Assessment of learning outcomes: validity and reliability of classroom tests

Maizam Alias

Kolej Universiti Teknologi Tun Hussein Onn
Johor Darul Takzim, Malaysia

ABSTRACT: Teachers in engineering routinely design and administer classroom tests to their students for decision-making purposes. To be of real value in decision-making, these tests must be valid and reliable. Test validity and reliability may be achieved by taking a systematic approach to test design. In this article, the author proposes and discusses measures that teachers could take in order to help them enhance the validity and reliability of their classroom tests, taking examples from the teaching and learning of structural design in civil engineering. A sample spreadsheet in Excel is provided that may be used by teachers to get a quick estimate of the reliability of their classroom tests.

INTRODUCTION

Assessment entails the systematic gathering of evidence to judge a student’s demonstration of learning. Teachers can then judge whether a student has learned what they are expected to learn by securing valid and reliable information through various assessment methods. The assessment method chosen would depend on the learning domain that is of interest, which could be the cognitive, affective or psychomotor domains [1]. Examples of learning in the three domains are given in Table 1.

Table 1: Examples of learning in the cognitive, affective and psychomotor domains [2].

<table>
<thead>
<tr>
<th>Domain</th>
<th>Learning Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>Choosing to learn from own and other peoples’ experiences by ensuring similar</td>
</tr>
<tr>
<td></td>
<td>mistakes are not repeated, and incorporating past successes into current</td>
</tr>
<tr>
<td></td>
<td>design where appropriate</td>
</tr>
<tr>
<td>Affective</td>
<td>Manoeuvring a computer mouse to produce the desired effect on the computer</td>
</tr>
<tr>
<td></td>
<td>screen when using a Computer Aided Design package for drawing</td>
</tr>
<tr>
<td>Psychomotor</td>
<td>Manoeuvring a computer mouse to produce the desired effect on the computer</td>
</tr>
<tr>
<td></td>
<td>screen when using a Computer Aided Design package for drawing</td>
</tr>
</tbody>
</table>

Assessment tools that can be used to assess learning include an achievement test for the cognitive domain, an attitude questionnaire for the affective domain and a checklist for the psychomotor domain. In this article, the author focuses on the assessment of learning in the cognitive domain since this is the most frequently assessed domain for classroom learning.

For the purpose of this article, the term classroom test will be used in place of achievement test to emphasise its classroom application. A classroom test is defined as any set of questions that is specifically designed by teachers to measure an identified learned capability (or set of learned capabilities) and administered by teachers to their students in classroom setting.

Classroom tests are routinely designed and administered by teachers to assess students’ learned capabilities, and output from classroom tests are often used to support decision-making, such as in giving grades to students or assigning students to remedial classes. In order to be of real value in decision-making, these classroom tests must possess two important characteristics, namely: validity and reliability. A discussion of some of the issues that teachers need to look into, plus some of the practical measures that teachers can take to enhance the validity and reliability of their classroom tests, is presented in this article.

VALIDITY AND RELIABILITY

Validity and reliability are two quality indicators for classroom tests. Validity refers to the degree to which a test is measuring what it is supposed to measure, while reliability is an indication of the consistency between two measures of the same test [3]. A test may be highly reliable but not necessarily valid, but a highly valid test is usually reliable.

Types of Validity

There are two types of validity that are most relevant to classroom tests, namely: face validity and content validity [3]. Face validity refers to the appearance of a test that looks like it is measuring what it is supposed to measure. Face validity is essential in ensuring that test-takers persevere and try their best on a test. A test that appears to be other than what it claims to be measuring – without face validity – may dissuade students from persevering with the test. Therefore, ascertaining whether a test possesses face validity does not require the opinion of an expert.
In contrast to face validity, a claim of content validity requires affirmation from an expert. The expert should look into whether the test content is representative of the skills that are supposed to be measured. This involves looking into the consistency between the syllabus content, the test objective and the test contents. If the test contents cover the test objectives, which in turn are representative of the syllabus, it could be said that the test possesses content validity.

For example, an English test paper is definitely not a valid instrument for measuring mathematical skills. An algebra test, on the other hand, is to a certain degree a valid measuring tool for mathematical skills because the ability to do algebra is an indicator of a person’s mathematical skills. Still, the algebra test is not highly valid because mathematical skills are not confined to the ability to solve algebra problems alone. Therefore, to make the test paper highly valid, other indicators of mathematical skills must be included in the test paper. If the test is valid and reliable, a student who shows good mathematical skills on that particular test should also do equally well on other mathematical tests of similar content and objective. In other words, students do not just possess skills to solve the mathematical items that are given in the specific test.

