Integrity of Electronic Testing of Mariners

John Tucker [Professor, M.Eng., P.Eng.]\(^1\), John Cross [Professor, ONII, M.Eng., P.Eng.]\(^2\)

\(^1\) Marine Institute, Memorial University of Newfoundland, john.tucker@mi.mun.ca
\(^2\) Marine Institute, Memorial University of Newfoundland, john.cross@mi.mun.ca

**Abstract.** With the Manila Amendments to STCW that introduce distance learning and web-based learning as valid course delivery systems there have been several initiatives aimed at developing distance courses for mariners. There are ways that content can now be effectively delivered to mariners that allow learning to take place outside the academic setting, most importantly on ship. However, electronic testing brings with it two potential problems:

- First, the test must be structured so that it provides a fair opportunity for assessment.
- Second, the test must maintain the academic integrity of the evaluation and course.

While electronic testing has been around as long as computers have been in common use, the technology and educational pedagogy is only now beginning to capture the sophistication of a typical test. The delivery of the test is fairly easy, but grading it so that it provides a fair and accurate assessment of a student’s capability is much more challenging.

The second problem is that the provider of the test must be certain that the student is displaying their capability. It is naive to think that students are not taking advantage of electronics or other methods to give them an unfair advantage in examination performance. If the examination is taking place at a distance the problem is even greater with the potential for everything from a student “googling” the answer to it not even being the student sitting and writing the test.

For electronic testing to reach its full potential these problems must be addressed. This paper begins with lessons learned from the author’s experience in developing electronic questions over the last 5 years and provides some insight and guidance for anyone interested in creating questions. It then goes on to look at how the Marine Institute is currently handling distance testing and the potential for new biometric processes to make distance invigilation stronger and more secure.

**Keywords:** Electronic Testing, Online Proctoring, Examination of Mariners

1. **Introduction**

The 2012 Manilla Amendments to STCW opened the door for national certifying authorities to permit Maritime Education and Training institutions (MET’s) to deliver certification level training using online or web based. The critical nature of maritime training training and the assessment that must accompany it for certification advancement resulted in this evolution taking a longer time to occur than in most other post secondary education institutions, but the authors feel that this evolution is inevitable.

With the MET transition to online or web delivery lagging most other higher education fields, we are afforded the opportunity to study and learn from online delivery of programs in other areas of higher education and apply those lessons learned to the evolution of our programs and the ways in which we evaluate our learners.

2. **The Categorization of MET**

Trenholme categorizes online deliveries as being of two distinct types; “Writing-Based” (WB) courses where learners are required to author papers and answer essay type questions, or “Math or Fact-Based” (MFB) style courses in which in which the learners are required to recall facts or perform calculations [1].
WB courses tend to be assessed in a more subjective way where presentation as well as creativity may have an impact on the performance of the learner in an evaluation. A test for academic dishonesty in this type of evaluation would typically be an assessment to determine plagiarism. There are a host of tools and devices to assist in ensuring the academic integrity of WB type evaluations ranging from online plagiarism checkers to student portfolio creation and maintenance tools incorporated into the majority of Learning Management Systems. Sadly, the development of a set of similar tools to prevent academic dishonesty in MFB type courses has not evolved at the same pace. MFB style course evaluations are more objective. The correct answer is, simply, the correct answer. As a consequence of this, an assessment of plagiarism in an MFB type evaluation is not valid.

MET would tend to fall into the category of a MFB type of delivery. Mariners are required to learn mission critical facts and perform objective evaluations in a consistent and timely fashion. This is reflected in the model course content presented in STCW, course deliveries provided by MET Institutions, and evaluations performed to permit certification level advancement as provided by either the MET Institutions or the national certifying authorities.

2.1 Cheating in MET Certification Level Evaluations

It is important to note that, at this time, all certification advancement level examinations of mariners conducted in Canada are conducted using a face-to-face method. In this system, examinees present themselves at an authorized examination centre with appropriate documentation, identification and credentials, and sit to write their examination in an invigilated setting.

