A LOW-PAIN ASSESSMENT MODEL FOR LABORATORIES AND TUTORIALS

David J. Low, Heiko Timmers

Presenting Author: David J. Low (d.low@adfa.edu.au)
School of Physical, Environmental and Mathematical Sciences, University of New South Wales at the Australian Defence Force Academy (UNSW@ADFA), Canberra ACT 2600, Australia

KEYWORDS: continuous assessment, laboratory, tutorial, attitudes, performance, workload.

ABSTRACT

Continuous assessment motivates students to become engaged with coursework, but presents challenges for staff due to the workload involved in marking and delivering feedback. Care must be taken to avoid students concentrating on simply accumulating marks, to the detriment of learning. We present a Checkpoint assessment model which strives to find a balance between these factors, and show how the Checkpoint model can be applied in tutorial and laboratory environments. Based on an evaluation of changes to student attitudes and performance within a large first-year physics course, we conclude that the Checkpoint model is effective at improving both student satisfaction and results.

Proceedings of the Australian Conference on Science and Mathematics Education, University of Melbourne, Sept 28th to Sept 30th, 2011, pages 199-204, ISBN Number 978-0-9871834-0-8.

INTRODUCTION

MOTIVATION

Continuous assessment is accepted as an important part of the modern tertiary experience. However, the demise of the comprehensive multi-hour exam as the sole or primary assessment task in a course has been accompanied by a concomitant increase in the semester-long workload for both students and academic staff. Much of the increase in staff workload centres on the review of, and allocation of marks to, student work; and, one way or another, workloads have a cost impact on institutions and individuals. Whether that impact is seen through increased sessional staff costs for marking, or via academic staff spending relatively more time on the 'grind work' of assessment rather than course development or research, it appears to be in everyone's interest to develop assessment techniques which minimise processing time, yet provide staff and students with suitable metrics and feedback regarding progress and understanding.

Our aims in this paper are to: (1) present a model for continuous assessment tasks which is easy to implement and mark, yet has sufficient granularity (when applied across a semester) to reflect variations in student performance; (2) give examples of how this model has been applied to the laboratory and tutorial components of a large first-year physics course; and (3) evaluate the model in terms of student attitudes (in the laboratory component) and performance (in the theory component).

BACKGROUND

Black and Wiliam (1998a, 1998b) review the modern implementation of in-class assessment, and highlight some key aspects of formative assessment in this regard. Some of their key findings include the negative impact of emphasising the awarding of marks over learning, and the dangers of students seeing marks as a competitive scoring system. They highlight the need to encourage questioning, search and discovery, via teacher-student and student-student interaction. More recently, these points have been reiterated by Gibbs and Simpson (2004), who note that good learning outcomes are achieved by quality student engagement, not by teachers doing lots of marking! Large class lectures may be time-efficient from a teaching perspective, but economies of scale are difficult to achieve in assessment, to the point that assessment costs rapidly overtake teaching costs in large classes. Gioka (2006) identifies that simply awarding grades, and calling that 'feedback', does not really solve this problem, but also stresses the benefits of oral feedback over written comments. The latter are time-consuming to prepare, yet without significant personalised detail they convey little of worth to students. As an example of addressing these concerns, Foster (2010) describes how changing physics tutorials to a more interactive model, without a 'testing' component, increased participation and reduced course non-completion rates.

For the laboratory environment, Beun (1971) describes a model which avoids marking altogether, by simply requiring students to spend time-on-task. It may be interesting to reflect whether implementing such a model today would result in 'prophecies…about disorganised administration, loitering students, silly experiments, lack of equipment, and unwilling, overloaded staff members' (p.1354) which would be as false today as they were four decades ago! At the other extreme, Ganiel and Hofstein (1982) give a detailed list of objective criteria against which a student could be assessed for each experiment, producing precise, repeatable results, albeit in a time-consuming manner. Patterson and Prescott (1980), following on from Prescott and Anger (1970), describe a middle-ground: a self-paced laboratory emphasising experimental design over following instructions, where students must complete experiments to a satisfactory standard, but better students complete more experiments in a given time and thus achieve a higher mark. In this model, experimental reports are assessed at the conclusion of the activity: student interaction with academic/demonstrating staff is limited to the final review stage. Rice, Thomas and O'Toole (2009), however, in their review of the place of laboratories in tertiary science, note the need for formative assessment in the laboratory environment: students having contact with, and feedback from, a demonstrator, during rather than only after an experiment.

