Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning

Ramraje SN¹, Sable PL²

¹. Department of Pathology, Grant Medical College, Mumbai, India
². Department of Preventive and Social Medicine, G.S. Medical College, Mumbai, India

Abstract

Background
People forget much of what they learn, therefore students could benefit from learning strategies that yield long-lasting knowledge. Yet surprisingly, little is known about how long-term retention is most efficiently mastered. We studied the value of teacher made in class tests as learning aids and compared two types of teacher-made tests (multiple choice and short-answer tests) with a no test (control) to determine their value as aids to retention learning.

Method
The study was conducted on two separate batches of medical undergraduate students. This study compared two types of tests [multiple choice questions (MCQs) and short answer questions (SAQs)] with a no test (control) group. The investigation involved initial testing at the end of the lecture (post instruction), followed by an unannounced delayed retention test on the same material three weeks later. The unannounced delayed test comprising of MCQs and SAQs on the same material was given three weeks later to all the three groups.

Results
In batch I, the MCQ group had a higher mean delayed retention score of 10.97, followed by the SAQ group (8.42) and the control group (6.71). Analysis of variance (ANOVA) test and least significance difference (LSD) post hoc test revealed statistically significant difference between the means of the three groups. Similar results were obtained for batch II.

Conclusion
Classroom testing has a positive effect on retention learning; both short-answer and multiple-choice tests being more effective than no test in promoting delayed retention learning, however, multiple-choice tests are better.

Key Words
Initial testing, delayed retention tests, retention learning

What this study adds:
1. Much discussion is already ongoing in this field. There are plenty of reliable references. Although studies have been published analysing knowledge retention in other fields of study, a literature search failed to yield any published research on this topic in the field of medical education.

2. Medical school curriculum is vast and information taught is difficult to retain in future years or in clinical practice. Classroom testing will definitely help in retention of information. Our data will add to the knowledge pool.
Background
Cognitive learning is best assessed with traditional classroom tests and it is important to maximise the learning value of the time spent in testing regardless of the type of tests used. Testing assesses what students have learned and also improves long-term memory. The act of taking a test helps move information from the short-term memory to a deeper level. Whether this effect is caused by the mere fact that taking a test provides one additional opportunity for rehearsal or there is some unknown factor (such as the kinesthetic act of writing the answers) at work has not been determined yet. Students remember information which has been tested more, than non-tested information, even without additional study of the to-be-tested material hence the need for undertaking such a study.

Method

Aims and objectives
1. To discover if initial testing post instruction helps retention learning.
2. To improve delayed retention learning by introducing post instruction multiple assessment systems like MCQs and SAQs.
3. To compare the effect of post instruction MCQ and SAQ tests on delayed retention learning.

Ethics committee approval was taken.

Definition of Terms:
“Initial testing” means testing which occurs at the time of instruction or immediately thereafter.

“Delayed retention tests” are research instruments which are administered two or more weeks after instruction and initial testing to measure retained knowledge.1-3

“Retention learning” means learning which lasts beyond the initial testing and it is assessed with tests administered two or more weeks after the information has been taught and tested.4

A delay period of three weeks was used in this study.

Design
1. Comparative study.
2. Target group: Two batches of students of Second MBBS (Pathology) class
3. Sample size: Batch I (n1 =84), divided into three groups:
   a) multiple-choice test (Group A, n=35);
   b) short-answer test (Group B, n=35);
   and
   c) no test (Control, Group C, n=14).

Batch II (n2 =72), divided into three groups:
   a) multiple-choice test (Group A, n=30);
   b) short-answer test (Group B, n=30);
   and
   c) no test (Control, Group C, n=12).

Total sample size= n1 + n2 =156. The study was conducted on two separate batches of medical undergraduate students.

This study compared two types of tests (multiple choice and short-answer) with a no test (control) group.

At the beginning of the lecture it was announced that an experimental study was being conducted. Participation was voluntary and participants should give their best.

All instructional and testing procedures were done in the same room, by the same teacher thus helping to control environment variables.

Process
The investigation involved initial testing at end of lecture (post instruction), followed by an unannounced delayed retention test on the same material three weeks later. The multiple-choice tests were a 10 item test and had four response alternatives. The short answer versions were
identical to the multiple-choice tests, however there were no alternatives and brief answers were required. The same information was reflected in both tests.

