Evaluating a pass/fail grading model in first year undergraduate computing.

ZARB, M., MCDERMOTT, R., MARTIN, K., YOUNG, T. and MCGOWAN, J.

2023

Evaluating a Pass/Fail Grading Model in First Year Undergraduate Computing

Mark Zarb
School of Computing
Robert Gordon University
Aberdeen, UK
m.zarb@rgu.ac.uk

Roger McDermott
School of Computing
Robert Gordon University
Aberdeen, UK
roger.mcdermott@rgu.ac.uk

Kyle Martin
School of Computing
Robert Gordon University
Aberdeen, UK
k.martin3@rgu.ac.uk

Tiffany Young
School of Computing
Robert Gordon University
Aberdeen, UK
t.young3@rgu.ac.uk

Jess McGowan
School of Computing
Robert Gordon University
Aberdeen, UK
j.mcgowan4@rgu.ac.uk

***Abstract*—This Innovative Practice Full Paper investigates the implications of implementing a Pass/Fail marking scheme within the undergraduate curriculum, specifically across first year computing modules in a Scottish Higher Education Institution. The motivation for this implementation was to ease stress and pressure on students entering higher education, which became particularly relevant following the COVID-19 pandemic. The study reports on the results of a survey that gathered feedback from Stage 1 and Stage 2 students who experienced the Pass/Fail implementation, and results shows that students generally appreciate the Pass/Fail model, although for many, the benefits only become apparent once they are exposed to alternative grading models. A number of recommendations are made for the implementation of similar marking schemes within computing in Higher Education curricula.**

Keywords—grading systems, pedagogy, assessment, cs1, computing, pass/fail

I. Introduction

The COVID-19 pandemic necessitated emergency changes to teaching practices in higher education around the globe. While responses differed at both national and institutional levels, students invariably needed to adapt to a mix of online teaching, blended learning, and hybrid learning in quick succession over the course of two years. During this period, higher education institutions sought to improve the student experience through leveraging new and varied pedagogical tools and styles. One area which saw significant research during this period was the exploration of different assessment methods and structures as a way of enhancing the student experience.

This paper presents a study which investigates the implications of implementing a Pass/Fail marking scheme within the undergraduate curriculum; specifically across first year computing (CS1) modules in a Scottish HEI. The implementation of this marking scheme is motivated by the goal of easing stress and pressure on students entering higher education. The results of a survey are reported, where Stage 1 students (who are experiencing the Pass/Fail implementation for the first time) and Stage 2 students (who have experienced the Pass/Fail implementation in their previous year of study and have now transitioned to a more traditional letter-grading system) react to this implementation.

The subsequent discussion shows that students generally appreciate the Pass/Fail model as a way to ease their entry into higher education, although for many, the benefits only become apparent once they are exposed to alternative grading models. Finally, we present a set of recommendations for any instructors who wish to implement similar models within their curricula, and set the stage for future work.

II. Background

A. Assessment and Grading

Assessment, its aims and objectives within an academic programme, its operational constituents, and the conclusions that are drawn from its outcomes are all much debated subjects within higher education [1-3]. A full discussion of all these substantial and important topics would not be possible within this paper. Instead, we will focus on the concept of grading [4-6], i.e., the assignment of values or scores to a student's work or performance based on a set of desired learning outcomes. For the purposes of this paper, we take learning outcomes to be a set of predetermined criteria which express, with varying degrees of specificity, what a learner should be able to do at the end of a period of instruction. The purpose of grading is to evaluate the student's overall performance and to provide a measure of achievement in a particular subject area or success in a particular course. Grading is therefore not synonymous with assessment, but a successful grading system should seek to further the aims and objectives of educational assessment. Consequently, to identify the criteria for a successful grading scheme, we need to clearly articulate the purpose, or purposes, of assessment. We start this process by discussing some of these issues, before looking at examples of grading schemes and how they facilitate the objectives of the assessment process.