THE CHECKPOINT MODEL

We outline here an assessment model that avoids a detailed review and assessment of every aspect of a student's work. Instead, the Checkpoint model breaks assessment tasks into relatively large chunks (Checkpoints). The assessor then evaluates each Checkpoint via a broad 'complete/incomplete' or 'competent/not-yet-competent' measure. The simplest realisation of this approach would be a binary 1/0 system, although in some cases there may be a justifiable case to implement a middle-class (i.e. adding '½' to the binary system). Care must be taken, however, to avoid implementing discrimination too early: after all, the aim here is not to return to a series of marks-out-of-ten! The Checkpoints should associate with sufficiently large sections of the assessed task that there are not too many in any particular piece of assessment, but should not be so large that they require a significant degree of internal granularity. A good guide seems to be that if one feels a need to go beyond the binary system, one should first question if the Checkpoints are appropriately sized (i.e. are there too few Checkpoints for the task at hand?). Finer detail is imposed at the end of the process by awarding a quality indicator that takes into account the whole of the student's work, and acts as an overall weighting factor. The simplest case here is a terminal letter-grade, with broadly defined interpretations and a numerical equivalent for weighting purposes.

The numerical mark awarded to the student for the assessment task is then the sum of the Checkpoints, weighted by the quality indicator. On any given item of assessment, there may not be much discrimination between students. This coarse grading may reduce the perception of competition between students, as only significant differences in performance result in different marks. However, over the entirety of a semester's continuous assessment, granularity will emerge.

From an assessor's point of view, this is a form of top-down assessment which is relatively fast to implement. Intellectual overhead on the part of the assessor is reduced by keeping the number of assessment decision-points low. On the other hand, by having more than a single assessment point, students receive an indication of what an assessor thought were the strengths and weaknesses of their work. The many face-to-face interactions which occur along with the Checkpoint assessment provide the opportunity for relevant personal feedback. Exemplars, giving an indication of the expected standards, allow students to self-review their work against a reference.

CHECKPOINTS IN THE LABORATORY

IMPLEMENTATION

First-year physics laboratories at UNSW@ADFA were assessed by a single mark awarded at the end of each experiment until 2000, at which point Checkpoint marking was introduced. It is worth noting that the way in which students record their achievements, and the basis of their written assessment, has been the same both before and after the implementation of Checkpoint marking: students record their progress through the laboratory in continuous 'logbook style', via a written notebook: there is no post-experiment write-up or report. Thus, the expectations on students in the laboratory have not changed; only the way in which their work is assessed.

Checkpoints were allocated to every experiment script, spaced with intervals of about one hour of expected workload. Typically, the first Checkpoint would occur after students have thought about

experimental design; one or two would occur during the 'measurement and analysis' phases; and a final Checkpoint would take place at the end of the experiment once conclusions had been drawn. As students reach each Checkpoint, they are required to consult a demonstrator before continuing. Not only does this ensure that students do not get too far off-track in an experiment, it also gives demonstrators the opportunity to discuss what they think is good (or not so good) in the students' work to that point.

Demonstrators are told that the assessment of each Checkpoint should take no more than a minute or two, for a total assessment time of about five minutes per experiment, all within the allocated laboratory time: there is no out-of-class marking. It was expected that demonstrators would spend significant additional time with students *during* the experiment, particularly before work was presented for a Checkpoint assessment, giving advice where appropriate. Thus, the workload and expectations of demonstrators is shifted from what might be termed low-level, post-facto, summative assessment, towards higher-level, continuous, formative assessment. The aim is for demonstrators to act more as facilitators and advisors, than as markers.

