability was determined by calculating coefficient alpha for the MEQ format and the MCQ format. Coefficient alpha is a measure of internal consistency, that is, how consistently a student responds on a multi-item test. To compare the reliability coefficients, the Spearman Brown prophecy formula was used to calculate the projected reliability coefficients for a 60-item questionnaire (alpha-60). To test validity, intercorrelations were made using the Pearson product-moment correlation among the MEQ, MCQ, and the in-hospital clinical reasoning scores.

RESULTS

The reliability coefficients for both the MCQ and MEQ are presented in Table 1. The $R$ values were specific to the number of items on each of the formats; the adjusted $R$ values were standardized to a 60-item questionnaire. The overall $R$ value for the MEQ (.39) was lower than we hoped. A principal reason for the magnitude of this coefficient was that it was calculated based on 18 items. The adjusted $R$ values for both the MCQ and MEQ were comparable. In addition, the adjusted $R$ value for the same problem at a different joint was comparable to the MCQ and MEQ value. The adjusted $R$ values were considerably higher for different problems at the same joint.

Table 2 presents the intercorrelation among the MEQ score, MCQ score, and clinical reasoning score (criterion). We found a modest positive correlation between the mean MEQ score and clinical reasoning score ($r = .33; p = .06$). The modest correlation between the mean MEQ score and the MCQ score ($r = .28$), however, was not statistically significant. We found no evidence of a correlation between the MEQ score and the in-hospital clinical reasoning score ($r = .10$).

DISCUSSION

The adjusted reliability coefficient (alpha-60) calculated in this study is consistent with those reported by others. Of particular interest is the similarity of the reliability coefficient for the same problem at a different joint to that of a different problem at a different joint. In contrast, the reliability coefficients for a different problem at the same joint are substantially higher. These values are of particular concern for educators who wish to advance the problem-solving ability of students. The results suggest that these students tended to learn content or tests, or both, based on the anatomical site rather than to learn concepts that are transferable across sites. In part, these results may be due to the process used by the students in reviewing for an examination. For example, if the students used the standard textbook, the content is organized based on anatomical sites rather than on common assessment concepts. Finally, because the site-specific reliability coefficients were based on only three items, we must wonder if the observed differences are real or merely an artifact of the small sample.

The correlation between the MEQ and MCQ is somewhat lower than those reported by others. Irwin and Bamber demonstrated correlation coefficients of .43 and .36 between the MEQ and MCQ in medical students' final doctor of medicine examinations at Queen's University in Belfast. The magnitude of these coefficients is similar to that reported by Berner et al who demonstrated a correlation of .38 between the SMP and MCQ for sophomore medical students. The content of the questions in our study was strongly oriented toward clinical reasoning rather than toward knowledge alone. Conceivably, the correlation between the MEQ and MCQ could have been higher if we had asked more questions on the basic sciences and clinical sciences.

Although the correlation between the MEQ and the in-hospital clinical reasoning scores was disappointing, this low correlation for validity was not altogether unexpected because of the design base. Ferguson reported that the maximum validity coefficient possible is restricted by both the reliability coefficient of the predictor variable (mean MEQ score) and the reliability coefficient of the criterion variable (in-hospital clinical reasoning), so that the validity coefficient cannot exceed the square root of the product of the reliability coefficients. Thus, to enhance the validity coefficient, the researcher must first enhance the reliability coefficient. In our study, the reliability coefficient of the predictor variable was .39. In the future, this coefficient may be enhanced by increasing the number of items on the MEQ examination. Our study did not examine the reliability associated with assessing clinical reasoning in the clinical setting. Previous information

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Reliability Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>No. of Items</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
</tr>
<tr>
<td>MCQ</td>
<td>50</td>
</tr>
<tr>
<td>MEQ*</td>
<td>12</td>
</tr>
<tr>
<td>MEQb</td>
<td>Knee</td>
</tr>
<tr>
<td></td>
<td>Ankle</td>
</tr>
<tr>
<td>MEQc</td>
<td>(ligamentous sprain)</td>
</tr>
</tbody>
</table>

* Different problem, different joint.  
* Different problem, same joint.  
* Different joint, same problem.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Intercorrelation (r) Among Mean MEQ, MCQ, and In-Clinic Clinical Reasoning Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCQ</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>MCQ</td>
<td>.28</td>
</tr>
<tr>
<td>MEQ</td>
<td>.33</td>
</tr>
</tbody>
</table>
suggests that under similar conditions, a reliability coefficient of .50 is reasonable. Numerous sources of variation, such as a different expectation for the different observers, different types of cases, and the ease with which the cues are elicited from the patient, influence the magnitude of this reliability coefficient. Accordingly, by applying Ferguson’s formula to our reliability figure, our optimal validity coefficient would not be expected to exceed \( .44 (\sqrt{.39 \times .50}) \).