B. Purposes of Assessment

Viewed from a general educational perspective, assessment is a rich area of study, with many interrelated academic, social, and cultural purposes and a range of subtle and nuanced outcomes and effects. Nevertheless, we can identify a number of broad purposes of the assessment process, consideration of which will be useful for subsequent discussions of grading. The first of these is the intuitively obvious one of trying to evaluate student competence in a specific subject area, i.e., the knowledge, skills, and dispositions that students have acquired over a course of study. This can be done in a variety of ways but in a university context, it is usually accomplished by

measuring student performance against specific learning objectives or standards. A key concept in this context is measurability, with student performance being evaluated against the criteria set out in the learning objectives, allowing monitoring of developmental progress. The concept of measurability itself introduces notions of validity and reliability, with validity referring to the extent to which an assessment measures what it is intended to measure, and reliability concerning the consistency of the assessment outcomes. Validity and reliability are both essential to ensure that the assessment results are meaningful, and consequently useful, in the context in which the assessment is being carried out, but they are also necessary when providing performance indicators to stakeholders. Accurate and robust measures of assessment are needed for feedback purposes, with results often used to motivate students to engage more deeply with the course content and put in the effort required to succeed [7]. They also provide a reward mechanism to encourage high performance and a deterrent to poor learning behaviour [8]. Transparency in the assessment process is required because students are more likely to be motivated when they understand what is expected of them and when they receive feedback on their progress [9].

Fairness is also a fundamental characteristic of good assessment practices [10]. While this is linked in part to issues of validity and reliability, the requirement for fair and equitable assessment goes far beyond these operational concerns. It is clear that an assessment, which failed to measure what it purported to do, would lead to students believing that they have attained a level of proficiency when this was not in fact the case. This would disadvantage individual students, and in some circumstances would be dangerous, but it would also be unacceptable to the civic institutions which provide public funding for specific purposes and which would also have been misled. Poor reliability, as indicated, say, by lack of consistency or reproducibility of results, would also disadvantage students who submit work similar to their peers but receive dissimilar grades. While these examples of unfairness are important, issues of equality in assessment have social and ethical implications beyond this. Fairness would demand that assessment processes should not disadvantage individual students or groups of students because of characteristics which are not relevant to the outcomes being evaluated [11]. The advent of mass higher education has been accompanied by an increase in student diversity, with learners from previously underrepresented or marginalised social groups becoming more represented and visible within institutions [12]. Educators are subject to the same sorts of cultural pressures and biases as other members of society and this needs to be considered when devising assessments which may be taken by a broad and diverse range of students. Similarly, with the increasing internationalisation of higher education, assessment of culturally diverse student cohorts has become a more complex and challenging issue. Different cultures and educational systems may have different expectations and approaches to assessment, and these can create challenges for international students and for institutions that are seeking to attract and retain those students [13].

Assessment also provides a mechanism for the dissemination of feedback to both students and teachers on areas of strength and weakness [14]. Students can use this feedback to identify parts of the curriculum where they need to improve and so make appropriate adjustments to their learning strategies, while teachers can use feedback to identify areas where adjustments need to be made to teaching strategies and delivery methods, to better support student learning. Data from assessment outcomes informs decision making at all levels, from that of the individual student thinking about course selection, to programme-level decisions about curriculum development, institutional decisions about the allocation of resources in response to political and social pressures, and decisions by national governments about political and economic strategy [15]. The distribution of appropriate feedback to students is one way in which assessment closely aligns with the pedagogical role of “assessment for learning”. This focuses on supporting and motivating students in their learning journey rather than the certification role of measuring outcomes to provide a summative measure of attainment (assessment “of learning”) [16].

At the societal level, assessment data is used to demonstrate that degree programmes are meeting accreditation and certification standards [17]. Overseeing and managing the process of certification, i.e. the formal recognition of that competence used as a mark of entry into a new developmental level or a specific field or profession, is a fundamental way in which educational institutions demonstrate accountability as part of their social contract with the state [18]. For example, data from assessment processes can be used to provide evidence in support of satisfaction of political and social demands that educational institutions provide an effective workforce [19], as well as inform a population which can contribute to the civic and cultural life of the society [20]. A related but distinct purpose for assessment is to provide a ranking mechanism either for learners themselves, or for the output of learning. Whereas certification seeks to ensure minimal requirements, the use of standardised assessments aims to create an order of proficiency within a cohort, or between similar groups of learners, usually through the deployment of some kind of standardised testing. Examples of this include the use of national tests for university entrance or the calculation of grade point averages.

