A novel approach to mathematics examination design and marking

David Easdown, School of Mathematics and Statistics, University of Sydney
Ruth Corran, Department of Mathematics, American University of Paris
Brad Roberts, School of Mathematics and Statistics, University of Sydney

Abstract: *We introduce, motivate and describe a new approach to the design and marking of mathematics examinations. The method is suitable for successfully mapping student performance to a range of grade thresholds in the context of holistic or criterion-based assessment. It uses just one examination to assess and distinguish the performance and achievement of students coming from a bimodal or multimodal distribution with regard to backgrounds, preparedness and aspirations. The method combines and reflects the two phases in the SOLO taxonomy that distinguish deep from superficial learning. This also highlights different learning behaviours and outcomes as students move through and beyond liminal space, in the theory of threshold concepts. The design has been implemented in the School of Mathematics and Statistics at the University of Sydney, particularly with regard to assessment of mathematics units of study taken by large numbers of first year students.*

1. Introduction

Mathematics examinations typically involve extended answer questions or problems that require students to write solutions that demonstrate their knowledge, understanding and ability to communicate and reason mathematically. This article is primarily concerned with a new design and assessment of extended answer questions, especially in the context of holistic theories of learning and criterion-based assessment. Our method has similarities to alternative techniques to traditional grading introduced by Brilleslyper *et al* (2012) and Brilleslyper and Schaubroek (2011), where educators introduce a “points-free” approach to assessment, combining holistic criteria and assessment paradigms that promote deep learning. Our method evolved quickly in response to practical difficulties associated with changes in assessment policies at the University of Sydney in years leading up to 2013, and has been widely used in marking examinations in First Year mathematics at the University of Sydney since 2014.

In Sections 2 and 3, we describe briefly traditional methods of marking, compare norm-referenced and criterion-referenced grading, and give some historical background that led to development of our method. In Sections 4 and 5, we briefly describe the SOLO taxonomy and theory of threshold concepts, which intertwine and provide a theoretical basis for understanding our method and how it applies in evaluating depth of student learning. The method, algorithm and an implementation are presented in Sections 6 and 7, with reference to examination questions provided in the Appendix. In Sections 8 and 9, we make brief comments about ‘hyperbolic’ as opposed to ‘linear’ grading, and discuss the possibility of paradoxes and moderation.

2. Traditional methods of marking mathematics examination scripts

Traditional methods of setting and assessing extended answer questions in mathematics examinations tend to rely on numerical grading, involving simple aggregation of marks. Examination questions take the form of problems or exercises, often broken down into steps or parts, each of which is allocated a specified number of marks. While writing an answer, the student may be aware of the mark value of a question or part of a question, which may be explicitly stated or easily inferred from instructions. Afterwards, the students' answers are assessed by the marker, who allocates marks, often using a marking scheme that has been prepared beforehand by the examiner, or may have evolved in a period of trial marking, or possibly left to the marker's discretion. If a marking scheme changes in the process of marking, then the marker should go back and ensure that the changes are applied fairly across the entire cohort, even if this involves remarking scripts.

The new method, described below, differs from the traditional method primarily in its use of letter grades, rather than numerical marks, utilising qualitative characteristics as well as non-linear aggregations of credit, which are then combined in a novel way when the examination has several questions.

3. Norm-referenced and criterion-referenced grading

The use of pre-determined proportions of a population of students, as the primary basis for the allocation of grades, is an example of *norm-referenced grading*, also known as "grading to the curve". This method aims to achieve fairness of grades between different disciplines, especially with diverse assessment practices and approaches to grading, and to maintain continuity of historical standards.

In the Faculty of Science at the University of Sydney, norm-referenced grading was widespread up until about 2012, and it was expected that, for large cohorts of students taking a unit or study, or a suite of related units of study, as a proportion of the passing cohort, at most

- 5% would achieve High Distinctions,
- 20% would achieve a Distinction or higher, and
- 55% would achieve a Credit or higher.

There was no explicit stipulation about what proportion of the class should pass or fail. The passing criterion had been at the discretion of the academic or team of academics involved in the assessment. If the criterion was not explicit it could, for example, be related to the academic's professional opinion and knowledge of the subject matter, and possibly influenced by historical practice with regard to relative perceived levels of difficulty from year to year. Consider then, in this framework, that a passing grade had been established in a unit of study with large enrolments. Without reference to criteria, one could simply take the rank order of the passing grades and then map, at most, the

- 45th percentile to 65, the minimum final grade for a Credit,
- 80th percentile to 75, the minimum final grade for a Distinction, and
- 95th percentile to 85, the minimum final grade for a High Distinction.