To summarise, the decision on what to include in a test paper will depend on what the content of the syllabus is, as well as what the test objectives are. It is of utmost importance for teachers to appreciate that the degree of test validity depends on the test’s coverage of the necessary objectives, which, in turn, depends upon the syllabus.

Types of Reliability

There are three types of reliability that are most relevant to classroom tests, namely: internal consistency, inter-scorer and intra-scorer reliability [3]. Internal consistency refers to the consistency of objectives among the items of a test. For example, consider a 10-item mathematical test that is supposed to measure students’ ability to solve two variable algebra problems. In this case, the question of internal consistency refers to the answer to the question: are the 10 items measuring the same skill (ie students’ ability to solve two variable algebra problems), or are the different items measuring something else entirely or others besides the stated objective?

Inter-scorer reliability refers to the consistency between the marks given by different teachers. Doubts upon inter-scorer reliability could arise when the same quality of answers is given different scores by different teachers. On the other hand, intra-scorer reliability refers to marks given by the same teacher on different occasions. An example of intra-scorer reliability at stake is when a teacher gets tired of marking and starts to give lower marks as time goes on. Consistent grading is essential in order to ensure the reliability of test scores.

Scorer reliability can be improved by a marking scheme or a scoring rubric that is prepared in advance and used to assist teachers in scoring answer scripts.

So what can be done to develop a valid and reliable test?

In order to achieve a certain degree of validity and reliability, the assessment and evaluation process has to be looked at in its totality, and the factors that may affect validity and reliability need to be identified. Typical activities in the classroom assessment and evaluation process are as follows:

- Deciding on a test’s objectives;
- Designing and developing a test;
- Evaluating the test;
- Administering the test.

At each stage, something could be carried out to enhance the validity and reliability of a test. The discussion below is based on these activities, starting with the decision on test objectives.

Deciding on a Test’s Objectives

Determining a test’s objective(s) is the first step in a test’s construction process. The test objective is the criterion that will be used in order to judge whether a test is sufficiently valid or not. This objective is general in nature, which can be represented by a set of more specific objectives or item objectives to be identified through an analysis of the syllabus. Three examples of test objectives are as follows:

- Measure final year students’ ability to solve calculus problems;
- Measure first year students’ understanding of concepts and procedures in circuit design;
- Measure civil engineering students’ ability to solve structural design problems that demand spatial visualisation abilities.

The key phrases, calculus problems, concepts and procedures in circuit design and structural design problems that demand spatial visualisation abilities determine the scope/content, item format and length of the test. If teachers as test designers are not clear of their objectives, they may end up measuring something other than what they wish to measure, that is, having an instrument that lacks validity.

Designing and Developing a Test

Designing a test is indeed a complex task. Many questions need to be asked and a lot of decisions need to be made at a number of stages along the way so as to increase the chances of meeting the criteria of a good test. In other words, the design and development stage of a classroom test holds the most possibilities for ensuring test validity and reliability. One of the most important steps in designing a test is constructing a table of specifications.

Constructing a Table of Specifications

As mentioned earlier, validity is concerned with how good a match is between what a test is supposed to measure and what it actually measures. Adequate content coverage is an important element of content validity. Constructing a table of specifications is one of the practical means towards achieving this objective.

A table of specification is a two-way table with the cognitive emphasis on the first row and contents in the first column. It can be constructed using a two level analysis described below.

The first level of analysis covers the following:

- Construct a two-way table with a list of topics in the first column and a list of cognitive emphases in the first row;
- Identify the topics/sub-topics and the corresponding cognitive emphasis to be tested;
- Estimate the percentage allocation for each topic.
The second level of analysis incorporates the following:

- Choose the appropriate item format (multiple choice (MC)/structured question (SQ)/long question or essay (LQ), etc) for the specific objective;
- Determine the number of questions for each specific objective;
- Check that the marks for each topic match the total weightage allocated.

An example of a table of specifications is given in Table 2 (taken from Alias [5]). This table of specifications forms the basis for designing a one-hour test on factual, concept and procedural knowledge of a beam, column and slab in structural design. The test covers the six cognitive skills as identified in Bloom’s taxonomy [4]. By constructing a table of specifications, teachers are forced to consider in a systematic manner the learning objectives that need to be covered by their tests. Therefore, a test that possesses content validity is ensured.