In summary, assessment is used for a number of different purposes, the main ones being evaluation of competence, to allow for certification and ranking of students, to provide data to demonstrate organisational transparency and inform decision-making at a range of levels from the individual student to society at large, and to promote good educational habits among students. The main problematic issues around assessment centre on the concepts of validity and reliability, and on the issue of fairness. Any discussion of grading should indicate how the grading system used should relate to these issues. There is obviously much more that could be said on the subject of assessment but we will now turn to the specific issue of the nature and purpose of grading.

C. Grading

Grading is usually defined in terms of the attribution of meaningful symbols related to student performance to individual pieces of work [4] to indicate levels of competence. Grades are clearly meant to represent some summative measure of achievement in a student’s course of study, but beyond that, they are important because of their close connection to wider aspects of the student experience

[21]. They act as predictors for future educational performance, such as admission to, and success in, higher education [22-24]; as well as correlating with measures of educational disengagement [25]. Sometimes this process of assigning grades is known as "marking" and the summary achievement measure is known as a "mark". In this paper, we treat the terms grade and marks as synonymous.

In their review of grading research, Brookhart et al. [4] state that the central question in research on grading is "what do grades mean", or at least, whether there is evidence to support "the intended meaning and use of grades as an educational measure". While differing levels of student performance at the end of a period of study cannot be reasonably disputed, a number of aspects of grading have proved controversial. The issues are broadly parallel to those discussed more generally in relation to assessment, namely the methodological basis of assessment, the validity of the concept, the reliability of the assignment process, both from an operational and ethical perspective, and the affective impact on the learner experience.

With regard to methodological basis, grading is usually taken to be either norm-referenced or criterion-referenced [26]. Norm-referenced grading refers to assignment of a summary achievement statistic based on performance of the individual student relative to the population of those being assessed, whereas criterion-referenced grading evaluates performance against a set of pre-specified qualities or criteria, without reference to the achievement of others [27]. The requirement that assessment, especially at university level, should provide a basis for certification means that criterion-based grading is more common, as it is difficult, and in some cases, practically impossible, to gather appropriate data on a population, with respect to which normative grading could be performed. However, in the absence of sufficient clarity concerning the criteria themselves, criterion-based grading has a tendency to devolve into normative grading of the specific assessment cohort, significantly limiting its effectiveness. We note in passing that a third alternative to norm-based and criterion-based assessment is ipsative assessment [28-29] , which seeks to compare current student achievement with past individual performance. This is not a common form of assessment but has been an object of enquiry in recent years [30-31].

The second aspect of grading that deserves attention is the validity issue, i.e. whether a grade really represents a measure of learning. A grade purports to be a summative indicator of achievement expressed as a single token. However, it is uncommon for modules or course units within a programme of study to have just one learning objective or outcome and so assessments rarely measure just one element of competence. This leads to questions about how the significance of different learning objectives are combined within a single assessment, and how this aggregation is reflected in a single summary statistic. In addition to problems with the inclusion of multiple learning objectives, it may also be the case that assessments include factors that do not indicate achievement in the domain they intend to measure, e.g., where the overall grade includes elements that either implicitly or explicitly give consideration to surface-level features of an assignment such as formatting of text. Even when tightly constrained by assessment rubrics based on achievement of stated learning objectives, there is some evidence that some assessors incorporate appreciation of affective factors such as the degree of effort, motivational elements, and other academic enablers when determining grades [4; 32] . These factors bring into question the validity of the grade as a summative indicator of learning.

Finally, there are issues of reliability, i.e. the consistency of the grade statistic and its assignment process. In order to be fair, grading should be internally consistent, i.e. grades assigned for similar pieces of work within the same assessment diet should be similar [33-34]. However, given that the results of assessment also provide justification for certification, there is an external consistency requirement which has both a locational and chronological component. For such assessments, grades for similar work - should, all things being equal - be similar, regardless of where and when the assessment took place.

