Medical Education 1992, **26**, 454–461

Learning, retention and recall of clinical information

J. C. SISSON, R. D. SWARTZ & F. M. WOLF†

Department of Internal Medicine and †Department of Postgraduate Medicine and Health Professions Education, University of Michigan Medical Center, Ann Arbor

Summary. A representative group of 33 medical students who were entering the junior year clerkships was tested for retention and recall of clinical information 3 months after taking an examination on the same subject. The students were not given an opportunity to review the subject. On 39 identical multiple choice test questions, the students' mean score declined 10 percentile points ($P < 0{\cdot}05$) from that on the original examination. On 40 comparable but previously unseen questions, the mean score fell 19 percentile points from that attained 3 months earlier. On open-ended questions of clinical reasoning, a third component of the assessment, the students performed at a level similar to those on the two multiple choice tests, but with greater variability. These assessments give data on retention and recall that have not previously been reported in the literature. Correlations among individual test components were moderate (r = 0·52–0·63). There was inconsistency of individual students in scores on the component tests, and, thus, variability in performance by students was marked. Retention and recall were weakly predicted by results on an initial multiple choice examination. In addition, on a subsequent assessment of knowledge, results from different types of tests were inconsistent, suggesting that these tests evaluate different forms of competence.

Key words: students, medical/*psychol; *recall, *retention; *clinical medicine; clinical clerkship; educational measurement; learning; clinical competence

Correspondence: James C. Sisson MD, Division of Nuclear Medicine, B1G412 University Hospital, University of Michigan Medical Center, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109–0028, USA.

Introduction

Can medical students efficiently recall clinical information learned months ago? The answer is assumed to be yes, for retention of knowledge is a foundation of medical practice. Of course, no one's memory is perfect, and there must be some decline in the ability to recite what was once learned. Yet, most evaluations in medical school are designed to require evidence of learning that has frequently taken place only a few hours or days beforehand. Intense cramming is standard for students, and remuneration, in the form of test scores including pass and fail, is apportioned for what may be fleeting knowledge. Academicians have little experience and no guidelines for expectations for prolonged retention. Yet, the problems posed will be even more vexing as advances in medical knowledge demand more from the already heavily taxed memories of students (Anderson & Graham 1980; Covell *et al.* 1985; Bordage 1987).

Retention of information is difficult to assess, and the educational system usually assumes that recall in the future, with or without review, will be proportional to what can now be remembered. However, it is possible that individual students will retain information disproportionately; there are few data on this subject. Moreover, a given student may remember certain categories of facts and principles better than others. Nevertheless, it is what the medical student, and eventually the doctor, can recollect over months and years that shapes the practice of medicine.

To assess the rates of decline in recall, we asked medical students to answer, without review, questions on clinical information they were previously asked to learn. Three components composed this assessment: questions that were identical to those taken on a final examination 3 months earlier, questions similar in content but not seen before, and questions asking abut the same information but written in a different format and requiring more reasoning. We report here the abilities of a cohort of students to recall clinical information and the variable and unpredictable retention of knowledge among students, giving new data and insights on the subject.

Methods

Second-year medical students at the University of Michigan Medical School took a final examination evaluating learning in the second term of a course designed to teach clinical information, 'Introduction to Clinical Science', in April 1991. The test group consisted of 33 of these 204 students. This cohort was selected because the students were meeting in a single site 3 months later (July 1991) as they entered the internal medicine clerkship for third-year students. The assignment of students to this cohort was based on several criteria (including student preference for this period), but there was no reason to believe that selection was biased towards student knowledge or ability. The performance of these 33 students on the final examination in 'Introduction to Clinical Science', 83·6 ± 4·8% (mean score ± the standard deviation), did not differ significantly from that of the entire class, 83·2 ± 7·1%. The students were unaware of the assessment in July until the day before the examinations were administered, and they were advised that review was unnecessary since the results would not alter their status but would be used to guide their future education.

The final examination in 'Introduction to Clinical Science' contained 157 multiple choice questions (MCQ). Forty questions (MCQ-April) were randomly selected from this examination (Arkin & Colton 1963) for a repeat assessment (MCQ-July); however, it was subsequently discovered that one question had been previously challenged by students, and this question was therefore discarded. Forty questions were also selected by random numbers from the makeup examination (MCQ-new July) that had been created from the same bank of questions developed for the course; none of the 33 students had previously seen the questions in this component.