**Deciding on Item Format**

The choice of item format depends upon several factors, with the item objective being the most important. Apart from the item objective, ease of scoring, ease of administration and the content coverage are also relevant factors in deciding on the item format. Common item format includes multiple choice, essay, structured and true/false. Certain formats are more suitable than others in meeting the item objective. For example, an essay question allows a student answering the question to demonstrate his/her depth of knowledge. On the other hand, essay questions are relatively more time consuming to mark and need greater efforts to ensure inter-scorer and intra-scorer reliability. In brief, when designing test items, a teacher has to balance the needs of the test objectives while also considering other practical constraints that may contribute to lower (or enhance) the test’s validity and reliability.

**Constructing Items**

Once the format is chosen, the teacher has to construct the test item. The language used, the context of the problem and ease of understanding can affect the reliability and validity of the test as a whole. Some common mistakes that contribute to a reduction in validity and reliability include the following:

- Ambiguous questions, i.e. questions that have multiple interpretations;
- Bias items, such as items that are favourable certain social backgrounds;
- The use of jargon that is not familiar to the target group.

Avoiding these mistakes should enhance the validity and reliability of the test scores.

**Test Documentation**

Once test items are constructed, they need to be assembled and documented for record and reproduction purposes. Apart from that, test documentation is also extremely important for evaluation and refinement purposes.

**Test Evaluation**

Test evaluation can be formative or summative. Formative evaluation can be carried out by administering a newly drafted test to a small group of students that is similar to the target group. The items are then analysed and the reliability of the scores are estimated. The results of this evaluation can be utilised to refine any test items found to be inadequate. A summative evaluation is performed in a similar manner but is based on the actual target group of students. In this case, test refinements can only be of benefit to the next batch of students.

**Item Analysis**

During the item analysis stage, a teacher can estimate the item quality indicators, specifically the item total correlation (ITC), which indicates the consistency of items, the difficulty index (Diff P) and items discrimination index (Disc D). These quality indicators can alert teachers to poor items. For example, an item that has a very high Diff P may be too easy. A Diff P of 0.5 is suitable for a norm-referenced test. An item that has low Disc D

<table>
<thead>
<tr>
<th>CONTENT</th>
<th>Knowledge &amp; Comprehension</th>
<th>Application &amp; Analysis</th>
<th>Synthesis &amp; Evaluation</th>
<th>Total (Content)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam design</td>
<td>10%</td>
<td>20%</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td>-Load assessment</td>
<td>-</td>
<td>MC (4 @ 1 mark each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Structural behaviour</td>
<td>SQ (2 @ 5 mark each)</td>
<td>SQ (2 @ 4 mark each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Design</td>
<td></td>
<td>SQ (2 @ 4 mark each)</td>
<td>LQ (2 @ 10 marks each)</td>
<td></td>
</tr>
<tr>
<td>Column design</td>
<td>10%</td>
<td>10%</td>
<td>0%</td>
<td>20%</td>
</tr>
<tr>
<td>-Axis of rotation</td>
<td>MC (2 @ 1 mark each)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Projection</td>
<td>MC (2 @ 1 mark each)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Plane of bending</td>
<td>MC (2 @ 1 mark each)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Effective height</td>
<td>MC (2 @ 2 marks each)</td>
<td>MC (2 @ 2 marks each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Structural behaviour</td>
<td></td>
<td>SQ (1 @ 6 marks each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slab design</td>
<td>0%</td>
<td>20%</td>
<td>10%</td>
<td>30%</td>
</tr>
<tr>
<td>-Load assessment</td>
<td>-</td>
<td>SQ (2 @ 4 marks each)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Structural Behaviour</td>
<td>SQ (2 @ 6 marks each)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Design</td>
<td>-</td>
<td>LQ (1 @ 10 marks each)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total (Cognitive Emphasis) 20% 50% 30% 100%

237
may not be discriminating between low and high achievers. A Disc D of 0.4 is considered adequate for classroom tests. By undertaking item analyses, teachers can identify some of the weaknesses of the items and thus improve upon them. The results of the try out test can then be used to refine the test.