The initial Checkpoint system, which ran from 2000-2003, simply replaced a final mark-out-of-ten for each specified experiment with three marked Checkpoints, one after the introduction to the experiment (marked out of two), one after the first few measurements had been made (marked out of three), and one at the conclusion of the experiment (marked out of five). In 2004, the first-year laboratory moved to a 'cafeteria' system (Patterson & Prescott, 1980), which involved students selecting an experiment from those available, carrying it through to satisfactory completion, then moving on to another experiment (etc.); the only time limitation on students being the total number of afternoons they are permitted to spend in the laboratory (i.e. experiments can run across multiple afternoons; and/or students can complete one experiment then start another in the same afternoon). Under this system, demonstrators simply signed off on each Checkpoint when it was completed to their satisfaction, ensuring that unsatisfactory work is not allowed to progress. At the end of semester, Checkpoints were totalled, and a mark issued based on the student's performance relative to the expectation of 'one Checkpoint per hour in the lab': completing a total of 12 Checkpoints in an allocated 15 hours would result in a mark of 12/15 = 80%.

Since 2006, the cafeteria-Checkpoint model has been further modified by the inclusion of a final letter-grade following the final Checkpoint, in order to reflect the quality of the logbook record of that experiment: 'A' indicating impressive work (weighting 0.9–1.0); 'B' – the expectation grade – signifying work at a good standard (weighting 0.7–0.8); while 'C' grades are awarded to work which is adequate yet has significant room for improvement (weighting 0.5–0.6). At the end of semester, weighted Checkpoints are tallied, and compared to the expected standards and extremes. For example, if students are allowed 15 hours in the laboratory, then 15 A-grade Checkpoints (reflecting working at the expected rate of one Checkpoint per hour, all of 'impressive' quality) may correspond to a laboratory mark of 100%. If the announced weighting factors were [A=1.0, B=0.8, C=0.6], then a student who accumulated [3A, 6B, 3C] over their 15 hours in the laboratory program would be awarded a laboratory mark of [(3×1.0)+(6×0.8)+ (3×0.6)]/15 = 9.6/15 = 64%.

EVALUATION: IMPROVING STUDENT ATTITUDES TOWARDS LABORATORY CLASSES

Each semester from 2000-2003, students were asked to rate both the 'Marked Checkpoint' and 'Marked at End' laboratory assessment models. The questions were asked independently, via two separate response items along the lines of, 'What is your opinion of the [...] assessment model?'. A total of 425 responses to each question were received over this time, and were evaluated on a five-point scale (where 1 = 'Great', 3 = 'OK' and 5 = 'Poor'). There was little variation in the statistics over each of the four year (eight semesters). The collated results are displayed in Figure 1, where it can be seen that the Checkpoint model has a sharply skewed positive-opinion distribution, while the end-mark model has a broader distribution skewed more to the negative-opinion end. Written comments from students backed up this numerical picture: the most common responses noted that students appreciated regular contact with demonstrators throughout each experiment, and that this was a feature of the Checkpoint system that was lacking in the 'Marked at End' model. Contact was deemed particularly important in terms of catching errors and identifying potential problems early. Very few students felt that the interactions with demonstrators were unwelcome interruptions!

Overall satisfaction with the first-year physics laboratory also improved with the introduction of the Checkpoint assessment model. Student feedback on the UNSW@ADFA first-year physics laboratory

through the early 1990's indicated general dissatisfaction with the program on a number of grounds, but 'rushed for time' was a common complaint. Staff responded in 1997 by dramatically cutting the work required for each experiment, mainly by reducing the number of repeated measurements and eliminating extension work beyond the main theme. The earliest numerical data we have been able to source covers the period 1996-1999, and the impact of the changes in 1997 can be seen in Figure 2, where responses to the question, 'Are you enjoying the laboratory program?' on a five-point scale (where 1='yes' and 5='no') improved from about 4.0 to about 3.1. With the introduction of marked Checkpoints in 2000, and the archiving of individual student feedback data, a further improvement to net-positive opinions of the laboratory program can be seen: responses to the 'Enjoy?' question averaged 2.3 ± 0.3 over 2000-2003, with standard deviations each year of about 0.9, indicating that the majority of the first-year science cohort were now taking away a positive experience from the laboratory program. While it would be unreasonable to explain this improvement as being entirely due to the change to the Checkpoint assessment model, it does indicate that any effects of introducing that model were likely to be positive.