A second factor that could have influenced the magnitude of the validity coefficient is that 2.5 weeks elapsed between the administration of the MEQ and the criterion measure. During this time period, we expected some learning by the students but not at a uniform rate for all students.

The MEQ only required the students to interpret information already provided, but the clinical performance required the student to elicit and interpret the data. Therefore, clinical performance scores would vary with ability to elicit data.

The MEQ has potential as both an ongoing (formative) and final (summative) evaluation tool. When used in the formative mode, we suggest that the items be constructed in a manner to discourage cumulative error. When used as a summative evaluation tool, the issue of cumulative error is more complex. At least one author has suggested that cumulative error may provide an appropriate penalty to the student to reflect the potential harm he could have inflicted on the patient.\(^7\)

Although other researchers have suggested that problems presented to the students by overhead projection result in unnecessary stress to the students, we did not find this to be the case provided that trial runs were performed to determine an appropriate response time.\(^2\) If the student to faculty ratio is high, overhead projection also may provide the only feasible method of discouraging academic dishonesty when the MEQ is used in the summative mode.

CONCLUSION

Our work has demonstrated that the reliability of the MEQ for physical therapy content is comparable to that described in the medical literature. To obtain a high reliability coefficient, the faculty member must consider the number of items to be placed on the evaluation. In addition to considering the number of items on the instrument, preliminary evidence suggests that reliability coefficients may be higher when the content is based on anatomical sites rather than on concepts concerning assessment and management across sites. Finally, to determine whether clinically significant correlations exist between the MEQ scores and the in-hospital clinical reasoning scores, several steps must be taken with respect to the study design. First, a sufficient number of items must be present on the MEQ. Second, steps must be taken to boost the reliability coefficient of the in-clinic evaluators. Third, the study must be designed to minimize the influence of potential confounding variables, such as the 2.5 weeks between assessments. We found that the MEQ appears to be a potentially attractive instrument that minimizes the cueing effect present in MCQs and maximizes the interrater reliability coefficient over the TEQs.

REFERENCES

Referral: Left knee pain. Assess and treat (plain x-rays negative). Adam Beck is a 38-year-old hydro worker who twisted his left knee two days ago.

Q.4.1. List five questions you would ask Adam to help identify his problem:

Adam tells you that this is the first time that he has ever injured his knee. He states that he was walking across a field when he stepped in a gopher hole and twisted the knee. The knee gave way and he fell to the ground. He was able to limp back to the car by himself; however, the knee was quite uncomfortable. On returning to the office about 30 minutes later, he noticed that the knee was very swollen. Although the entire knee is sore, the discomfort is focused over the medial aspect of the knee. The knee has not locked.

Q.4.2. Suggest what immediate swelling to the joint is indicative of:

Q.4.3. Adam’s knee did not lock. Suggest the significance of a positive history of locking:

Q.4.4. Suggest three potential hypotheses for Adam’s problem:

Q.4.5. List one test you might perform to differentiate between a capsular and a meniscal lesion:

Q.4.6. List two areas where you would expect to find discomfort on palpation if he had dislocated his patella:

A bit of the physical examination.

Adam limps into your office on a flexed knee full weight bearing.

On inspection, the left knee is severely swollen.

The knee is warm to touch and tender over the medial joint line and over the medial epicondyle.

Passive ROM is from 0-degree flexion to 84-degree flexion—pain at the extremes.

Active ROM is from 12-degree flexion to 76-degree flexion.

There is no discomfort or weakness in stressing the left hamstring muscles.

There is marked inhibition and weakness of the left quadriceps femoris muscle when stressed at 12-degree flexion; however, you are unable to break his resistance at 45-degree flexion.

There is no discomfort with stress.

There is no apparent ligamentous laxity or discomfort with stress. There is discomfort when externally rotating the tibia on a fixed femur. This discomfort is noted over the medial aspect of the knee.

His sensation to temperature is intact.

Q.4.7. Suggest two likely hypotheses for his problem:

Q.4.8. List four management steps you would take immediately with Adam:

Q.4.9. Suggest why you are able to break Adam’s quadriceps femoris muscle resistance at 12-degree flexion but unable to break his quadriceps femoris muscle resistance at 45-degree flexion (two reasons):

Q.4.10. Suggest why the medial meniscus is at greater risk of injury than the lateral meniscus: