Abstract - The paper discusses the usage of multiple-choice and open-ended constructed response questions for measuring the performance of undergraduate students in computer science courses. A goal of this research is to investigate if “guessing” plays a significant role in multiple-choice answering. An experiment was constructed in two academic institutions over a two-semester period involving students in four introductory programming classes. The quizzes consisted of pairs of similar questions, where each pair contained a multiple-choice and an open-ended constructed response question of equal difficulty and weight.

Index Terms - Assessment, Multiple-choice questions, Open-ended constructed response questions.

INTRODUCTION

In the past, much research has been done in the field of educational methodology. Some has focused on finding objective and reliable examination techniques for measuring student performance. This paper investigates multiple-choice (MC) and open-ended constructed response (OECR) questions. OECR question consists of a problem that has to be addressed by developing an answer, rather than selecting a provided choice [6]. The advantages and short-comings of both testing tools are well known. MC questions are “well suited for assessing students’ recall of factual knowledge and their ability to solve problems that have unambiguous right answers” [5]. In addition, they are reliable and objective measures [7]. Also, according to [1], MC tests are not time consuming for students and graders, can cover a large portion of material, and last but not least, are effective preparation and self-assessment tools. However, on the down side [4], MC questions encourage guessing and trivialize educational measurement [3]. According to [2, 5], OECR questions are appropriate for measuring thinking and problem solving strategies.

PROBLEM DEFINITION

In general, the different methods of testing are not homogeneous and assigning equal weights to both types of questions and including them in the same quantity may be a potential problem. This issue has been discussed in detail in [7]. The authors conclude that MC questions are the most reliable and objective tool for testing, whereas OECR are more subjective. When combined with MC questions, they (OECR) introduce “noise” in the scoring and reduce the overall reliability of the examination. The exceptions to this rule, however, also exist. These exceptions are apparent in some fields of mathematics and computer science (CS), where no statistically significant difference in the reliability has been discovered when both types of questions are mixed. Therefore, using both MC and OECR questions is beneficial because it offers diversity of testing methods and preserves the reliability of the test.

In our research, use of MC and OECR questions is further developed in the context of some introductory CS courses. The main problems addressed in this case could be formulated as follows:

- Investigate if “guessing” plays a significant role in MC tests in introductory programming courses;
- Determine if there is a significant difference between the number of MC and OECR correct responses given by the better students compared to the rest of the students.

The research supporting the solutions to the above problems consists of several steps. First, the experiment is described in the Methodology section. Then, the Logic Model section develops the research hypothesis and the Descriptive and Empirical Results section shows the statistical results from the tests. Finally, the Conclusion section of the paper contains the summarized findings.

METHODOLOGY

In search of the solution to the two problems mentioned above, an in-class experiment was conducted over two semesters, and involved students from two undergraduate institutions who voluntarily elected to take an introductory CS course (CS I). The study was constructed in the following way. A bank of paired questions was developed that included both MC and OECR questions. A pair of sample questions is described in the Methodology section. Then, the Logic Model section shows the statistical results from the tests. Finally, the Conclusion section of the paper contains the summarized findings.

Abstract - The paper discusses the usage of multiple-choice and open-ended constructed response questions for measuring the performance of undergraduate students in computer science courses. A goal of this research is to investigate if “guessing” plays a significant role in multiple-choice answering. An experiment was constructed in two academic institutions over a two-semester period involving students in four introductory programming classes. The quizzes consisted of pairs of similar questions, where each pair contained a multiple-choice and an open-ended constructed response question of equal difficulty and weight.

Index Terms - Assessment, Multiple-choice questions, Open-ended constructed response questions.

INTRODUCTION

In the past, much research has been done in the field of educational methodology. Some has focused on finding objective and reliable examination techniques for measuring student performance. This paper investigates multiple-choice (MC) and open-ended constructed response (OECR) questions. OECR question consists of a problem that has to be addressed by developing an answer, rather than selecting a provided choice [6]. The advantages and short-comings of both testing tools are well known. MC questions are “well suited for assessing students’ recall of factual knowledge and their ability to solve problems that have unambiguous right answers” [5]. In addition, they are reliable and objective measures [7]. Also, according to [1], MC tests are not time consuming for students and graders, can cover a large portion of material, and last but not least, are effective preparation and self-assessment tools. However, on the down side [4], MC questions encourage guessing and trivialize educational measurement [3]. According to [2, 5], OECR questions are appropriate for measuring thinking and problem solving strategies.

PROBLEM DEFINITION

In general, the different methods of testing are not homogeneous and assigning equal weights to both types of questions and including them in the same quantity may be a potential problem. This issue has been discussed in detail in [7]. The authors conclude that MC questions are the most reliable and objective tool for testing, whereas OECR are more subjective. When combined with MC questions, they (OECR) introduce “noise” in the scoring and reduce the overall reliability of the examination. The exceptions to this rule, however, also exist. These exceptions are apparent in some fields of mathematics and computer science (CS), where no statistically significant difference in the reliability has been discovered when both types of questions are mixed. Therefore, using both MC and OECR questions is beneficial because it offers diversity of testing methods and preserves the reliability of the test.