In practice, this may not be so simple, especially if one has to balance numbers of students between units of study that serve different purposes, such as preparation for honours courses. Nevertheless, however one achieves the final grades, the starting point in norm-referenced assessments is to limit the award of higher grades to certain predetermined proportions.

Assessment policies at the University of Sydney evolved over about ten years preceding 2013, moving away from norm-referenced grading. From 2014 onwards, *criterion-referenced* grading was, in principle, universally adopted at the University of Sydney: *Coursework Policy 2014*, item (3) of Part 14, Clause 63 stipulates that

> *Students' assessment will be evaluated solely on the basis of students' achievement against criteria and standards specified to align with learning outcomes.*

How 'evaluation' should take place is left open to interpretation and professional practice in the relevant discipline. Practices within the School of Mathematics and Statistics, leading up to 2014, worked very well: simple criteria were used to determine which raw marks should map to particular grade thresholds. For example, it was common for a raw aggregate of about 40 marks to map to a final Passing grade threshold of 50, on the basis that this achievement demonstrated routine knowledge across a spectrum of topics and important ideas in the course. It was also common that a raw aggregate of about 90 would map to a final High Distinction grade threshold of 85, on the basis that this achievement demonstrated complete or close to complete mastery of the material. Having a flexible raw mark range of about 40-90 mapping to 50-85 had the virtue that academics could set a wide range of challenging assessments tasks, and had plenty of room to make a professional judgment to match raw marks to criteria.

From 2014, however, the practice of mapping raw marks was forbidden. All academic staff were instructed that students' raw aggregates of numerical marks must translate directly into final grades. Thus, an academic had to design assessment so that, at one extreme, a raw total of 50 exactly matched the Passing criteria, and, at the other extreme, a raw total of 85 exactly matched the High Distinction criteria. Only a very narrow range, from 50-85, was available to distinguish the relative performances of students.

This produced a conundrum. On the one hand, assessments had to be easy enough that weak students, who just deserved to pass, would accumulate 50 raw marks: but then many students who did not deserve high distinctions might easily accumulate 85 or more raw marks. On the other hand, assessments had to be difficult or challenging enough that a strong student, who just deserves a High Distinction, would accumulate 85 raw marks: but then many weak students who deserve to pass might not accumulate 50 raw marks. The purpose of the exam design described in this article is to successfully resolve this conundrum and create an exam where the numerical raw marks exactly match, or match as closely as possible, the criteria for the different grades.

4. The SOLO Taxonomy

The SOLO (Structure of Observed Learning Outcome) taxonomy was devised by Biggs and Collis (1982) as a tool for classifying learning and teaching activities and outcomes, and is useful in practical applications of the theory of constructive alignment (see, for example, Biggs and Tang (2007)). It is especially useful for understanding the underlying framework for the new marking method proposed below.

The SOLO taxonomy uses three basic categories to describe the level of a student's understanding or comprehension. In the *prestructural* phase, a student has not properly grasped or understood anything significant related to the subject matter. The level of cognition could be described as amorphous, tending towards chaotic, and without clear identifiable structure or coherence.

In the *quantitative* phase, the student may have grasped or mastered isolated pieces of information or technique, but does not see or understand how these come together to produce a coherent whole. This phase can be broken up into an initial *unistructural* subphase, where a student has successfully focused or mastered just one aspect of the topic, followed by the *multistructural* subphase, where the student manages to focus on more than one, and possibly many, aspects. The second subphase could, in principle, split up into a succession of subphases, as a student includes more and more aspects to his or her repertoire. In a certain sense, the student is aggregating expertise. However, this aggregation remains essentially disconnected and individual items retain, in the mind of the learner, the characteristic of being isolated from one another.

In the *qualitative* phase, the student begins to see how ideas and techniques from the quantitative phase come together to form an integrated whole, where individual parts coordinate, as in an orchestra, to work together to produce a powerful concept or method. There is an effect of precipitation, or crystallisation, where suddenly, or very rapidly, a coherent and complete "whole" emerges, which becomes "greater than the sum of the individual parts". Aspects of learning, arising in the quantitative phase, may seem useless, inanimate or sterile when viewed in isolation (and indeed students often complain "what is the use of this?"), but only derive their power and full sense of purpose from being integrated into a fully functioning, live, complex system.