Recall involving a greater degree of application and synthesis of information and perforce a higher level of reasoning was sought in 49 newly created questions that made up the third component of the assessment. The questions were in the form of clinical vignettes describing classic presentations of diseases that were to be learned along with the information covered in the two MCQ components, and were answered by filling in blanks and thus were open-ended questions (OEQ-July). The diseases were from oncology, cardiology–angiology, pulmonology, nephrology, endocrinology and gastroenterology. Answers to the questions included varingly: pathogenic mechanisms, historical and physical findings important to the diagnosis, or diagnostic conclusions. The queries were similar to those that would be encountered by the students as clerks on a clinical service. A typical question was: 'A 65-year-old woman nurse has developed crushing anterior chest pain that has been continuous for 2 hours. Her EKG shows ST elevation in leads V-2 to V-6. What is your diagnosis?'

The students were given 2·5 hours to complete the three parts of the assessment. Multiple choice questions were graded by computer; one of us (JCS) graded each of the items in the OEQ-July for each student. The results were expressed as per cent correct. Differences in mean performance on the three components in the July assessment were analysed using repeated measures ANOVA to determine levels of retention and recall compared to the MCQ-April performance. These data were graphed as box plots (Cleveland 1985) for additional interpretation. Moreover, students' scores on each component were intercorrelated to detect the consistency of performance, i.e. did the students who scored higher on MCQ-April tend to attain higher scores 3 months later on the July components? χ^2 analyses also were performed to ascertain whether students who received passing scores did so consistently among the components.

Results

For the 39 questions of the MCQ-April, the students scored 81·1 ± 8·0%, and 88% of these

Table 1. Performance of 33 medical students on three measures of recall and retention of knowledge (MCQ) and one measure of clinical reasoning or synthesis that requires the application and integration of knowledge (OEQ)

Type/Date of assessment	% correct Mean (SD)†	% correct Range	% (*n*) students with passing scores (≥70%)
MCQ-April‡	81·6 (8·0)	66·7–92·3	88 (29)
MCQ-July‡	72·1 (8·4)	56·4–87·2	55 (18)
MCQ new-July§	63·3 (8·8)	42·5–77·5	30 (10)
OEQ-July¶	67·8 (11·6)	37·4–85·0	52 (17)

† F (3, 131) = 45·5, $P < 0·0001$, repeated measures ANOVA. All means are significantly different from each other except that 'OEQ – July' does not differ significantly from either 'MCQ – July' or 'MCQ new July' based on Scheffé *a posteriori* comparisons, $P < 0·05$.

‡ Identical set of 39 multiple choice questions given in April and July.

§ New set of 40 multiple choice questions given in July; questions were drawn from the same pool as MCQ-April.

¶ Uncued, open-ended questions based on clinical vignettes given in July.

students attained a score of 70% (the traditional passing level) or higher (Table 1). For these same questions answered later (MCQ-July), the mean and standard deviation were: 72·1 ± 8·4%, and 55% attained a score of 70% or higher (Table 1). For the 40 questions of the MCQ-new July, the attainments were less: 63·3 ± 8·8% and 30% with at least a 70% score (Table 1).

For the 49 OEQ-July, the students attained a mean and standard deviation of 67 ± 11·6%, but the range of results was broader than on the MCQ assessments (Table 1). In fact, four students had scores of less than 50%, whereas this was true for only one student in the MCQ assessments. Overall, 52% scored 70% or better with the OEQ-July.

Students' mean performance in July declined significantly ($P < 0·05$) on all the measures from their original performance on MCQ-April. They did significantly worse on both the identical items of MCQ-July and on the previously unseen questions of MCQ-new July; moreover, average performance in July on the original items was significantly better ($P < 0·05$) than on the new items. Average performance on the OEQ-July questions that required clinical reasoning was consistent with the mean performances for both MCQ-July and MCQ-new July, and did not differ significantly from the results on either of these components.

Greater variability appeared among students within the July components of the assessment, particularly in the OEQ-July. Insight into this variability can be gleaned from the notched box plots of performance (Fig. 1). Each box plot is derived from five percentiles that summarize the distribution of all the students' performance. The top and bottom of the box represent the 75th and 25th percentiles, respectively, which, for example, reflect scores of 76·5% and 60·5% correct for OEQ-July. The middle 50% of students' scores are represented within the box, with the line representing the median, or a score of 70·1% correct for OEQ-July. The thickness of each of the boxes graphically represents the degree of homo- or heterogeneity of performance of the middle 50% of the class on each measure. The MCQ-April and the OEQ-July boxes are thicker, or somewhat more heterogeneous than the other two, indicating that there was a wider range of scores for the middle half of the class.