An initial test of MCQs and SAQs was given at the end of lecture (post instruction) to the MCQ and SAQ group respectively. The control group was not given any test initially.

An unannounced delayed test comprising of MCQs and SAQs on the same material was given three weeks later to all the three groups.

**Results**

Only the delayed test scores were statistically analysed. Statistical analysis was done using SSPS (Version 17). Mean test scores for all the groups were calculated. Statistical difference between different groups was calculated using ANOVA and post hoc comparisons.

The total students (n=84) participating in the project were divided into three groups: MCQ, SAQ and control group. The MCQ group (n=35) had a higher mean delayed retention score of 10.97, followed by the SAQ group (8.42) and the control group (6.71). The total mean retention test score was 9.20 (n=84) (see Table 1).

The GLM procedure of ANOVA was used to discover the difference between the three groups. It showed a statistically significant difference between the means of delayed retention test scores of the MCQ, SAQ and control group (p < 0.05) (see Table 2).

LSD post hoc test (post comparison method) was used to calculate the difference between the means of the three groups. Statistical analysis revealed significant difference between the groups (p value < 0.05). Hence, the mean score of the MCQ group is higher than the SAQ group as well as the control group (see Table 3).

In batch I, the MCQ group had a higher mean delayed retention score of 10.97, followed by the SAQ group (8.42) and the control group (6.71). ANOVA test and LSD post hoc test revealed a statistically significant difference between the means of the three groups. Similar results were obtained for batch II (see Figure 1).

The total students (n=72) participating in the project were divided into three groups: MCQ, SAQ, and control group. The MCQ group (n=30) had a higher mean delayed retention score of 10.20, followed by the SAQ group (8.16) and the control group (5.66). The total mean retention test score was 8.59 (n=72) (see Table 4).

The GLM procedure of ANOVA was used to discover the difference between the three groups. It showed a statistically significant difference between the means of delayed retention test scores of the MCQ, SAQ, and control group (p < 0.05) (see Table 5).

LSD post hoc test (post comparison method) was used to calculate the difference between the means of the three groups. Statistical analysis revealed significant difference between the groups (p value < 0.05). Hence, the mean score of the MCQ group is higher than the SAQ group as well as the control group (see Table 6).

**Figure 1: Mean delayed retention test scores**

![Figure 1: Mean delayed retention test scores](image-url)
Discussion

Testing in education, i.e. time devoted to testing and the effects of testing, are becoming very important. Testing assesses students’ knowledge, and it improves long-term memory. This retention of course-related information resulting from test-taking is called the ‘testing effect’. The testing effect has been studied in laboratory and classroom settings, dating back to the early 20th century and continuing to the present day.

Much research on testing has been done in recent years and has been focused on standardised tests. In schools, teacher-made tests are used for evaluation. Data is available on the quality of these tests creating doubt as to whether teachers are able to perform evaluation effectively or not. But teacher-made tests are important in the classroom for evaluation of taught material. Teacher-made tests are an important part of the educational system and should be researched.

Evaluation by teacher-made tests in schools is an important and needed part of the educational system and a crucial area for research.

The use of short-answer tests is popular. Short-answer tests are used more often and are more effective with lower level types of learning. They have certain qualities, such as they are relatively easy to prepare and scored more quickly. However, they are not objective like multiple-choice tests and cannot adequately test students who have read the subject well. Despite this inadequacy, they are useful because many teachers cannot make good multiple-choice items.

Multiple-choice tests promote retention. Literature has shown that only test-taking enhances retention learning. The announcement of an upcoming test did not have a positive effect on retention learning. Hence increased studying for a slated test did not result in better retention, only the act of taking the test did.

Conclusion

Examinations and assignments are the two most commonly used approaches to assessment in education. Since testing consumes a large amount of students’ and teachers’ time, the evaluation process should be made more efficient.
Classroom testing has a positive effect on retention learning; both short-answer and multiple-choice tests being more effective than no test in promoting delayed retention learning, however, multiple-choice tests are better.