This higher qualitative phase may be viewed as breaking up into two subphases: the *relational* subphase refers to the initial coming together or integration of parts. It marks a critical point or threshold (and see discussion about threshold concepts below), where the fog miraculously lifts. The student may experience an epiphany moment, where meaning and significance become apparent and the subject matter transforms. The student shifts position from, previously, being a superficial learner to, now, becoming, or having the potential to become, a deep learner and expert. Learning potential may expand rapidly, even explode, as the student moves into the highest *extended abstract* subphase: here integration leads to further conceptualisation, or elevated levels of abstraction and generalisation, giving rise to surprising and spectacular insights, breakthroughs and applications. By moving into, and thereafter inhabiting, the extended abstract phase, the student becomes a master of his or her own learning, reaching at least one major peak or viewpoint, from which other peaks also become visible, and may be in a strong position to then embark on research, exploring new ideas and making discoveries.

5. Threshold concepts and liminality

The theory of threshold concepts was introduced and developed by Meyer and Land (2003) (and see also Meyer and Land (2005) and Land *et al* (2005)), in order to explain and inform processes that lead to successful and deep learning. A *threshold concept* is a key idea or notion associated with a particular discipline that has transformative and integrative properties, opening up pathways or *portals*, to new and otherwise inaccessible knowledge and understanding. One hopes to identify threshold moments, when the student's understanding or perception crystallises, empowering the student. By reaching these new vantage points, entire vistas and panoramas open up, facilitating rapid progress and exploration. In order to move towards and reach such portals, students embark on journeys along pathways that may be problematic, frustrating or troublesome, involving twists and turns, possible backtracking and repetitive behaviour, making and recovering from mistakes. Before reaching a particular portal, a student is said to be in *liminal space*.

The educator's principal task is to create or facilitate an environment in which the student is prompted first to move into liminal space (possibly from an initial state referred to as *preliminal*), and then successfully navigate his or her way through it until the relevant threshold concept is mastered. This can involve a great deal of time and effort. The effect of mastering a threshold concept is so powerful that the changes in the learner's mind become irreversible: "once learned, never unlearnt".

This underlying theoretical framework encourages the educator to move away from a linear, indiscriminating list of topics, and instead focus attention on the most important knotty or problematic aspects of a particular discipline that lead to the most profound, far-reaching and accelerated learning. The theory is relatively undeveloped in mathematics, though there are some exploratory articles (see Easdown (2007), Wood *et al* (2011), Easdown (2011a), Easdown (2011b), Jooganah (2009), Pettersson (2011) and Easdown and Wood (2014)).

The relevance to this article is, firstly, that preliminal and early features of liminal space may correspond roughly to the prestructural and quantitative phases of SOLO. Measures of progress in these phases tend to relate to an accumulation or aggregation of disconnected or isolated skills or pieces of information. Secondly, the act of reaching the portal associated with a given threshold concept, and then unlocking the power of the underlying ideas or techniques, corresponds roughly to moving into the relational and extended abstract phases of SOLO. Measures of success are now expressed in terms of mastery, fluency and depth of learning, and may typically be associated with grades that reflect distinction, honours and research potential.

6. The new method and marking algorithm

In the new method described here, the underlying marking algorithm is the same for all questions on the examination, though there could be question-specific points of interpretation. Each written extended answer receives one letter grade from amongst the following list, in descending order of quality:

AA, **A**, **BB**, **B**, **CCC**, **CC**, **C**, **D**, **E**, **F**, **Z**.

In the implementation described below, this spectrum of letter grades turns out to be practical and adequate. One could add further refinements (for example, introducing **AAA** or **BBB** grades), but doing so risks making the marking algorithm too complicated or difficult to apply rigorously and fairly. The following four letter grades are *superior*:

- **AA** and **A**,

corresponding to a student response within the extended abstract phase of SOLO, where **AA** corresponds to complete mastery, and **A** to almost complete mastery; and

- **BB** and **B**,

corresponding to the relational phase of SOLO, where **BB** corresponds to a student response exhibiting excellence that remains within the relational phase, but does not quite reach the extended abstract phase, and **B** corresponds to a student response that just reaches the relational phase. The *passing* letter grades are the following:

- **CCC**, **CC** and **C**

corresponding to a student response within the multistructural phase of SOLO, indicating routine but meritorious competency across a spectrum of ideas or techniques associated with the topic, but which falls short of entering the relational phase; the measure of quality is indicated by the number of distinct positive attributes in the answer, ranging from four positive attributes for **C**, five positive attributes for **CC**, and six or more positive attributes for **CCC**. The *inferior* letter grades are the following:

- **D**, **E**, **F** and **Z**,

corresponding to a student response within the multistructural, unistructural and prestructural phases of SOLO, indicative of isolated pockets of competence; where **D**, **E** and **F** correspond to three, two and one positive attributes respectively, and **Z** corresponds to a complete absence of positive attributes.