Looking again at the plot of OEQ-July, the distribution is not symmetrical in as much as the median is not near the exact middle of the box, which it tends to be for the MCQ components. The lines, or 'whiskers', extend above the boxes to the 90th percentile and below to the 10th percentile, which are 81·5% and 49·3% correct scores, respectively, for OEQ-July. The similarity or difference in length of these 'whiskers' for each measure reflects the degree of symmetry of the distribution of student scores. The fact that the longer 'whisker' for OEQ-July is associated with poorer performance indicates a negatively skewed distribution. The small circles denote

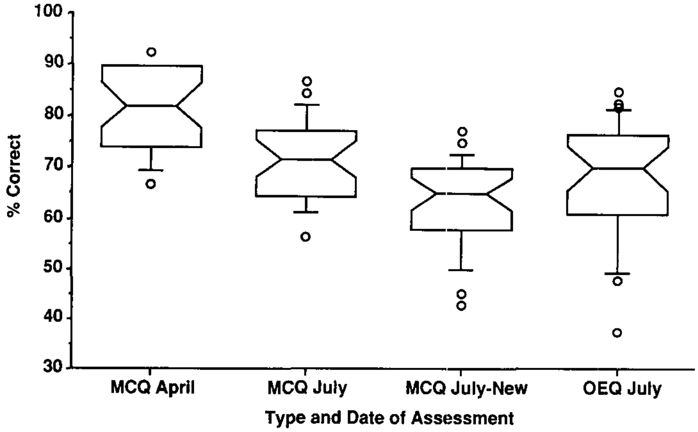

Figure 1. Notched box plots of medical student performance on three measures of recall and retention (MCQs) and one measure of clinical reasoning and synthesis (OEQ). Each box plot is derived from five percentiles that summarize the distribution of all the students' performance. The top and bottom of the box represent the 75th and 25th percentiles, respectively. The middle 50% of students' scores are represented within the box, with the line representing the median. The thickness of each of the boxes graphically represents the degree of homo- or heterogeneity of performance of the middle 50% of the class on each measure.

scores above the 90th percentile or below the 10th percentile. The relatively greater distance below the end of the 'whisker' for the 10th percentile on OEQ-July indicates relatively poorer performance for students on this component than on the MCQ components. These extreme scores explain the larger standard deviation for OEQ-July compared with those of the MCQ components.

The results on the individual components were also correlated (Table 2). Individual student scores on the MCQ-April were plotted against the scores on the MCQ-July (Fig. 2); the correlation coefficient, *r*, was 0·58. The correlation was somewhat higher when the scores on the MCQ-July were compared with those of the MCQ-new July component (Fig. 3), *r* = 0·63. The correlation was weaker between the results of the OEQ-July and those of the aggregate MCQ-July and MCQ-new July (79 questions), *r* = 0·52 (Fig. 4). The correlation coefficients between the OEQ-July results and the individual MCQ components were even lower (0·45–0·49) than that obtained for the aggregate MCQs of July.

Table 2. Pearson product–moment intercorrelations among the various assessments for 33 medical students on measures of recall/retention and the application/integration of clinical knowledge

Assessment†	MCQ-April	MCQ-July	MCQ new July	MCQ aggregate ‡ July
MCQ-April	—			
MCQ-July	0·58	—		
MCQ new-July	0·52	0·63	—	
MCQ aggregate-July	0·61	0·89	0·91	—
OEQ-July	0·42	0·45	0·49	0·52

†Abbreviations and identifications of components as in Table 1.
‡79 questions of MCQ-July and MCQ-new July.

All correlations are statistically significant, *P* < 0·05, two-tailed tests.

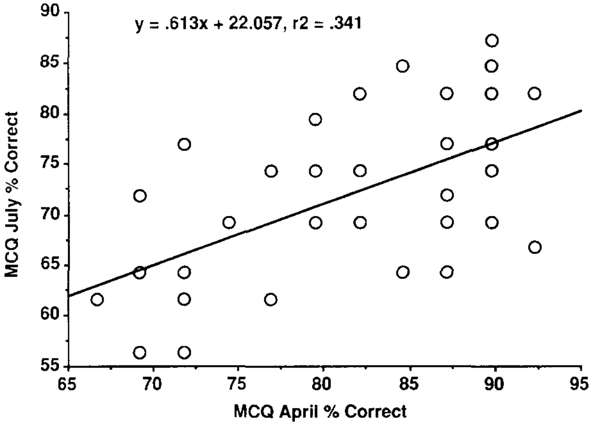

Figure 2. Scattergram depicting the correlation between medical student performance on the identical MCQ examinations in April and again in July.