In high stakes evaluations such as these, academic honesty is a well-guarded feature of the entire certification process. It is conceivable, however, that even with such a strict and time-tested system, the pressures to perform well on these evaluations can lead individuals to use any means at their disposal to assist them in succeeding in their evaluations.

In the 1980’s, the authors observed a method that was employed to better the chances of success in MET evaluations that was a consequence of poorly designed evaluation databases. It was the case at this time that a limited database of static (unchanging) examinations was employed for certification purposes in Canada. Mariners would write a certification level examination offered by the national certifying authority that was drawn from this limited database, and upon completion they would leave the exam and immediately note the exam questions on that version of the examination. These questions were solved, shared and compiled over time as a set of exam questions that you could expect to see in the certification level examination. Some learners would memorize a set of questions and then take the examination multiple times until they received the exam with the questions they had committed to memory…and pass.

The examination database in Canada has experienced a serious restructuring and updating over the last few years and now draws on a more flexible bank of new questions in the hopes of alleviating the above mentioned issue.

In considering the possibility of offering MET certification level advancement examinations online, as part of a course or not, we must re-evaluate the methods we use to administer and secure the examinations and the writing of the examinations. Providing a less strictly controlled environment for the writing of examinations must be avoided, as it would provide a more tempting opportunity for cheating.

Three factors should be considered as influencing mariners when faced with the perceived opportunity to cheat in an online evaluation:

1. Increasing pressure to advance in certification level. This can come from the opportunity for monetary gains, employment opportunities or status shift.
2. Shifting socio-economic views on the acceptability of cheating on an evaluation. From the examinee’s perspective, the mission is to obtain the advancement in certification level. What it takes to obtain this advancement is somewhat irrelevant. From the perspective of the certifying authority, it is imperative that the examinee demonstrate knowledge and expertise in the examination topic area.

3. Anonymity is a factor unique to the administration of online examinations. In the case of online evaluations it is more the perception of anonymity that the examinee may feel when writing an examination with no one else in attendance than the actual fact of anonymity.

Trenholme goes on to add that academic dishonesty observed in an online or web based delivery of a MFB type course is observed as either:

1. Use of an alternate individual, typically an expert, to complete evaluations or assessments.
2. Unauthorized collaboration on evaluations; and
3. Unauthorized coaching on evaluations.

In the context of MET, these evaluations typically consist of requiring the mariner to provide an evaluation or factual data in which the correct answer must be provided in a clear and concise fashion.

Plagiarism is not a factor in MFB assessments as the correct answer is the correct answer. However, the three above-mentioned concerns for academic dishonesty are. For online course deliveries as well as online evaluations to take place, a paradigm must be established whereby there is complete confidence in the academic integrity of all evaluations in the online MET delivery.

3. Structuring Electronic Tests

The purpose of any test is to provide a fair assessment of a student’s knowledge and ability. At the most basic level, this is just a recall of facts, however at the higher cognitive levels it requires both a recall of facts as well as assessment of pertinent information and application of a solution algorithm.

In fact testing by electronic machines has been around for over 80 years. The IBM 805 was developed in 1934 and would correct a score sheet placed in it [2]. It would apply a small voltage across the correct multiple choice answer and if a pencil mark was present in that place a small current would flow. The machine would measure the total current flowing and thus the total score for that sheet.

However the digital age has allowed much greater flexibility and sophistication to be employed. The authors were trying various means of incorporating electronic assignments into their classes as a means of reducing plagiarism prior to 2010 [3]. However with the adoption of MapleTA in 2011 for assignments the potential to use it for tests and examinations was first realized. It should be noted that while we use MapleTA, there are a number of similar software packages available that have similar functionality.