In terms of the theory of threshold concepts, a student whose learning remains in liminal or preliminal space with respect to the topic at hand, is almost certain to receive a passing or inferior grade. By contrast, a student who has successfully navigated through liminal space to reach the portal is likely to perform at a level that receives a superior grade. If the student has made active use of this transformative knowledge then he or she has a reasonable expectation of achieving the highest possible grade.

The marking algorithm to be followed by the marker, for each examination question, is precise and proceeds in two phases:

Marking Algorithm:

1. ***First Phase:*** if the student's answer is of superior quality award a grade of **AA**, **A**, **BB** or **B**, in descending order of quality;

2. ***Second Phase:*** if the student's answer is not of superior quality, then look for an accumulation of distinct positive attributes, awarding a grade of **Z** (for zero), **F**, **E**, **D**, **C**, **CC**, **CCC**, in ascending order, with a ceiling of **CCC** (for six or more positive attributes).

If an answer is of superior quality, then the marker remains only in the First Phase, the Second Phase is avoided and there is no need to try to identify positive attributes. If an answer is perfectly integrated and masterfully written, with at worst only minor or trivial blemishes, then the highest **AA** rating is applied. If there is at least one substantial defect, but the answer is clearly in the extended abstract phase of SOLO, then the next highest rating **A** is applied. If the answer is relational but exhibits sufficiently many defects or omissions so as to not qualify as extended abstract, then either **BB** or **B** is awarded, in decreasing order of quality.

Only if an answer is not deemed to qualify as having superior quality, then the marker moves into the Second Phase, looking for an accumulation of distinct positive attributes. It is important, in this phase, that the marker is not simply trading off negative and positive points or attributes, which typically happens in more traditional marking schemes. A student may write some or a lot of nonsense, that disqualifies him or her from a superior grade, in the First Phase. However, this defective material should not then, in the Second Phase, prejudice the student from receiving at least some credit for exhibiting isolated pieces of knowledge, technique or understanding.

It is important to note, in evaluating the student's work, that the roles of the First and Second Phases are different: in the First Phase, one may find deficiencies that contribute towards the impression that the student is not operating in the relational or extended abstract phases of SOLO (or has not in fact mastered the corresponding threshold concept); yet, the same piece of work, in the Second Phase, may possess one or more distinct positive characteristics that are counted towards a lower grade. This is not a contradiction or paradox, but simply recognition of different phases of learning, and finding and rewarding credit where it is due.

The Marking Algorithm described above, of course, may be supplemented, or fleshed out, by providing feedback, in the case that the assessment has formative properties, to any degree of detail that the educator feels will be helpful, to explain how the final grade was obtained and to support the student's ongoing development and learning.

7. An Implementation

We describe a recent implementation of this method in a summative examination for *MATH1111 Introduction to Calculus*, at the University of Sydney. For MATH1111, and many other mathematics units of study, levels of academic performance are rewarded with the following passing and higher grades:

- **High Distinction (HD):** complete or close to complete mastery of the topic;
- **Distinction (DI):** excellence, but substantially less than complete mastery;
- **Credit (CR):** creditable performance that goes beyond routine, but less than excellence;
- **Pass (PA):** routine knowledge or understanding across a spectrum of ideas or concepts.

When a grade is awarded for an entire unit of study, the student receives a numerical grade within the spectrum 0 to 100, according to the following thresholds and ranges:

- 85-100 for High Distinction;
- 75-84 for Distinction;
- 65-74 for Credit;
- 50-64 for Pass;
- 0-49 for Fail.

In describing this implementation, we are only considering the extended answer section of the MATH1111 examination. A discussion of how this grade combines with other formative and summative assessments, to form final grades for the overall unit of study, raises delicate issues that are beyond the scope of this article. The four examination questions appear in the Appendix, and test a variety of concepts and techniques from the course. They were designed from the point of view that, for any given question, at least four substantial defects or omissions in a student's answer would have the effect of 'disconnecting' it, so that it should not be regarded as evidence of successful learning in the qualitative phase of the SOLO taxonomy. Equivalently, the questions were designed so that at least four substantial defects or omissions in a particular answer translates into evidence that the student remains in liminal or preliminal space and has not successfully mastered the relevant threshold concept or concepts. Thus, in this case, Step 1 (First Phase) of the Marking Algorithm simplifies: the marker awards a superior grade according to how far the answer is from a demonstration of complete mastery, *by looking for up to three substantial defects, errors or omissions.* If there are four or more substantial defects, the marker moves into the Step 2 (Second Phase), and then looks for an accumulation of positive characteristics. In realising the algorithm in this instance, therefore, the markers received the following technical instructions:

Marking Instructions:

1. Draw a line down the side of each page that you have looked at.
2. On the side of the line away from the student's work, place a cross against a substantial error or omission, in the first phase when deciding to award **AA**, **A**, **BB** or **B**. Do not put ticks in the first phase. As soon as you find four crosses, move directly into the Second Phase (no need to spend time locating all errors or omissions).
3. If moving into the Second Phase, place a tick against each distinct positive attribute on the side of the line away from the student's work, when deciding to award **Z**, **F**, **E**, **D**, **C**, **CC**, **CCC**.
4. Do not place any markings on the work itself, only nearby on the side of the line away from the student's work.
5. Place the letter grade at the end of the student's answer for that question, on the side of the line away from the student's work.

These instructions minimise writing on the script, simplify the process of checking the marking later, if necessary, and also assist the student in understanding how the grade was determined, in the case that he or she requests a review or makes an appeal.

A given student would receive four letter grades, one for each of the four questions. To convert these to a numerical aggregate with a maximum score of 4 × 8 = 32, the letter grades were converted to numerical scores as follows:

- **AA**=8, **A**=7, **BB**=6, **B**=5.5, **CCC**=5.5, **CC**=5, **C**=4, **D**=3, **E**=2, **F**=1, **Z**=0.

The equivalent grade thresholds were then given the following natural interpretations, which correspond exactly or very closely to the final minimum numerical grade thresholds:

- **High Distinction** = **A + A** + **A** + **A** = 28/32 = 87.5%,
- **Distinction** = **BB** + **BB** + **BB + BB** = 24/32 = 75%,
- **Credit** = **B** + **B** + **B** + **B** = 22/32 = 68.75%,
- **Pass** = **C** + **C** + **C** + **C** = 16/32 = 50%.

The following points should be noted:

1. The numerical conversion of the **B** and **CCC** grades in this implementation coincide exactly with 5.5. One could of course adjust the conversions so that the passing grades are worth slightly less. For this cohort and academic context, however, even though a student may not have reached the relational phase of SOLO, an answer including six or more distinct positive attributes on a given question was deemed sufficient evidence of meeting criteria for a Credit.
2. A student could produce any combination of letter grades for the four questions, which are then aggregated numerically using the above conversion (and see the discussion below about the possibility of paradoxes). For example, a student could achieve a High Distinction with the combination **AA+AA+A+CC=**28/32,

reflecting the relative levels of difficulty of the questions. Similarly, a student could achieve a Pass with combinations such as **BB+CC+D+E=**16/32, or **A+B+C+Z=**16.5/32.

3. The design of the questions, in increasing difficulty, in this implementation, leads naturally to combinations of letter grades that tend to be nonincreasing. If the examiner, on reflection, feels that the aggregates are too high or low, corresponding to the published criteria, then they can be adjusted. For example, in this exam, Q1 is very easy, and Q4 is very difficult, and for this particular examination, one might have considered **AA+A+A+C** as corresponding to the minimum threshold for a High Distinction and **CC+C+C+F** as corresponding to the minimum threshold for a Pass, and so on.
4. The numerical conversions given above turn out to be convenient and appropriate for this particular implementation (especially using fractions of 8 for each question), but of course one can use any numerical conversion of letter grades (and fractions of some other numbers) that the examiner deems most appropriate for interpretation of criteria. In this particular implementation the numerical translations for Credit and High Distinction fall slightly above the respective final minimal numerical thresholds, whilst the translations for Pass and Distinction produce exactly the respective final numerical thresholds. This produces a slight numerical bias against awarding Passes and Distinctions, though the examiners in this case felt for this exam that was appropriate. For a more difficult examination, one might aim for numerical translations that land further inside the range of the particular grade, well above the minimum threshold (and see the previous comment).
5. This example uses four questions. Obviously the method can be adjusted for any number of questions. It simplifies considerably if, for example, there are just one or two questions in an in-term assignment, for which the method could be used to award letter grades. In the case of a formative assignment, the instructions to the markers can, of course, be supplemented to include the provision of detailed feedback.
6. This implementation was for a major piece of summative assessment, marked carefully and quickly, involving several hundred students. It was not expected that students would normally receive or expect feedback. However, formative aspects arise naturally when students contact the lecturer to have explained to them how their grades were determined.