D. Types of Grading System

Before detailing the grading system that is the focus of this paper, we give a brief overview of the main forms of grading scheme that can be found in higher education. These usually fall into two types, categorical or numerical, depending on whether the grade is a qualitative symbol (e.g. A to F) or quantitative score (e.g. a percentage).

Although the earliest grading schemes for individual pieces of work (rather than, say, an overall classification for a degree or programme) were based on categorical scales [35], over the last hundred years, university education has seen widespread adoption of numerical scales, often based on some percentage score. Percentages provides a simple scale to describe results (e.g. 0 to 100) and allow for the easy identification of a single cut-off point for success (e.g. 40%). It also affords a straightforward mechanism for giving weights to different elements of the assessment based, for example, on perceived significance to the learning outcomes, or time required for completion (e.g. Q1 is worth 5%, Q2 is more important and so is worth 20%, ...). The method of combination for these subscores is confined to simple arithmetical addition. This, in its basic model at least, gives rise to features which may or may not be a desirable feature of the assessment, e.g. good performance in one part of the assessment will automatically compensate for poor performance in another. Also, while not an inherent feature of a numerical grading scheme, the ubiquity of those based on percentages may lead to an issue with the assumed precision of the assigned scores or subscores, i.e. whether there is a significant difference between performance if numerical scores differ by a few percentage points, and whether the assessment has been constructed so that the markers are able to reliably and consistently make judgements of competence based on such margins. This issue is exacerbated when a categorical system is superimposed over the numerical scheme, e.g. to provide summative letter grades for feedback. In this case, small differences at grade boundaries give rise to significantly different grades, especially when rounding occurs.

Categorical grades, e.g. letter grades from A to F, are also often used to provide summary information on performance. While the process of assignment of categorical symbol grades should be qualitatively different from that of numerical scores, in reality, the more symbols are used on a single achievement scale, the more the former resembles the

latter, especially if there is some kind of mapping between the symbolic grade and some overall quantitative aggregate, such as a percentage range or a grade point score. While it is possible to retain some categorical character to the grade on, say, a six-point A to F scale, it is more difficult to do this when the number of categories exceeds twenty. For example, some universities in the UK routinely use twenty-three to twenty-five point categorical scales combined into seven or eight bands based on A to G or H grades [36-37]. The twenty-five point scale includes five grade A subdivisions, three divisions at grade B and C, four divisions at grade E and F (which include separate resit pass/fail grades at postgraduate level) and three G grades which denote various kinds of minimal or non-submission. Moreover, these bands, and the grades within them, are then mapped onto a numerical grade point ranging from 22.00 to 0.00. Given that the purpose of the grade point mapping is to allow the calculation of a numerical grade point average which characterises course-level performance, and that this is done by averaging the individual grade points for modules or course units, it is clearly challenging to ensure that this kind of grading scheme does not devolve into a proxy for a numerical scheme that has slightly less granularity than the more conventional percentage scoring system. Note that from the assessment regulations associated with these grading schemes, it is clear that the different grades within each band are not meant to measure different subcomponents but only the degree to which the piece of work has achieved a level of attainment greater than or less than the midpoint of the band. All grades are commensurable and we thus have a simple linear grading scale.

Even when fewer categorical grades, say A to F, are used, there may be some element of formal or informal mapping to a nominal quantitative scale, e.g. grade A is mapped to scores above 70%, grade B to scores in the range 60% to 69%, C in the range 50% to 59%, etc. One way around this reliance on pseudo-numerical grading is to use grade profiles where assessment components are given categorical subgrades and the aggregation of these subgrades to the overall assessment grade is performed using a grade profile. For example, if an assessment task involves the assignment of eight subgrades in the range A to F for assessment subtasks, the overall grade would be calculated by specifying minimal grade counts. An overall grade A might be awarded if the grade profile was, say, equal to, or exceeded, 4 grade As, 2 grade Bs and 2 grade Cs, an overall grade B would be awarded if the student did not have the requisite subgrades for a grade A but had achieved a minimal threshold of 4 grade Bs, 2 grade Cs and 2 grade Ds. More important assessment components can be accommodated by assigning them some form of higher integer weighting. Such a system retains the categorical nature of the grading scheme but is clearly challenged by edge-cases, e.g. a student who achieves subcomponents 7 grade As and a grade D for the assessment would appear to be have a grade B profile despite a significant preponderance of A subgrades in their submission.