An example of item analysis results using Excel and taken from Alias is given in Table 3 [5]. The example in shown in Table 3 is based on data taken from 30 students, and the formula for Disc D is given as Disc D = (U - L)/n, where U is the 50% of upper scores and L is the 50% lower score. The formula for Diff P is given as Diff P = (U+L)/Total N. Both formulae are from Black [3].

Estimating Reliability

In addition to the items analysis, a teacher should gain some estimate of the reliability of his/her test’s scores as part of the evaluation process. The reliability for norm-referenced classroom tests may be estimated using various methods, with the Cronbach Alpha method being the most common method used. The Cronbach Alpha method provides estimates of internal consistency based on all possible split halves, while the split-half method provides an estimate of internal consistency based on two equivalent halves. The Cronbach Alpha coefficient, α, may be estimated using Equation 1,

$$\alpha = \frac{N}{N-1} \left( \frac{\sum S_i^2}{\sum S_i^2} \right)$$

(1)

where, \(N\) is the number of items (or identifiable parts of essay questions), \(S_i^2\) is the variance of individual questions (or parts) and \(S^2\) is the variance of whole test. An alpha coefficient of around 0.7 can be considered adequate for classroom tests. A lesser value may be obtained if heterogeneous items are used.

The author managed to obtain a good reliability estimate of 0.74 based on the table of specifications shown in Table 2 [6].

Estimating Cronbach Alpha Coefficient Using Spreadsheets

The use of spreadsheets in estimating test reliability may greatly reduce the workload associated with repetitive hand calculations for teachers. Table 4 is the suggested spreadsheet format for reliability calculation adapted from Black [3]. The shaded cells constitute scores that the teacher has to key in while the rest are computed using formulae in Excel. Having once set up the spreadsheet, it is readily available for future usage.

Similarly, inter-scorer reliability can also be easily estimated using the template in Table 4. In this case, Items are replaced by Examiners, and the scores for each item in Table 3 (shaded cells) are replaced by total scores for each student.

Administering the Test

Tests that are well designed but not administered in an appropriate manner may still fail to produce reliable results, ie the results produced are not representative of students’ actual capabilities. An example of instances where a test fails to be administered in an appropriate manner is when poor invigilation allows cheating among students. In this case, students’ performance may be higher than actual scores and the results are not valid because there is inconsistency between actual and obtained scores. Therefore, even at this stage, care should be taken in order to avoid raising doubts over validity and reliability.

Table 4: The Alpha calculation for a norm-referenced test.

<table>
<thead>
<tr>
<th>Item</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>(S_i^2)</th>
<th>(S^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>10.927</td>
<td>12.409</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>16.034</td>
<td>13.355</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>16.959</td>
<td>14.595</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>16.959</td>
<td>14.595</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>16.959</td>
<td>14.595</td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
<td>125</td>
</tr>
<tr>
<td>(\bar{x})</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25.000</td>
<td>25.000</td>
</tr>
<tr>
<td>(\bar{S}_i)</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25.000</td>
<td>25.000</td>
</tr>
<tr>
<td>(\bar{S})</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25.000</td>
<td>25.000</td>
</tr>
</tbody>
</table>

CONCLUSION

Classroom tests are routinely designed and administered by teachers. To be of real value they must be valid and reliable. Test validity and reliability may be achieved from the steps taken throughout the design and administration stages. Two of the most effective methods that could be employed to enhance reliability and validity are constructing a table of specifications and carrying out a pilot study on the newly designed test. For increased efficiency, teachers may decide to work in teams to design and develop classroom tests. Lastly, although following the recommended measures previously discussed does not provide a guarantee for a perfect and valid test, it can certainly help teachers from getting it totally wrong.

REFERENCES

The effect of the blended Problem-Based Learning method on the acquisition of content-specific knowledge in mechanical engineering

Maizam Alias & Hasanul H.M. Saleh
Tun Hussein Onn University of Malaysia
Johor Darul Takzim, Malaysia

ABSTRACT: Problem-Based Learning (PBL) on its own has not been very successful in ensuring the learning of content-specific knowledge in engineering. The purpose of this study was to determine the effect of blending PBL and conventional teaching methods on students' achievements in mechanical engineering. The quasi-experimental design method was used with two classes of first year mechanical engineering students who registered for the Fluid Mechanics I course at Tun Hussein Onn University of Malaysia in Johor Darul Takzim, Malaysia. The experimental group (n=28) was prescribed the blended PBL method while the control group (n=52) used the conventional method for completing group tasks. A pre-test and two post-tests on selected topics in Fluid Mechanics I were administered to both groups before the study in weeks 4 and 11, respectively. The results showed that the mean scores on the achievement tests of the blended PBL group was statistically significantly higher than the conventional group in both instances (weeks 4 and 11) with the effect sizes ranging from 0.56 to 1.1. It was concluded that blended PBL is better at developing content-specific knowledge compared to the conventional method.