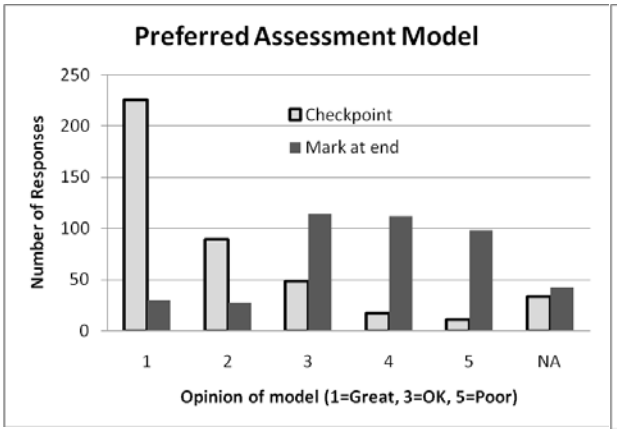

Figure 1: collated student responses over four years/eight semesters (2000-2003; n = 425) to the questions, 'What is your opinion of the [Checkpoint / Mark at End] assessment model?'

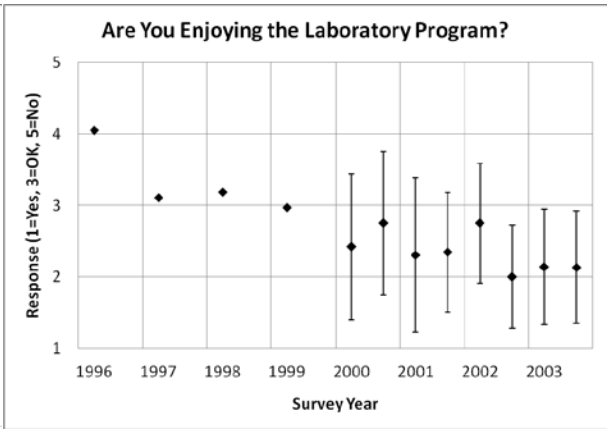

Figure 2: averaged science student responses to the 'Enjoy?' survey question (n = 30 ± 10 each year). Results are shown by semester for 2000-2003. No distribution data is available for 1996-1999.

With regards to staff experiences with the Checkpoint model, most colleagues found the experience to be positive. In particular, the demonstrator workload was spread more evenly throughout the three-hour afternoon, rather than being concentrated at the end, and outside the laboratory, which reduced staff time-on-task. The increased number of interactions with students which is forced by the Checkpoint model was also seen as a positive experience for both staff and students.

CHECKPOINT TUTORIALS

IMPLEMENTATION

A Checkpoint model was trialled in first-year physics tutorials at UNSW@ADFA during the first semester of 2011. The aim was to encourage student preparation and participation in the tutorial program, without resorting to additional summative-style assessment such as quizzes or tests. Students were supplied with a set of questions related to the current lecture material, about a week before each tutorial, and were expected to attempt all of them prior to the tutorial. However, one question was specified as requiring written submission at the start of the tutorial. This 'preparation' question counted as the first Checkpoint (CP1) of the tutorial, and was given either a '1' (for a serious attempt at the problem), '½' (for an incomplete attempt), or '0' (in case of no serious attempt at the problem). The second Checkpoint (CP2) was awarded for in-class work, as documented in a separately-submitted workbook which was regarded as an important learning tool in its own right. CP2= '1' indicated a consistent effort, with well-documented notes or solutions; CP2='½' was awarded for efforts which fall short of expectations in either quantity *or* quality of documented effort; while CP2='0' signified either a lack of effort, or failings in both quantity *and* quality of the submitted work. Finally, the tutor awarded a grade to reflect overall standard and to give extra granularity: 'A' (weighting 0.9–1.0) for impressive work; 'B' (weighting 0.7–0.8) for work that was sound but had

identifiable shortcomings; or 'C' (weighting 0.5–0.6) indicating that major improvements were possible.