We note that a limiting case of the assignment of both numerical and categorical grades is a binary scheme based on assignment of either a pass or a fail grade. This presents the marker with a straightforward choice about whether or not the student's work has satisfied the minimal conditions necessary for success in the assessment.

Given the range of grading schemes, we can articulate some general operational characteristics that should apply. Firstly, a grading scheme should allow for the evaluation of competence in whatever context the assessment takes place. At this stage, we do not state how they should do this but only that this feature is clearly an essential and necessary requirement. Such an evaluation must be present for certification purposes and also for the sensible return of feedback to learners. Secondly, for validity purposes, the grading scheme must evaluate either a direct demonstration of competence or clearly defined and professionally agreed proxies for that competence. Thirdly, the range of the grades must allow suitable distinction to be made between various appropriate degrees of proficiency. 'Appropriate' here means that the inferences about precision should be transparent and should neither be too limited nor excessive. Fourthly, both the grades used and the assignment process used by teachers should foster appropriate pedagogical goals, such as supporting assessment for learning, provision of useful feedback, etc. Fifthly, the reception of grades by students should encourage appropriate educational dispositions, such as encouraging good learning habits and providing motivation for continued engagement. Finally, the grading scheme should be fair and not disadvantage any individual or group based on factors that are not being assessed.

E. The Pass/Fail (Binary) Grading Scheme

The Pass/Fail system of grading is any scheme in which the assessor evaluates a piece of work and comes to a decision about whether it satisfies the criteria for minimal success given the learning objectives of the assessment. This can, of course, be seen just as an extra coarser-grained wrapper placed upon a more fine-grained categorical or numerical grading system, in which case it adds very little to the process and loses important information about levels of proficiency in assessment performance. However, a more interesting and valuable example of this kind of grading is when it is applied to the evaluation of work in so-called holistic assessment.

The term "holistic assessment" (and by extension, holistic grading) has been used in a number of ways in educational research and practice over the past fifty years [38-39]. These include the assessment of "holistic" competencies (i.e. what are often called "soft skills") [40] as well as a mechanism for the assessment of writing and oral presentation skills [41]. In both cases, the "holistic" epithet refers to an assessment process which seeks to give an evaluation of work based on an academic judgement of its overall merit rather than using some reductive procedure. Note that there is nothing about these forms of holistic assessment which dictates the use of a binary marking scheme, but, as we will demonstrate, the use of a non-trivial implementation of a Pass/Fail grading system necessitates an evaluation of work based on holistic principles. This does not mean that individual components of the assessment cannot be marked separately and then aggregated into a final summative grade but rather that any such aggregation process must take into account the way that different elements combine into an integrated whole.

III. Method

A. Institutional Context

The Robert Gordon University is a higher education institution based in Aberdeen, Scotland. This study was carried out in the School of Computing, which offers a range of undergraduate programmes with Stage 1 entry (including BSc Computer Science, BSc Computing and Creative Design and BSc Cyber Security). These undergraduate programmes share a foundation year (referred to in this paper as "Stage 1"), where all students study the same core modules, then select electives to complement their chosen course. Completion of the foundation year awards students with 120 SCQF (60 ECTS), and allows them to progress to Stage 2, which is more tailored to their chosen course of study.

As of the 2021-22 academic year, the design of Stage 1 was updated to incorporate a Pass/Fail assessment scale across all assessments and modules, thus promoting a greater focus on feedback rather than grades. This was designed and implemented with the intention of allowing students more flexibility and creativity in how to complete their work beyond targeting minimum requirements, therefore better preparing students for more granular grade distribution from Stage 2 onwards. Furthermore, the use of a Pass/Fail model, with its constrained and binary use of grading outcomes, allows for a greater focus on feedback. This approach helps to shift the focus away from a narrow focus on letter grades and towards a more comprehensive understanding of the student's strengths and weaknesses.

For each module of study, students would have been presented with coursework in week 1 of the semester, accompanied by a marking grid that adhered to a rigid template. This template provided guidance on the requirements for each assessment, as well as the guidance on what would constitute a Pass grade, and guidance on what would constitute a Fail grade.