INTRODUCTION

It is well accepted that engineers of the future need to develop additional skills to be able to cope with the continuously changing, technologically advanced, socially and politically complex working environments. Thus, besides having a good grasp of basic engineering principles, engineers also need to have additional attributes to be:

- Good communicators;
- Innovative;
- Creative;
- Able to manage people as well as systems;
- Life-long learners;
- Adaptable [1].

Therefore, there is an urgent need for teaching and learning innovations in engineering education to achieve the goals of future engineering education that support not only the need to acquire content-specific knowledge, but also the development of the broad knowledge and skills that are demanded of future engineers. As such the emphasis on Subject-Based Learning (SBL), with the teacher being the centre of learning process, is being replaced or complemented with a more appropriate learning approach.

PROBLEM-BASED LEARNING

Problem-Based Learning (PBL), where the problem is an important element in the learning process, has been attracting much interest from engineering educators recently due to its potential in promoting the development of wholesome engineers of the future. The PBL method is based on constructivism, which proposes that learning is a process wherein the learner actively constructs knowledge. Learning results from a learner’s actions; instruction plays a role only to the extent that it enables and fosters constructive activities [2]. According to constructivist theory, learning occurs as a result of a process where students actively construct their own knowledge from their experiences. Learning occurs when students are able to make connections of new information with knowledge and experiences that they have already assimilated. Learning becomes an act of discovery as students examine the problem, research its background, analyse possible solutions, develop a proposal, and produce a final result.

Knowing about knowing or metacognition, which refers to the ability of knowing how one knows or learns, affects learning. Good students can detect when they understand – or do not understand – new information, and know when to use different strategies to make sense of new knowledge and experiences. They are able to judge the difficulty of problems and assess their own progress in resolving them.

Social and cultural factors affect learning and, therefore, the emphasis is on learning within a context to ensure greater understanding by making connections between learning materials and real-life applications. In PBL, students are dealing with problems that are designed to be as close to real-life situations as possible. The social interaction imposed through PBL (working in groups) is not only instrumental in ensuring that learning occurs, but probably helps in the longer retention of knowledge. In one such study on medical students, the author found that PBL improved learning with effect sizes as high as 0.5; not only that, he also found that medical students who used PBL still retained their knowledge even as long as two years after it was learned [3]. In general, the characteristics of the PBL method appear to provide the right learning experiences that have the potential to develop excellent analytical skills and the ability to deal with complex engineering problems.

In PBL, students learn about contents through challenges in the form of problems relevant to their future practice [4]. This is
the opposite of the conventional teaching method that teaches content to help students solve problems. PBL uses the problem to challenge, motivate, focus and initiate learning [5]. In the PBL method, real-life problems that are not defined in engineering terms are posed to students. Therefore, problem analysis, definition and formulation in engineering terms are the critical prerequisites of the problem-solving process [6].

According to Woods, the fundamental difference between PBL and content-based learning or Subject-Based Learning (SBL) is in the starting point of the learning cycle [7]. In PBL, the learning process starts when students are given a problem following which they identify what they need to know in order to solve the problem. In SBL, students are given what they need to know to solve a problem following which they are given the problem.

PROBLEM STATEMENT

Although much support has been found for the efficacy of PBL in developing generic skills, its effect on the learning of content-specific knowledge is not as positive [8]. PBL has caused some students to feel... unsure how much self-directed study to do and what information is relevant and useful [9]. Learning the course content was found to be the main challenge and students perceived that they learnt less content in PBL compared to learning in a traditional course [10].

The student-centred nature of the PBL method has also caused frustrations among some students, who have strongly believed that the only way to learn is by... attending lectures and by listening to faculty [10]. This study also found that students who were new to PBL required longer periods of orientation to solve problems compared to those students using the conventional method and who received more inputs from their lecturers. Most students tended to have the opinion that the lecturer is the source of knowledge, thus the change from teacher-dependent to self-directed learning requires a change of paradigm which is not something that is easily acquired.