Each tutorial thus returned a weighted mark between 2 (for [1-1-A], equivalent to 100%) and 0. For example, with announced weighting factors of [A=1.0, B=0.8, C=0.6], an assessment of [1-1-B] = 2×0.8 = 1.6 would correspond to a mark of 1.6/2.0 = 80%, while [1-½-B] would be equivalent to 1.2/2.0 = 60%. At the end of semester, these weighted marks were tallied, and converted to a mark for the tutorial component as a whole. The tutor was not required to do the marks conversion: their responsibility was simply to return the [CP1-CP2-Grade] triad, corresponding to three decision points, each of three options. This process was considered simple enough that a tutor could determine it for each student following a brief review of the submitted workbook.

Example solutions were provided to students on a weekly basis, following the tutorials. This reduced in-class pressure to obtain complete solutions, and allowed the emphasis to be on the process of solving problems. It also encouraged students to revisit the material as a form of self-review.

EVALUATION: IMPROVING STUDENT PERFORMANCE IN SUMMATIVE ASSESSMENT

The merit of the Checkpoint tutorials has been evaluated by comparing student performance in summative assessment under otherwise equivalent conditions. Here, we compare the performance of the 2010 (n = 175) and 2011 (n = 190) cohorts of ZPEM1501 Physics 1A. We concentrate on the first half (Force, Motion and Energy; FME) of this first-semester course for Science and Engineering students. The only difference in teaching and assessment methods across these two instances was in the tutorial component: in 2010, tutorials were not assessed (although attendance was compulsory); while in 2011, tutorials contributed 15% towards the semester grade via the Checkpoint model described in the previous section. The textbook for this course was Halliday, Resnick, and Walker (2011), while FME tutorial questions were based on selections from the workshop tutorials published by Wilson, Sharma, and Millar (2002). Summative assessment for FME involved two 50-minute, 25-question multiple-choice-question (MCQ) tests, broadly sourced from the textbook test bank; students had access to (different) practice MCQs from the same test bank, and most FME lectures used a classroom response system to facilitate peer discussion of other MCQs (Mazur, 1997).

Figure 3 shows a direct comparison of 36 summative assessment questions which were common to the 2010 and 2011 tests. The majority of data points are clearly above the reference diagonal, indicating a clear bias towards better performance by the 2011 cohort. The FME average test performance over the three years 2008 (n = 114), 2009 (n = 106) and 2010 was (54 ± 4)% (yearly standard deviations were about 16%); in 2011, this improved to 67% (standard deviation 14%). Across the entire semester (FME, Thermodynamics and Waves), the 2008-2010 test average was (57 ± 3)% (standard deviations each year about 16%); this improved to 64% in 2011 (standard deviation 13%). In the absence of any evidence to suggest that the 2011 cohort is significantly more academically able than previous years, these results indicate a significant improvement in student performance, with the tutorial model being the only difference of note.

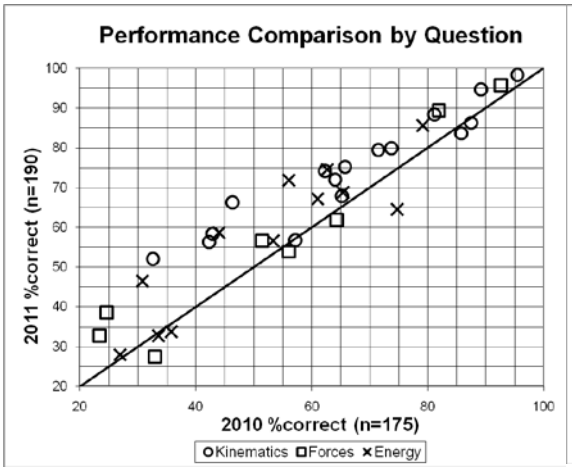

Figure 3: Comparison of student performance across 36 common summative assessment questions in first year mechanics.