In our research, use of MC and OECR questions is further developed in the context of some introductory CS courses. The main problems addressed in this case could be formulated as follows:

- Investigate if “guessing” plays a significant role in MC tests in introductory programming courses;
- Determine if there is a significant difference between the number of MC and OECR correct responses given by the better students compared to the rest of the students.

The research supporting the solutions to the above problems consists of several steps. First, the experiment is described in the Methodology section. Then, the Logic Model section develops the research hypothesis and the Descriptive and Empirical Results section shows the statistical results from the tests. Finally, the Conclusion section of the paper contains the summarized findings.

METHODOLOGY

In search of the solution to the two problems mentioned above, an in-class experiment was conducted over two semesters, and involved students from two undergraduate institutions who voluntarily elected to take an introductory CS course (CS I). The study was constructed in the following way. A bank of paired questions was developed that included both MC and OECR questions. A pair of sample questions is described in the Methodology section. Then, the Logic Model section shows the statistical results from the tests. Finally, the Conclusion section of the paper contains the summarized findings.
two hard to spot. Overall, the number of questions used in the present experiment was 44.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>SAMPLES OF PAIRED MC AND OECR QUESTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MC</strong></td>
<td><strong>OECR</strong></td>
</tr>
<tr>
<td>A1. What is the output of the following program fragment?</td>
<td>B1. What is the output of the following program fragment?</td>
</tr>
<tr>
<td><code>cout &lt;&lt; 10/4*5 + 17%4 - 4/5;</code></td>
<td><code>cout &lt;&lt; 12/5*4 + 17%3 - 5/6;</code></td>
</tr>
<tr>
<td>a. 11</td>
<td>e. none of the above</td>
</tr>
<tr>
<td>b. 14</td>
<td>f. 13.5</td>
</tr>
<tr>
<td>c. 13.5</td>
<td>d. 12.7</td>
</tr>
</tbody>
</table>

A total of 62 people were involved in the experiment. A division of the students was done according to their final course grades. Students, whose final grade was equal or above the class average, were categorized as *A-students*, and the ones who earned a score lower than the averages were labeled as *B-students*.

**Logic Model**

To investigate the problems described in the previous section of the paper, the following hypotheses concerning difference of means were tested:

**H1:** Test Objectivity Hypothesis \([H_a: \mu_A > \mu_B]\) To have reliable results, the administered tests have to be objective i.e. on average A-students should perform significantly better than B-students.

**H2:** MC and OECR Difficulty Hypothesis \([H_a: \mu_{MC} > \mu_{OECR}]\) If H1 is valid, then we expect, on average, all students to have answered MC questions correctly more often than OECR questions, due to their perceived difficulty, and not due to their actual difficulty.

**H3:** B-Student Response Hypothesis \([H_a: \Delta_{MC} < \Delta_{OECR}]\) If H2 is false, then students do not differentiate between the two types of questions. This suggests that no “guessing” is used by B-students. If MC questions are answered correctly more often than OECR questions, then B-students are not solving the problem but rather selecting a response that seems to be correct.

**Descriptive and Empirical Results**

Each of the above hypotheses was tested at the 5%-level of significance \((z_{0.025} = \pm 1.96)\) and the resulting value was a z-score. For H1 the z-score of 10.07 suggested that the quizzes were valid, and indeed, A-students performed better on average. For H2, the calculated z-score was 7.72, which once again confirmed our expectations, i.e. on average, both type of students score better on MC questions; therefore there is a perceived difference within the pair. Finally, for H3, a z-score of -7.30 suggested that A-students solve both types of problems better and also that B-students select correctly an MC answer but have difficulty solving the corresponding OECR, therefore “guessing” is involved.

In addition to the above tests, the following simple linear regression models were tested:

\[
FG = \alpha_1 + \beta_1Ave(MC) \quad \text{and} \quad FG = \alpha_2 + \beta_2Ave(OECR)
\]

where \(FG\) is the final grade of a given student and \(Ave(MC)\) and \(Ave(OECR)\) are the averages that the same student received for the corresponding types of questions.

The above models yielded adjusted \(r^2\) (coefficient of determination) of 19.93% and 64.38% respectively. These percentages indicate the variability in final grades that is explained by the respective variable, i.e. the two types of questions. Consistent with the initial hypotheses, the OECR questions exhibited a higher correlation, thus they explain to a higher degree the variability in the final grades. In the current context, this means that the OECR questions are better measurement of knowledge.

**Conclusion**

The statistical analysis performed on the sample data confirmed our initial expectations that OECR questions are equally objective as MC questions and actually evaluate students’ knowledge more accurately.

Overall, students answered MC questions much better than OECR questions, despite their equivalent difficulty. The average number of OECR questions that an A-student answered correctly was significantly higher compared to the correct responses of a B-student. Therefore, “guessing” took place. In addition, the OECR questions explained significantly better the variability of the final grades compared to MC questions.

To complete the work, more data has to be collected in order to increase the validity of the results. Based on the initial finding that OECR questions are more accurate in measuring knowledge, an empirical model will be developed, which will examine the predictive power of OECR questions with regard to the final grades of the students.

**References**