Multiple-choice tests are good in promoting retention learning of the information actually contained in the immediate post-tests as compared to the short-answer tests because the correct answer to each item is provided along with the distracters in the multiple choice items. The advantage of short-answer tests is that students do not have to choose from a set of responses. Teachers who find constructing multiple-choice items difficult could make do with short-answer items which are relatively easy to prepare. Their usefulness in the promotion of retention learning should be researched. It is recommended to choose the type of test which meets the educational objectives best. Continued research must be conducted and the best ways to test sought so as to maximise retention of important learning in all disciplines.

References

Tables

Table 1: Means, Standard deviations and sample sizes.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ</td>
<td>35</td>
<td>10.9714</td>
<td>1.12422</td>
<td>.19003</td>
<td>10.5852</td>
<td>9.00</td>
<td>11.3576</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAQ</td>
<td>35</td>
<td>8.4286</td>
<td>1.06511</td>
<td>.18004</td>
<td>8.0627</td>
<td>6.00</td>
<td>12.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>14</td>
<td>6.7143</td>
<td>1.13873</td>
<td>.30434</td>
<td>6.0568</td>
<td>5.00</td>
<td>8.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>9.2024</td>
<td>1.94985</td>
<td>.21275</td>
<td>8.7792</td>
<td>5.00</td>
<td>13.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: GLM procedure of Analysis of Variance (ANOVA)

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig. (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>217.160</td>
<td>2</td>
<td>108.580</td>
<td>89.380</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>98.400</td>
<td>81</td>
<td>1.215</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>315.560</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: LSD multiple comparisons

<table>
<thead>
<tr>
<th>(I) Test category</th>
<th>(J) Test category</th>
<th>Mean Difference</th>
<th>Std. Error</th>
<th>Sig. (p value)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(I-J)</td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>MCQ</td>
<td>SAQ</td>
<td>2.54286</td>
<td>.26347</td>
<td>.000</td>
<td>2.0186</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.25714</td>
<td>.34854</td>
<td>.000</td>
<td>3.5637</td>
</tr>
<tr>
<td>SAQ</td>
<td>MCQ</td>
<td>-2.54286</td>
<td>.26347</td>
<td>.000</td>
<td>-3.0671</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>1.71429</td>
<td>.34854</td>
<td>.000</td>
<td>1.0208</td>
</tr>
<tr>
<td>Control</td>
<td>MCQ</td>
<td>-4.25714</td>
<td>.34854</td>
<td>.000</td>
<td>-4.9506</td>
</tr>
<tr>
<td></td>
<td>SAQ</td>
<td>-1.71429</td>
<td>.34854</td>
<td>.000</td>
<td>-2.4078</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.
Table 4: Means, Standard deviations and sample sizes.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAQ</td>
<td>30</td>
<td>8.1667</td>
<td>.91287</td>
<td>.16667</td>
<td>7.8258</td>
<td>8.5075</td>
<td>7.00</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>5.6667</td>
<td>.88763</td>
<td>.25624</td>
<td>5.1027</td>
<td>6.2306</td>
<td>4.00</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>8.5972</td>
<td>1.86638</td>
<td>.21995</td>
<td>8.1586</td>
<td>9.0358</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Table 5: GLM procedure of Analysis of Variance (ANOVA)

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig. (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>185.686</td>
<td>2</td>
<td>92.843</td>
<td>103.940</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>61.633</td>
<td>69</td>
<td>.893</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>247.319</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: LSD multiple comparisons

<table>
<thead>
<tr>
<th>(I) Test category</th>
<th>(J) Test category</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig. (p value)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>MCQ</td>
<td>SAQ</td>
<td>2.03333*</td>
<td>.24403</td>
<td>.000</td>
<td>1.5465</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>4.53333*</td>
<td>.32282</td>
<td>.000</td>
<td>3.8893</td>
</tr>
<tr>
<td>SAQ</td>
<td>MCQ</td>
<td>-2.03333*</td>
<td>.24403</td>
<td>.000</td>
<td>-2.5202</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>2.50000*</td>
<td>.32282</td>
<td>.000</td>
<td>1.8560</td>
</tr>
<tr>
<td>Control</td>
<td>MCQ</td>
<td>-4.53333*</td>
<td>.32282</td>
<td>.000</td>
<td>-5.1773</td>
</tr>
<tr>
<td></td>
<td>SAQ</td>
<td>-2.50000*</td>
<td>.32282</td>
<td>.000</td>
<td>-3.1440</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.