The inconsistency in student performance across the various components is reflected not only in the above correlations, but is apparent in students who received a passing score (≥70%) on one measure, for example the MCQ-April, and did not consistently attain passing scores on the other measures. The pattern of performance, in fact, appears to be quite random when analysed, with χ^2 (1) < 3·0 and non-significant for all comparisons. For example, 12 of the 29 students (41·4%) who passed MCQ-April failed the identical MCQ-July, and one of four students who failed MCQ-April actually passed the MCQ-July examination! Almost half (47·1%) of the students who passed the OEQ-July did not obtain passing scores on the aggregate of MCQ-July and MCQ-new July.

Discussion

A decline in medical students' recollection of information over time was expected, but, before analysis, the magnitude was uncertain. We showed that, for factual information sought in identical MCQs, the mean scores fell 11·6%, or nine percentile points (from 81 to 72%), over 3 months.

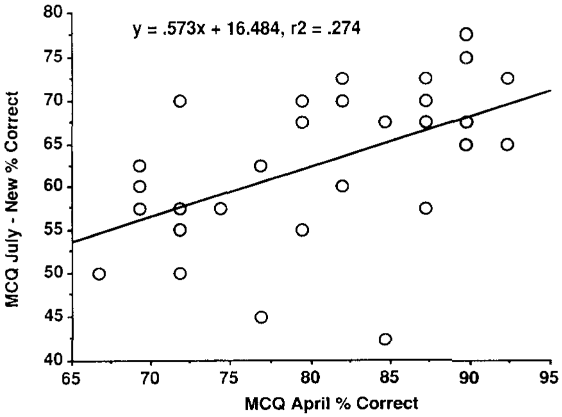

Figure 3. Scattergram depicting the correlation between medical student performance on different MCQ examinations in April and July that covered the same content.

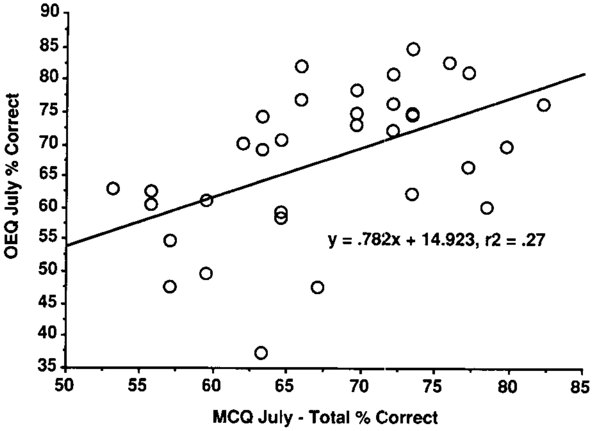

Figure 4. Scattergram depicting the correlation between medical student performance on the identical MCQ aggregate total examination and the OEQ exam, both administered at the same time in July.

When these students were asked to recall similar types of information in MCQs not previously seen, the mean performance was poorer. Without memory aids and without review, the assessment, based on results with new but comparable questions answered 3 months later, showed diminished performance: a decline of 22·4%, or 18 percentile points, from the mean of 81% on the MCQ-April to 63% on the MCQ-new July.

Assuming that the information required was comparable in the two MCQ assessments in July, the significant difference in performances in July requires explanation. It is likely that being exposed to the one set of MCQs in the final examination in April was a stimulus to memory so that the information in the identical questions could be more fully dredged up. If we assume that 70% correct answers is the lower level of competence, then, over a 3-month period, a substantial number of students, 45% on one assessment and 60% of the other, became less than qualified. However, this decline must be tempered with the fact that the students had no opportunity to prepare specifically for these examinations.

The students had no prior experience with clinical OEQs, and the answers required a higher level reasoning than did the MCQs. Nevertheless, the mean score for the OEQs, 68%, was comparable to those attained on the MCQ assessments. However, the range of scores was somewhat broader for the OEQs than for the MCQs, possibly indicating that some students have unusual difficulty in responding to requests for greater application and synthesis of information. Since the OEQs may represent more faithfully the information demanded in the practice of medicine, this type of assessment should be explored further to determine if some students need special help in their education. A somewhat similar test item format using extended matching questions is currently part of the revised examination of the National Board of Medical Examiners (Case *et al.* 1988; National Board of Medical Examiners 1990).