The first use of MapleTA for testing was its use to administer a multiple choice type examination in a material science class. It was an obvious candidate since previous examinations in this course were developed from an existing bank of multiple choice questions. However MapleTA allowed a couple of beneficial options. The first was that we could create multiple copies of similar questions (for example a diagram showing a particular plastic manufacturing process and asking what the process was). We could also group related questions together. For example we might have 10 questions related to corrosion. When it came time to make up the exam we would choose the topics we wanted and then the number of questions from each topic (for example we might want 4 questions on corrosion) and MapleTA would create each student an individual test based on how we specified (for example in the case outlined the computer would choose a random 4 out of the 10 corrosion questions). This way tests could be created to target certain areas but also prevent any potential for copying since each student had an individual exam.
Lately, our experience with MapleTA has allowed us to create more sophisticated questions. For example we now have a course in Applied Mechanics running which tests exclusively using MapleTA. A typical question is shown below in Figure 1 where a student is asked to find the centroid and moment of inertia of a complex shape. The dimensions for the shape are randomized within a certain range. For example the thickness of shape 2 can range between 0.5 inches and 1.0 inches in increments of 0.1 inches. In all there are 1,764 possible variations for this question which means in a typical class of 25 students, there is only a 15% chance that two students will get identical questions. However all questions follow the exact same procedure! We call these questions dynamic questions since they change with every student as opposed to a static question where every student answers the same question.

![Figure 1 Sample Dynamic Question from MapleTA](image)

The math engine behind MapleTA is very powerful and has allowed us to create questions that do not normally have a calculation based answer. For example to create questions about phase diagrams, the curves were approximated by 4th order polynomials which allow solutions sufficiently accurate to correct answers from students reading tables. Similar things were done in fluid mechanics as well as thermodynamics.

The creation of a static question is a relatively straightforward process where the instructor makes up a question and then answers it. Usually there is not a problem, but if the instructor doesn’t like the way the question turns out (too easy or too hard for example) they can change it before giving it to the students.

The development of a static question follows the same process to start with – a static question is created and coded into MapleTA and the solution checked. However then certain variables must be randomized and this requires a number of additional tests. The following is a checklist questions that we apply to the conversion of static questions to dynamic questions.

- Will any selection of the randomized values cause division by zero?
• Will any combination of the randomized values create a non-realistic answer?
• Will any combination of the randomized values cause other physical factors to come into play?
• For multiple choice questions, will any combination of the randomized values result in two or more of the calculated distractors and/or the correct answer to have the same value?

While the creation of dynamic questions is a much longer process than the creation of static questions we see significant benefits in them. To begin with, students can brainstorm as a group yet every individual is forced to do their own work. This is very important since teamwork is a critical skill for them to learn, but at the same time each individual must develop enough confidence to solve problems on their own. In addition the same question can be used multiple times, for example the instructor can solve it in class and then assign the same question for homework but each student will be forced to work through the problem step by step for their individual question. Finally there are tools available that will allow instructors to identify weak or poorly answered questions – this will be expanded upon in Section 4.

4. Assessing Tests and Good Educational Practices

The creation and correction of a test is too often thought of as the end of the instructor’s involvement. However due to the importance of training mariners, it is important for MET instructors to continuously review their tests to ensure the evaluation is performing the way it should. To quote Matlock-Hetzel “… some best practices in item and test analysis are too infrequently used in actual practice” [4].

She recommends the following item statistics be used: the p-Value, the d-Value and the p-Biserial. These will be discussed below, but they all require some post-test analysis which is at best tedious and potentially complicated.

When looking at item statistic analysis, electronic testing has two strong advantages. First the records are available electronically and are able to be stored easily, securely and indefinitely. The second is that the digital format makes it very easy to perform the item analysis.

In regards to keeping of student records, electronic testing provides the opportunity to keep student work around as long as you wish. The information is collected as the student performs the exam (note that in our system, the answers are stored on the server so even if the student’s computer crashes the information is safe). The storage space is minimal and in our experience the only reason to not keep student records is due to privacy issues. It is hard to imagine a circumstance where student records would be needed after 5 years so we tend to purge student records after five years.

In regards to tracking the performance of test items electronic testing provides some extremely useful statistics which can yield valuable information. By keeping a record of every response to a question as well as the overall grade, a user can identify questions that are possibly misleading.

The question generation software we use generates 5 useful statistics out of which we use 4 to identify potential problem questions. The 4 are: Success rate, p-Value, d-Value and p-Biserial.