The purpose of this study is to understand students' perception of this model, both from a Stage 1 perspective, and from a Stage 2 retrospective. Evaluating the model from the student perspective allows us to inform future pedagogy and refine how the model is implemented in future academic cycles. We use these findings to make recommendations for other instructors looking to implement similar models.

B. Survey Design

A survey was created via Microsoft Forms by the researchers and validated by the School's Foundation Year Coordinators. It was distributed via mailing list to all Stage 1 and Stage 2 students, with no remuneration for its completion.

A copy of the survey can be seen in Appendix A. First, students were asked to report which Stage they were reflecting on (Q1) in order to allow the researchers to slice the data. No demographic data was collected. Whilst the Microsoft Form required students to log in for validation purposes, in order to preserve anonymity this information was not saved, therefore individual responses can not be attributed to individual students. In the second section, the researchers included three questions which used a three-point Likert scale for students to select the most appropriate response for their situations. The questions covered: the level of effort students applied to assessments with Pass/Fail grading (Q3), the impact of Pass/Fail on their stress levels (Q7), and how Pass/Fail influenced expectations for subsequent study (Q8). Students were given the opportunity to expand upon each question with free-text to explain their choices (Q4, Q5, Q6). Finally, in section 3, participants were given the opportunity to add further information as free-text (Q9).

IV. Results and Discussion

A. Participants

A total of 30 students (Stage 1: n=22, Stage 2: n=8) completed the survey over a three-week period in February 2023, representing 35% of the Stage 1 cohort, and 15% of the Stage 2 cohort.

B. Data Analysis

The results of this survey are addressed in two parts. First, an analysis of the students' perception of how the Pass/Fail model impacted upon their effort, stress and preparedness for later years is presented. Second, a sentiment analysis analyses the final open-text response to identify avenues for future work.

To prepare the data for analysis, values were assigned for the three potential responses for Questions 3, 7 and 8: the negative response was assigned a value of 1, the neutral response a value of 2, and the positive response a value of 3. This allowed for some preliminary statistical analysis sliced by the Stage of the responding students. Furthermore, free text responses from Questions 4 - 6 were used to provide further commentary and justification where necessary.

a) Effort

The results for Question 3 (*In terms of effort, when working on an assessment with a Pass/Fail Grade do you… (1) put in minimal effort to pass; (2) put in a good effort but not necessarily my best; (3) put in your best effort*) indicate that both year groups indicate a neutral response ($M = 2.32$, $SD = 0.68$). Stage 1 ($M = 2.33$, $SD = 0.66$) and Stage 2 ($M = 2.29$, $SD = 0.76$) did not report any significant difference in the level of effort put into their assessments ($p > 0.05$), with both groups indicating a tendency towards putting in good effort, but not necessarily pushing themselves beyond the minimum requirements. Fig. 1 visualises this data.

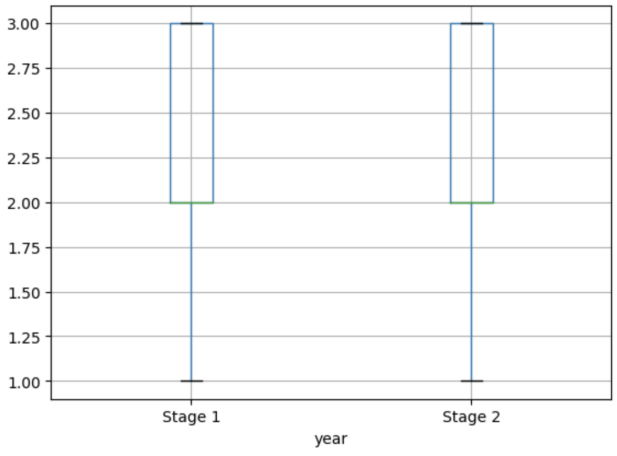

Fig. 1. Box-plot showing the results for Question 3 grouped by year

Some Stage 1 students commented on the fact that the Pass/Fail system provides leniency: "the grade system is a little bit more forgiving from past experiences" / "it could be

good for people starting out and to eliminate competitiveness" / "a Pass/Fail system for first year only is a good approach". For the students who indicated that they pushed beyond the minimum requirements, the qualitative text provides some further context. A common response was that students exceeded the minimum requirements to be certain that they would achieve the required grade: "[I wanted to] make sure that I definitely pass and not just scrape past each module" / "Making sure I actually pass" / "Just because it is pass/fail does not mean that the equivalent score of a D is required to pass". One student further expanded on this line of thought, by indicating that "the anxiety of failing" was their primary motivation to push past the goalposts of a basic Pass.