8. Hyperbolic versus linear marking

In traditional marking of students' work, numerical marks tend to accumulate linearly, or approximately linearly, when equivalent work, effort or insight across the assessment task correspond to equivalent marks. Our new method of marking deviates from 'linearity' in at least the two following respects:

1. There are two distinct marking phases, that distinguish performance or achievement in the two main phases of SOLO (and corresponding learning spaces in the theory of threshold concepts). Our new method is particularly adept at successfully assessing a student cohort that may be bimodal or multimodal with just one examination. The *deep* learners, or those who have mastered the relevant threshold concepts, tend to receive combinations of

superior grades. The *superficial* learners tend to receive combinations of passing or inferior grades. Though the effects of aggregation of grades across multiple questions may cause some blending, the marking design recognises a qualitative leap that is less obvious in traditional 'linear' marking.

2. In the Second Phase, the marker is looking for positive attributes, and the accumulation is not expected to be linear. The first positive attribute, to achieve at least an **F**, finding any response from the student that deserves credit, may come easily. Each subsequent distinct positive attribute may become more difficult to achieve. There is an absolute ceiling in the Second Phase of a **CCC** grade, representing at least six positive attributes. One may describe this accumulation of credit as 'hyperbolic', never able to formally break over the Pass boundary. In the numerical conversion in the implementation described earlier, each of the first five distinct positive attribute contributes one unit each and the sixth positive attribute, if found, contributes an extra half of one unit, reflecting this nonlinear tapering of credit in the Second Phase. One could, of course, make this tapering more precise numerically.

9. Paradoxes and moderation

A *paradox* occurs if a given assessment produces a higher grade for a weaker student than for a stronger student. There are many natural reasons why paradoxes might occur in any assessment setting, especially related to a student's attention to the task at hand, concentration, health, distractions or accidents. The most serious paradoxes occur however if there are fundamental flaws in the design of the assessment.

As explained in the previous section, the accumulation of credit is not linear. However, if the examination comprises only one question, then the following rank order of grades is achieved:

$$\mathbf{Z < F < E < D < C < CC < CCC \leqq B < BB < A < AA}$$

In the numerical conversion used in the implementation described in this article, **CCC = B**, but one could use a conversion where **CCC < B**.

There is no apparent paradox here: if the single question is well-designed, one expects the grades to correspond to the learning achievement of the student in the appropriate phase of SOLO or space in the theory of threshold concepts.

The possibility of paradoxes arises in aggregation when the examination comprises more than one question. Consider, for example, an examination with four questions, as in the Appendix, with the numerical conversion described in the earlier implementation. Suppose the learning of Student X has not progressed beyond the quantitative phases of SOLO, takes this exam and achieves grades **CC+CC+CC+CC**=20/32, in the order of those questions (from easy to difficult). Suppose Student Y and Student Z have learning achievements in the relational or extended abstract phases of SOLO with respect to most topics and achieve grades **AA+AA+E+F**=19/32 and **D+D+BB+A**=19/32 respectively on the same exam. It might not be fair to claim that Student X deserves a higher grade than Student Y or Student Z. The performance of Student X is well-balanced, consistently within the passing spectrum, possibly spending roughly equal amounts of time on the questions, picking out easier parts to answer. Student Y, by contrast, may have worked very hard and long on the first two questions, providing perfect model answers, and then ran out of time to properly attempt the last two questions. Student Z, by further contrast, may have attempted the questions in the reverse order of difficulty, investing time in the most difficult parts of the exam and ran out of time or energy to fully attempt the easiest parts.

Such apparent inconsistencies are typically related to time pressures in an exam. The best way to avoid this, would be to provide more than enough time for any student to thoroughly attempt every question. However, if there is a difficult question (like the last question in the Appendix) that could swallow up any amount of time, then even generous time allowances may still result in a paradoxical result, such as that of Student Z. The examiner must therefore have moderation strategies. One might consider adjusting weightings in the aggregation, that recognise the relative difficulty of questions, even just for students for which the examiner is sure the outcome has produced an unfair result. However, this may introduce complications, especially in explaining to students how their final grades were calculated, and also in ethical considerations and consistency when telling students in advance the mechanism for determining final grades.

The most effective solution, if one suspects that a paradox has occurred, would be to withhold the examination result for a particular student, until the examiner has had a chance to enquire about his or her performance in the exam, find out for sure what happened, and perhaps probe at interview the student's knowledge and understanding of particular questions, or ask the student to resit another version of the exam or part of the exam. At the end of any moderation process, the ultimate aim is to provide students with fair grades that reflect their learning achievements and do not disadvantage any other student.

Appendix: Exam Questions

The following four exam questions were used for MATH1111 Introduction to Calculus in First Semester 2017. The first question is routine, and tests basic algebraic manipulation of a cubic polynomial function, integration and the following threshold concepts:

- application of derivatives to curve sketching;
- application of even and odd functions to integration.