Of interest were the moderate levels of correlations between performances on the assessments. Since the OEQs differed from the MCQs in type of intellectual probe, a correlation coefficient in the range of 0·5 between the results of these two types of assessments was not surprising. However, this observation challenges the concept that a collection of MCQs that require recall of narrowly defined facts constitute an instrument that adequately evaluates student clinical knowledge. The level of correlation, giving only about 25% common variance for the OEQs and MCQs, suggests that the two forms of evaluation assessed different sets of competencies.

The correlation between the results of the

MCQ-July and those of the MCQ-new July was somewhat higher (about 0·6) but still less than expected for tests that cover the same type of information and use the same testing format. Although the information on the two assessments was randomly derived from the same large pool, the emphasis on some areas of clinical medicine differed in the two tests. It is likely that individual students retain information from some areas of learning, because of personal interest or aptitude, better than others, and that this pattern varies among students. Thus, one student may have answered correctly more questions on cardiology and another more on gastroenterology, and depending upon the relative representation of these areas on a given examination, these students would individually do better or worse. The number of questions in the different areas were too few to test this concept in the above data, but future assessments should examine this possible determinant of variability in student performance.

Unanticipated was a correlation coefficient as low as 0·6 between results of answering the same MCQs after an interval of 3 months. A reasonable concept holds that students will forget information at similar rates or that the rates will vary with initial performance, being faster for students who have more difficulty learning. Indeed, we reward students with highest levels of attainment on an examination, not so much for the short-term gain as for the predicted retention of their knowledge that will be useful in the future. From the MCQ assessment, a forecast of knowledge retention carries substantial unreliability, at least over a 3-month period. For example, one student scored 92% when the MCQ-final was taken in April (at the top of his cohort of students), but only 67% in July (in the lower third of the cohort). Typically, students cram over the few days and hours before an examination, and the short-term memory may not translate proportionally into retention of information for each student. We must ask if an educational system that promotes and rewards such study patterns is rational.

Our testing retention of knowledge over months lays a basis for determining if a change in methods of teaching and learning attains improvement in student comprehension in a meaningful way. A rise in test scores on examinations given at the completion of a learning session provides an immediate appraisal but is an insufficient indicator of success if recall over a prolonged period is the goal. Therefore, those who modify curricula should assess progress by comparing results of testing students before and after the intervention, but also immediately following and at a time more remote from the educational experience under analysis.

In summary, new data are provided on medical student retention and recall of knowledge. The students retained information over a 3-month period in a somewhat unpredictable pattern. Students' retention of clinical knowledge should receive more attention when considering rewards for student performance and when evaluating change in methods of teaching and learning; an assessment of retention such as described in this communication should be useful. The results of students answering open-ended types of questions requiring application and synthesis of knowledge differ sufficiently from scores on multiple choice questions so that educators should at least explore instruments of evaluation other than standard multiple choice questions.

Acknowledgements

The authors are indebted to the medical students who cooperated in this study. Mrs Pamela Gibb was most helpful in organizing the administration of the test components, and the Medical School Academic Information Systems and Ms Gloria Picasso were invaluable in the scoring of the tests. Mrs Karen Grahl gave expert typing and editing aid with the manuscript.

References

Anderson J. & Graham H. (1980) A problem in medical education: is there information overload? *Medical Education* **14**, 4–7.

Arkin H. & Colton R.R. (1963) In: *Tables for Statisticians*, 2nd edn, pp. 63. Barnes & Noble, New York.

Bordage G. (1987) The curriculum: overloaded and too general? *Medical Education* **21**, 183–8.

Case S.M., Swanson D.B. & Stillman P.L. (1988) Evaluating diagnostic pattern recognition: the psychometric characteristics of a new item format. In:

Proceedings of the Twenty-Seventh Annual Conference on Research in Medical Education, pp. 3–8. Association of American Medical Colleges, Washington, DC.

Cleveland W.S. (1985) *The Elements of Graphing Data*. Wadsworth, Monterey, California.

Covell D.G., Ulman G.C. & Manning P.R. (1985) Information needs in office practice: are they being met? *Annals of Internal Medicine* **103**, 596–9.

National Board of Medical Examiners (1990) *Part I/II/III Examination Guidelines and Sample Items*. National Board of Medical Examiners, Philadelphia, Pennsylvania.

Received 21 *November* 1991; *editorial comments to authors* 25 *February* 1992; *accepted for publication* 1 *May* 1992