On top of the demand for increased students’ efforts to acquire new knowledge and skills, students also need to develop self-assessment skills to monitor their own progress. The inadequacy of PBL in promoting content knowledge developments has been further supported by Kirschner, Sweller and Clark, who provided the theoretical foundations on why minimal guidance as provided in PBL is not effective in learning content-specific knowledge or problem solving [11].

Therefore, according to Savin-Baden, a blended PBL method, where some components of the conventional method are retained, would be more appropriate, especially with learners who are new to PBL where the goal of teaching and learning is to develop competent applications of knowledge in problem solving [12][13]. One such model has been proposed by Fink where conventional components (lectures, tutorials and experiments (laboratory assignments)), are blended with the following PBL components, namely:

- Problem analysis;
- Literature reviews;
- Field studies;
- Group work;
- Problem-solving;
- Reporting [6].

The purpose of this study is to determine the effectiveness of Fink’s PBL model on the achievements among mechanical engineering students. The null hypothesis given below was formulated to guide the study:

- Ho: There was no statistically significant difference in the means of test scores between the PBL and the control group.

The independent variable is the teaching method (PBL or conventional teaching method) and the dependent variable is the learning gain operationalised as the score on the achievement test in the Fluid Mechanics I course offered at Tun Hussein Onn University of Malaysia in Johor Darul Takzim, Malaysia. The conceptual framework for the study is shown in Figure 1.

![Figure 1: The conceptual framework.](image)

Learning achievements
- Knowledge;
- Understanding;
- Application [15].

METHODODOLOGY

Research Design

The quasi-experimental design method was used with two classes of first year students from the Mechanical Engineering programme at Tun Hussein Onn University of Malaysia who registered for the Fluid Mechanics I course as the samples. One group was asked to use a prescribed PBL method (n=28) to complete their group assignments while the other group used the conventional group method (n=52). The two classes were taught by experienced but different lecturers.

Instruments

Three achievement tests, namely pre-test, Test 1 and Test 2, were used to gather data on students’ content-specific knowledge of Fluid Mechanics I. The items in these tests were designed at the knowledge, understanding and application levels based on Bloom’s taxonomy of learning in the cognitive domain [14].

The pre-test consisted of 13 items on the topics of liquids in equilibrium, Test 1 had six items on the same topic and Test 2 consisted of three items on basic equations in fluid mechanics, which is the last topic in the Fluid Mechanics I syllabus. These achievement tests were developed by two subject matter experts. A table of specifications was constructed as a planning tool to ensure content validity and standardised marking schemes were developed in order to ensure intra-rater and inter-rater reliability.
Procedure

The study was carried out over 11 weeks in duration. Both groups were pre-tested before the intervention to determine their baseline knowledge and skills. Both groups were asked to complete three group tasks: Task 1, Task 2 and Task 3. The pre-test was administered before Task 1. Test 1 was conducted after Task 1 (at four weeks into the semester) and Test 2 after Task 3 (at 11 weeks into the semester).

For the PBL group, the tasks were completed using the following steps:

- Respond to the problems (triggers) – FILA table;
- Generate hypothesis;
- Research – KND chart;
- Problem solving;
- Presentation;
- Documentation;
- Assessment.

The time given for students to complete their tasks was one week for each task. An orientation session was provided to help students familiarise themselves with the PBL method. Students worked in groups of four or five with members of their choice.

The same tasks were also given to the conventional group that worked in groups of four or five members of their choice. The favoured working strategy for the conventional group was to divide a given task among group members and combine their efforts later before submission.

RESULTS

Group Equivalence

The Welch t-test was used to compare the means on the pre-test because the samples had an unequal variance as shown by the Levene’s test results in Table 3. The pre-test scores of the PBL group (\(\bar{x}=8.77, s=5.92\)) and the conventional group (\(\bar{x}=8.73, s=3.34\)) were very similar and not statistically significant at the 0.05 level (\(p=0.975\)) as shown in Table 1. Therefore, it can be concluded that at the start of the study, the two groups were equivalent with respect to their cognitive skills.

Table 1: Results of the Welch t-test on the pre-test means.