The second question is easy but less routine, testing the following threshold concepts:

- the Fundamental Theorem of Calculus;
- tangent line approximations to a curve.

The third question is harder, in two unrelated parts, both of which involve some nontrivial aspects of problem solving and the following threshold concepts:

- Riemann sum approximations;
- concavity;
- application of calculus to minimisation.

The fourth question is a difficult modelling exercise, and tests the following threshold concept:

- the Chain Rule,

embedded in an application, requiring ability to interpret information and fluency with pronumerals.

Extended Answer Section

Answer these questions in the answer book(s) provided.
Ask for extra books if you need them.

1. Consider the function f defined by the rule

$$f(x) \;=\; x^3 - 12x\ .$$

(a) Evaluate $f(0)$, $f(2)$ and $f(-2)$.

(b) Factorise the expression $x^3 - 12x$ and therefore find all x such that $f(x) = 0$.

(c) Find $f'(x)$ and $f''(x)$ and draw sign diagrams for each of them.

(d) Find the local maximum and minimum values taken by $y = f(x)$.

(e) Sketch the curve $y = x^3 - 12x$. Locate the point of inflection.

(f) Is the function $y = f(x)$ even, odd or neither. Explain briefly.

(g) Use the Fundamental Theorem of Calculus to find

$$\int_0^2 f(x)\, dx\ .$$

(h) Use your answers to the previous two parts, or otherwise, to evaluate

$$\int_{-2}^2 f(x)\, dx \qquad \text{and} \qquad \int_{-2}^0 f(x)\, dx\ .$$

2. (a) Sketch the areas represented by the following definite integrals and evaluate them exactly:

(*i*) $\displaystyle\int_0^{\pi/2} \cos x \, dx$. (*ii*) $\displaystyle\int_1^{\ln 10} e^{-x} \, dx$.

(b) Sketch the following curves:

(*i*) $y = x^3$ (*ii*) $y = x^{1/3} = \sqrt[3]{x}$ (*iii*) $y = \sqrt[3]{x+1}$

(c) Show that the tangent line to the curve

$$y = x^{1/3}$$

at the point $(8, 2)$ has equation

$$y = \frac{x}{12} + \frac{4}{3} .$$

(d) Use the tangent line from the previous part to estimate

(*i*) $\sqrt[3]{9}$ (*ii*) $\sqrt[3]{7}$ (*iii*) $\sqrt[3]{8.2}$

Evaluate $\sqrt[3]{8.2}$ also directly using your calculator, and comment on the accuracy of the tangent line estimation in part (*iii*).

3. (a) A car travelling initially at 27 m/sec comes to rest five seconds after the driver applies the brakes. The following velocities are recorded:

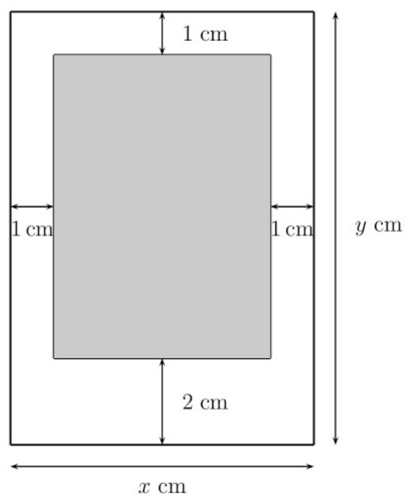
time since brakes applied (sec)	0	1	2	3	4	5
velocity (m/sec)	27	16	9	4	1	0

(i) Draw a rough sketch and use it to decide if the curve of velocity against time is concave up or concave down.

(ii) Give lower and upper estimates of the distance the car travelled after the brakes were applied.

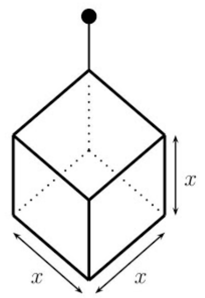
(iii) Take the average of your lower and upper estimates of the distance travelled. Do you expect this to be an overestimate or underestimate of the true distance travelled? Explain briefly.

(b) A rectangular poster is to have an area of 180 square centimetres with 1 cm margins at the top and sides and a 2 cm margin at the bottom.



Find the dimensions (width x cm and height y cm) of the poster that maximise the printed area (that is, the area inside the margins, shaded in the diagram).

4. A decorative ice sculpture in the shape of a perfect cube is formed using 100 litres of water, and suspended from the ceiling of a gallery. The cube is in a warm room maintained at a uniform temperature. It is melting away and after 3 hours have elapsed, 10 litres of water have been collected in a tray beneath the cube, so that, at that moment, exactly 90% of the volume of the cube remains.