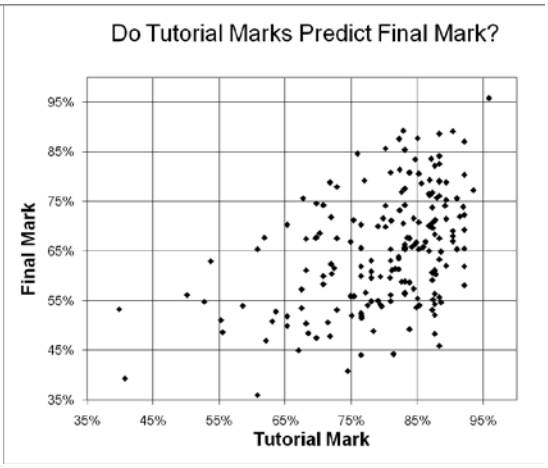

Figure 4: Comparison of tutorial marks and final marks for the 2011 cohort (n=190).

DISCUSSION AND CONCLUSIONS

One matter for future consideration and study involves the variations between the performance expectations of different assessors. In the laboratory environment, where many demonstrators share the assessment load, students will be exposed to a variety of staff, and to a natural human variation in what counts as 'satisfactory', 'good' and 'excellent'. It is useful for students to be exposed to such differences, which they will undoubtedly encounter outside of the academic environment. It is, however, an ongoing requirement to train demonstrating staff to ensure a consistent baseline expectation. Tutorials present more of an issue, as a class usually has the same tutor throughout a semester. Differences between individual tutors thus become more apparent in the final marks distribution, and may require some degree of moderation at semester-end.

The Checkpoint-based continuous assessment model has been shown to be popular with students. The model is relatively straightforward to implement and is not burdensome on staff tasked with applying it in practice. In laboratories, it has encouraged interaction between demonstrators and students; in tutorials, it has encouraged students to prepare and participate without requiring staff to spend an undue amount of time in marking. Overall, it appears to have encouraged and improved student engagement with course material, with linked improvements in student attitudes and results. Black and Wiliam (1998b; p.147) noted that assessment is, '...associated with very powerful feelings of being overwhelmed, and of insecurity, guilt, frustration, and anger.' We suggest that the Checkpoint model helps both staff and students feel better about assessment, arguably one of the most contentious aspects of education in both theory and practice.

REFERENCES

Beun, J. A. (1971). Experiences with a free undergraduate laboratory. *American Journal of Physics*, *39*(11), 1353-1356.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. Assessment in Education, *5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan 80*(2), 139-148.

Foster, G. (2010). Aligning learning and assessment through adaptive strategies in tutorials in physics at the University of Auckland. In *Proceedings of the 16th UniServe Annual Conference*, (pp. 29-34), Sydney, NSW: UniServe Science.

Ganiel, U., & Hofstein, A. (1982). Objective and continuous assessment of student performance in the physics laboratory. *Science Education*, *66*(4), 581-591.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, 3-31.

Gioka, O. (2006). Assessment for learning in physics investigations: assessment criteria, questions and feedback in marking. *Physics Education*, *41*(4), 341-346.

Halliday, D., Resnick, R., & Walker, J. (2011). *Fundamentals of Physics. Extended 9th Edition*: Wiley.

Mazur, E. (1997). *Peer instruction: a user's manual*. Upper Saddle River, NJ: Prentice Hall.

Patterson, J. R., & Prescott, J. R. (1980). Self-paced freshman physics laboratory and student assessment. *American Journal of Physics*, *48*(2), 163-167.

Prescott, J. R., & Anger, C. D. (1970). Removing the 'Cook Book' from freshman physics laboratories. *American Journal of Physics*, *38*(1), 58-64.

Rice, J. W., Thomas, S. M., & O'Toole, P. (2009). *Tertiary science education in the 21st century*. Canberra, ACT: Australian Learning and Teaching Council.

Wilson, K., Sharma, M., & Millar, R. (2002). *Workshop Tutorials for Physics*. Sydney, NSW: UniServe Science.