<table>
<thead>
<tr>
<th>Levene’s test</th>
<th>Welch t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F)</td>
<td>(p)</td>
</tr>
<tr>
<td>5.805</td>
<td>0.018</td>
</tr>
</tbody>
</table>

The Effect of PBL on Cognitive Skills

The results of the independent t-test on the difference between the mean scores in Test 1 are shown in Table 2. The obtained \(p\)-value was smaller than 0.05; therefore, the null hypothesis of no difference was rejected. This means there was a statistically significant difference between the mean scores, with the PBL group scoring higher than the control group for Test 1.

An effect size of \(d=1.113\) was obtained using the Cohen method [15]. It was found that 86% of the students in the control group obtained scores below the mean of the PBL group.

Table 2: Results of the t-test based on Test 1.

<table>
<thead>
<tr>
<th>Group</th>
<th>(\bar{x})</th>
<th>(s)</th>
<th>(t)</th>
<th>(df)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBL</td>
<td>18.7 (62.3%)</td>
<td>4.80</td>
<td>4.68</td>
<td>78</td>
<td>0.000</td>
</tr>
<tr>
<td>Control</td>
<td>13.6 (45.2%)</td>
<td>4.60</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the independent t-test on the difference between the mean scores in Test 2 are shown in Table 3. The obtained \(p\)-value was smaller than 0.05; therefore, the null hypothesis of no difference was again rejected. This means that there is a statistically significant difference between the mean scores, with the PBL group scoring higher than the control group for Test 2.

Table 3: Results of the t-test on long-term learning.

<table>
<thead>
<tr>
<th>Group</th>
<th>(\bar{x})</th>
<th>(s)</th>
<th>(t)</th>
<th>(df)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBL</td>
<td>11.2 (56.1%)</td>
<td>5.10</td>
<td>2.29</td>
<td>78</td>
<td>0.025</td>
</tr>
<tr>
<td>Control</td>
<td>8.7 (43.4%)</td>
<td>4.50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The effect size was 0.56 for Test 2. Although smaller in size compared to the effect size in Test 1, according to ref. [16], a size of above 0.50 is considered to be moderately significant. The effect size value of 0.56 shows that the mean of the PBL group is at the 71st percentile of the control group.

DISCUSSION

The consistently higher achievements obtained by the PBL group on cognitive skills is in contrast with Glew and Wood [8][9]. The differing results may be explained in terms of the specific PBL model used in the current study. In the current study, the PBL method was not used to replace the existing method, but rather to complement it. The existing conventional teaching and learning method, namely lectures, laboratory assignments, etc., were still used hand-in-hand with the PBL method.

Such an introduction to the PBL method probably reduces the cultural shock that students have often faced when using the PBL method for the first time. Furthermore, the current PBL model may provide greater opportunities to meet students’ varied learning preferences. This could be the reason why the students in the PBL group obtained better results on the achievement tests as compared to the conventional group. The meeting of learning style needs is still a hypothesis that remains to be tested in the future.

The effect of novelty is a common problem in teaching and learning studies. Nevertheless, such an effect in some aspects is good, as stated by Hawthorn (in Cohen and Manion): Put people in a novel situation and observe them and they will work harder (for a time) [16]. This means any new item in teaching and learning will attract attention and become a catalyst for a better result even though only for a short period. However, if the increased achievements in the PBL group were the result of novelty, then the level would have declined steeply after the initial excitement was over, which is not so in this case. The PBL group was still better at week 11, while the conventional group remained at the previous level. This could mean that the PBL method is indeed more effective than the conventional method.

Another major finding from this study is that the learning obtained is not only statistically significant, but also academically significant as the effect size was consistently valid.
The large effect size ($d>1.0$) is not only an indication of the effectiveness of the PBL method, but also indirectly a reflection of how poorly the outcome of the conventional method is because if the conventional group had done better, the observed effect size ($d=1.113$) would not have been possible. In medical education, for example, where students' performance is generally very good, the effect size obtained rarely reaches above 0.8 [3].

CONCLUSION

In this research, the authors set out to determine if learning effectiveness in engineering can be enhanced by complementing the conventional teaching method with PBL. The data generated supported the conclusion that the learning effectiveness of engineering subject matter content can be enhanced when PBL is used to complement the current teaching method. The findings were not only statistically significant, but also academically significant meaning that the results obtained were worth the effort.

ACKNOWLEDGEMENTS

The authors would like to express their appreciations towards the lecturers of the subject DDA 3033-Fluid Mechanics I at the Tun Hussein Onn University of Malaysia, who made this study possible.

REFERENCES