Students also reported that they were challenging themselves and exploring their learning journeys: "I believe the best way to learn is to challenge yourself and change your way of thinking" / "By pushing myself I can become more confident in what I'm taught" / "[I want to] make sure I have the same level of work ethic in the following years that don't use a pass/fail system".

In particular, one student responded that "there isn't an incentive if you pass and there's no sense of urgency since there isn't a scale to tell how badly it needs improving", and another one "wasn't aware of how good my work was in grade wise to take feedback [seriously] or not." A point was raised regarding how subgrades are combined, and how that can impact on student engagement "as so many people just passed the first [submission] and [didn't] try for the second."

b) Stress

The results for Question 7 (*When comparing Pass/Fail to letter grades (A-F), do you think that Pass/Fail… (1) adds to my stress levels; (2) makes no difference to my stress levels; (3) reduces my stress levels*) indicate that Stage 2 students (M = 2.57, SD = 0.77) noted a bigger reduction in stress when reflecting on their experience with the Pass/Fail model compared to Stage 1 (M = 2.14, SD = 0.85). Fig. 2 visualises this data.

It is interesting to note the variance across the Stage 1 students, indicating a higher degree of uncertainty in this population.

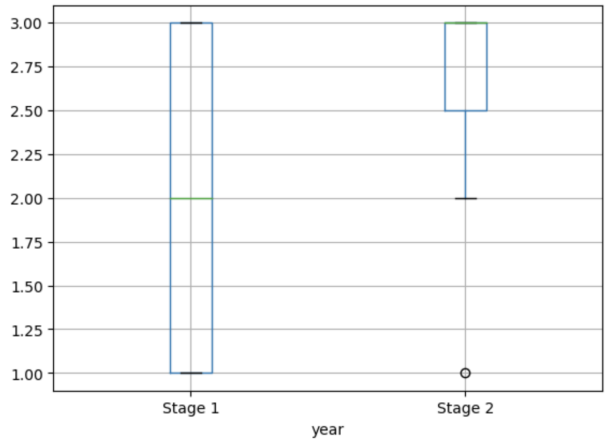

Fig. 2. Box-plot showing the results for Question 7 grouped by year

A Stage 1 student stated that "[..] the pass/fail system is harder to understand and not less stressful as you have actually more pressure." Conversely, other Stage 1 students said: "it's a good system [..] as there is less stress", "[the] pass/fail grading system has helped with transitioning to university level studies", and "the pass or [fail] grade is still the best [otherwise] you increase the pressure in students". These comments indicate that Pass/Fail is polarising across Stage 1 students, and this may unwittingly lead to this grading model inducing stress in some.

Stage 2 students felt more positively towards the Pass/Fail model when they reflected on their transition to letter grades. One response stated "the nebulousness of a Pass or Fail eased that anxiety somewhat because I wasn't so fixated on the semantics of a specific letter, and I just got to focus on doing my best." A similar sentiment was expressed by other Stage 2 students / "I prefer pass/fail as it has the same outcome but a lot less stress and I felt my mental health and attitude was better with pass/fail. [Letter grades] make me stress about passing whereas pass/fail makes me want to learn and makes uni more enjoyable".

The authors posit that this divide in experiences between the Stage 1 and Stage 2 students may be indicative of the fact that Stage 2 students are reflecting on their experience after having experienced both the Pass/Fail model in their previous year, and letter grading in their current year.

c) Preparedness for Letter Grades

The results for Question 8 (*When comparing Pass/Fail to letter grades (A-F), do you think that Pass/Fail… (1) feels too vague regarding future grade expectations; (2) undecided; (3) presents a good balance in terms of preparing you for future grade expectations*) indicate that Stage 1 students do not feel the Pass/Fail model helps them form realistic expectations of how they are likely to perform in subsequent years (M = 1.43, SD = 0.68). The response from Stage 2 students is more positive (M = 2.43, SD = 0.98), presumably due to the fact that they are able to look at their experience in hindsight. A Spearman's correlation coefficient analysis shows a significant difference between the setting of expectations between Stage 1 and Stage 2 (p = 0.01). Fig. 3 visualises this data.