The Success rate is the one statistic that Matlock-Hetzel does not use. The reason is that for a test where all questions are either right or wrong then it is the same value as the p-Value. However we use testing for questions where there are multiple parts and thus find both the Success rate and the p-Value useful. The success rate is simply the average score (normalized to a value between 0 and 1) on any particular test item (or question). You can think of it as the average score from all students who were graded on that question. It is useful as a quick check of the validity of the test item. If no one is getting the question wrong (i.e. the Success rate is 1) then perhaps the question is too easy. If no one is getting the question right (i.e. the Success rate is 0) then perhaps the question is too hard or has some other issue with it such as no correct answer.
The p-Value is a measure of the ratio of fully correct responses to the test item to all the responses. Taken with the Success rate it can provide information about the internal structure of the test item. If the p-Value is the same as the Success rate then there are no partial marks being awarded because all students get the question completely correct or completely wrong.

The d-Value is the discrimination of the test item. It is calculated by finding the difference of the p-Value of the top half of students (based on overall exam scores) and the p-Value of the bottom half of the students. It can be a value between -1 and +1. This is useful as rough measurement of how well good students do on a question as opposed to no-so-good students. We would expect students who score highly on the test to be knowledgeable about the subject material and thus would score well on any given item. The converse is true for students who are in the lower half of exam scores. A high positive value is desirable for a typical test item.

The last statistic is the p-Biserial or point biserial correlation coefficient. It is a more sophisticated measure of how well good students do as opposed to the no-so-good students, so it is a better indicator of how good students (and how no-so-good students) do on a particular question. It ranges in value potentially between -1 and +1. A higher positive value generally indicates a test item that is accurate at measuring student performance.

It is important to note that any of the above statistics for the results from an individual class are, by themselves, only an indication of whether a problem exists. An instructor needs to examine potential problematic questions to see if there really is an issue with the question. Figure 2 shown the statistics from a typical class of 21 students.

**Item Statistics :**

<table>
<thead>
<tr>
<th>Question</th>
<th>Description</th>
<th>Success rate</th>
<th>p-Value</th>
<th>d-Value</th>
<th>p-Biserial</th>
<th>r-Biserial</th>
<th>Count</th>
<th>Correct</th>
<th>Partial</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Problem 1.1 (4)</td>
<td>0.833</td>
<td>0.81</td>
<td>0.096</td>
<td>0.54</td>
<td>0.78</td>
<td>21</td>
<td>17</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>(2)</td>
<td>Problem 1.3 (4)</td>
<td>0.714</td>
<td>0.524</td>
<td>0.24</td>
<td>0.434</td>
<td>0.544</td>
<td>21</td>
<td>11</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>(3)</td>
<td>Problem 1.5 (4)</td>
<td>0.81</td>
<td>0.81</td>
<td>0.298</td>
<td>0.707</td>
<td>1.021</td>
<td>21</td>
<td>17</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>(4)</td>
<td>Problem 1.7 (4)</td>
<td>0.861</td>
<td>0.857</td>
<td>0.173</td>
<td>0.717</td>
<td>1.113</td>
<td>21</td>
<td>18</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(5)</td>
<td>Problem 1.10 (6)</td>
<td>0.967</td>
<td>0.429</td>
<td>0.087</td>
<td>0.429</td>
<td>0.541</td>
<td>21</td>
<td>9</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>(6)</td>
<td>Problem 1.19 (3)</td>
<td>0.857</td>
<td>0.857</td>
<td>0.376</td>
<td>0.804</td>
<td>1.247</td>
<td>21</td>
<td>16</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>(7)</td>
<td>Problem 1.25 (8)</td>
<td>0.592</td>
<td>0.361</td>
<td>0.618</td>
<td>0.491</td>
<td>0.825</td>
<td>21</td>
<td>6</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>(8)</td>
<td>Problem 1.39 (6)</td>
<td>0.857</td>
<td>0.857</td>
<td>0.376</td>
<td>0.73</td>
<td>1.132</td>
<td>21</td>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