Denote the side-length of the cube by x units, where

$$x = x(t)$$

is a function of time t, where t is measured in hours from the moment the cube is first suspended. Let $V = V(t)$ denote the volume of the cube at time t, so clearly

$$V = V(t) = x^3 .$$

(a) Explain briefly why

$$A = A(t) = 6x^2 ,$$

where $A = A(t)$ denotes the surface area of the cube at time t.

(b) Throughout the melting, the rate at which the volume of ice is melting is proportional to the surface area, that is,

$$dV/dt = kA$$

for some constant k. Use this fact to show that

$$dx/dt = 2k .$$

[Hint: apply the Chain Rule.]

(c) Explain briefly why

$$x = 2kt + C$$

for some constant C.

(d) Show that

$$C = \frac{6k}{\sqrt[3]{0.9} - 1} .$$

[Hint: $V(3) = \frac{9}{10}V(0)$.]

(e) Show that the ice cube disappears completely after 87 hours, to the nearest hour.

References:

Biggs, J.B. and Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy.* New York: Academic Press.

Biggs, J.B. and Tang, C. (2007). *Teaching for Quality Learning at University, 3rd Edition.* New York: Society for Research into Higher Education & Open University Press.

Brilleslyper, M. and Schaubroeck, B. (2011). Alternatives to Traditional Grading to Enhance Student Learning. Minicourse, MAA MathFest, Mathematics Association of America, Lexington, Kentucky. [online: https://www.maa.org/mathfest-2011-program].

Brilleslyper, M., Ghrist, M., Holcomb, T., Schaubroeck, B., Warner, B. and Williams, S. (2012). What's the point? The benefits of grading without points. *Problems, Resources and Issues in Mathematics Undergraduate Studies* 22(5), 411-427.

Easdown, D. (2007). The role of proof in mathematics teaching and the Plateau Principle. *Proceedings of the Assessment in Science Teaching and Learning Symposium*. Alexandra Hugman, Ian Johnston and Mary Peat (eds.). Sydney: Uniserve Science.

Easdown, D. (2011a). Transforming practice using threshold concepts. ALTC Professional Development Workshop. University of Wollongong, September 29-30, 2011. [online: http://www.austms.org.au/ALTC+Workshop+2011].

Easdown, D. (2011b). Excursions to and from semantic oblivion. *Proceedings of Volcanic Delta 2011*, Volcanic Delta 2011: The Eighth Southern Hemisphere Conference on the Teaching and Learning of Undergraduate Mathematics and Statistics. John Hannah and Mike Thomas (eds.). University of Canterbury.

Easdown, D. and Wood, L. (2014). Novel threshold concepts in the mathematical sciences. *Threshold Concepts: from personal practice to communities of practice.* Proceedings of the National Academy's Sixth Annual Conference and the Fourth Biennial Threshold Concepts Conference, National Academy's Sixth Annual Conference and the Fourth Biennial Threshold Concepts Conference, Catherine O'Mahony, Avril Buchanan, Mary Rourke and Bettie Higgs (eds.), NAIRTL (Irish National Academy for Integration of Research, Teaching and Learning), Cork, Ireland, 44–50.

Jooganah, K. (2009). Proof as a threshold concept for university mathematics: an exploration of student identity and transition. British Education Research Association Annual Conference. University of Manchester.

Land, R., Cousin, G., Meyer, J.H.F. and Davies, P. (2005). Threshold concepts and troublesome knowledge (3): implications for course design and evaluation. *Improving Student Learning – equality and diversity* (C.Rust, Ed.). Oxford: OCSLD.

Meyer, J.H.F. and Land, R. (2003). Threshold concepts and troublesome knowledge - linkages to ways of thinking and practising. *Improving Student Learning - Ten Years On* (C. Rust, Ed.). Oxford: OCSLD.

Meyer, J.H.F. and Land, R. (2005). Threshold concepts and troublesome knowledge (2): epistemological considerations and a conceptual framework for teaching and learning. *Higher Education* 49(3), 373-388.

Pettersson, K. (2011). The threshold concept of a function – A case study of a student's development of her understanding. Stockholm University. [online: http://www.mai.liu.se/SMDF/madif8/Pettersson.pdf].

Wood, L., Vu, V., Bower, M., Brown, N.R., Skalicky, J.L., Donovan, D., Loch, B., Joshi, N. and Bloom, W. (2011). Professional development for teaching in higher education. *International Journal of Mathematical Education in Science and Technology* 42(7), 997-1009.