**Figure 2 Question Statistics obtained from MapleT.A.**

As can be seen, no question had either a 100% success rate or a 100% failure rate which is an indicator that it is testing student knowledge. The p-Value is showing that there are a suitable range of partial marks being awarded for each question. The d-Values are all positive, but question 1 and 5 show very low values indicating that students in the top half of the class do about the same as students in the bottom half of the class. While not a problem by itself, it does indicate potential problems so we look at the p-Biserial value. The p-Biserial values are 0.54 and 0.43 which both indicate a question which is functioning as it should.
5. Online Proctoring

The issue of proctoring exams is one of the most critical and culturally complex issues in the delivery of online certification level courses. Historically, students taking distance courses were required to travel to an authorized examination centre to write tests and examinations in a proctored or supervised setting. The challenge being that many learners participate in these courses because they lack the ability to conveniently travel to such a centre to participate in course evaluations. This contradicts with the rationale for online delivery and, until recently, resulted in post secondary institution struggling with the question of trusting students to write examinations on their own computers in an unproctored setting or requiring the students to travel to an examination centre.

Advancements in technology have provided some interesting solutions to this challenge. There are many companies who, today, provide online proctoring services. All of these services have standard, common features [5]. The process of taking an online test would typically begin with some form of authentication. This could be in the form having the student input a username and password but may also involve verification of the person's identity using a photo ID and a webcam or facial recognition software. The test itself typically involves an actual proctor, which could be human or some form of technology which would observe the test taker during the evaluation event. This observation could be of the form where the testee's computer (screen) is being observed or recorded as well as the individual being observed using a webcam and a computer microphone. This permits the proctor to observe the test paper, to watch the individual to ensure that they stay at their desk during the evaluation, and it also permits the proctor to listen to the person being tested to ensure they are not in communication with anyone who is unauthorized during the test event.

MUN recently conducted a study and has implemented a policy to manage online proctoring of examinations. Online exam proctoring is now the default method of invigilation for examining students that reside outside the province of Newfoundland and Labrador. Professional proctors use webcams and microphones to remotely supervise students writing exams during their examination period.

The students may write their examinations at home or any quiet location with an active internet connection. The exam is written using a secure lock down browser and the scheduling of examinations is flexible but has to be set up beforehand.

Similar to writing exams in a classroom, students are required to show both their MUN Campus ID Card as well as a government-issued photo ID to the proctor prior to taking the exam. To ensure the privacy of student information, all data is stored in a secure data center in Canada and may be temporarily stored in secure locations outside Canada to facilitate transfer.

There are specific technical requirements that must be met, but they are not onerous or requirements that a student participating in an online course would not already be expected to have.

We can expect uncertainty about electronic examinations outside of a campus to remain a concern, but new technologies are constantly alleviating problems.

More sophisticated and secure means of authenticating and proctoring critical evaluations is that of using biometric techniques to monitor students. The three main biometric systems are: finger prints, facial recognition and retina scans. All provide some degree of certainty that the correct person is sitting at the computer, however there are still potential problems with having unmonitored students taking examinations. In addition, Canada has some of the strongest privacy laws in the country and some of these methods are considered intrusive.
6. Conclusion

With online distance courses becoming almost the norm these days, the use of electronic testing for distance courses must be seriously considered. It provides the convenience of allowing mariners to progress without the challenges of geography that mariners often face. However it comes with a number of potential problems that must be weighed against the benefits.

The problems are that confidence in the test administration and delivery process must be gained at all levels on the certifying side. Everyone from those who administer the evaluations to the certifying authority and IMO must have full confidence in the delivery of online evaluations before this type of testing will be permissible. An additional challenge to be considered is from the perspective of the mariner being evaluated in that they must understand the safeguards that have been put in place to ensure the academic integrity of the test and not feel that, although they are alone in the room while taking the evaluation, they are not being proctored.

The benefits are that it allows great flexibility in administering tests and has the potential to make tests more secure by allowing individual tests to be created, administered and corrected automatically. In addition it allows the instructor to continuously monitor the test’s performance to ensure that it is evaluating the way it is supposed to.

While the proctoring issue is significant, we feel that this will soon be solved in an acceptable manner and then electronic testing will allow the convenience and flexibility that mariners will find useful as they progress through their careers.

References