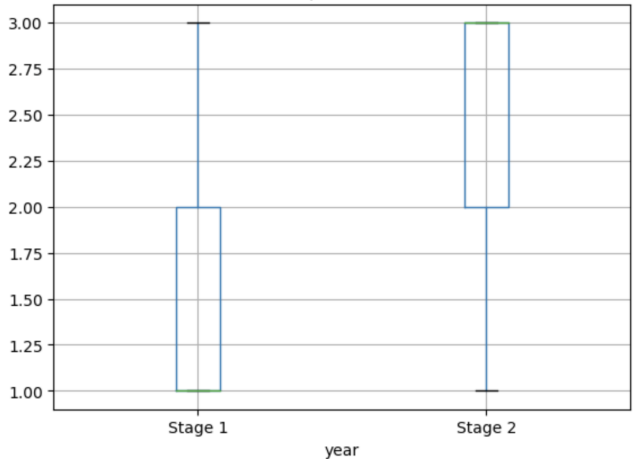

Fig. 3. Box-plot showing the results for Question 8 grouped by year

In the qualitative comments, Stage 1 students typically highlighted concern when thinking about the transition to letter grades in future years of study: "It will be a little tough to begin with" / "stating they "may feel more pressur[ed]" when trying to achieve "the best possible grades", although some students also highlighted the fact

they were “excited” and looked forward to “a better understanding of the work you put in” / “it will give me something better to strive towards”.

Stage 2 students considering the transition in hindsight commented on the switch between a focus on feedback to a focus on letter grades: “[letter grades are] good, but feedback for this year was not as strong as the first”, showing that they appreciated the higher focus on feedback possible due to the Pass/Fail model. One of the guiding principles the course team used when moving to a Pass/Fail model was to ensure that the grade awarded came with a more detailed level of feedback to the student, to help guide them and enhance their learning experience.

To this end, the qualitative questions were analysed to determine the students’ perception of feedback. In one student’s words: “While Pass/Fail comes across as more lenient because a more precise 'level' is not specified, it still gives accurate feedback by putting the focus on the written feedback and holistic quality of work.” Multiple students mentioned making use of the feedback for reflection and improvement: “to look back on work I had done, and have looked at what I can do in the future to help myself develop myself further to get better scores” / “I used the feedback to correct anything that I did wrong in the modules” / “self-reflection looking at the feedback and seeing where it applies to the coursework”. One student further explained that their feedback focussed on formatting and annotations, and that they therefore “put extra work into making my commentaries and annotations more thorough in the next creative submission”. However, some students felt that the feedback was too vague to help them in these areas. One student reported that “it is hard using the feedback as I don't know how well I passed or failed and therefore how much the feedback means”, and another lamented that it was not personalised enough - that the feedback “felt like it was just copied and pasted.”

C. Sentiment Analysis

Students were invited to end the survey by giving any other feedback of their experience of the Pass/Fail model. Twelve students completed this section. A pre-trained sentiment analysis system, available through the SpaCy library[1], was used to calculate the sentiment of each response. Mean sentiment was then calculated by summing the score for each response and dividing by the number of responses (to prevent long responses having more weight). The mean positive sentiment score was 75.6% (and the mean negative sentiment score was 24.4%), from which it was deduced that overall student opinion was positive.

These results were expanded upon by conducting an Aspect-Based Sentiment Analysis (ABSA) of the responses. ABSA is a sub-task of sentiment analysis, where the goal is to mine opinion regarding specific entities present in the text [42]. The term frequency was captured for all words in the responses and the most popular terms (i.e. any terms mentioned more than five times in the corpus) were extracted. The resulting list contained 13 words which were stemmed to their root-form (grade, pass, first, year [Stage 1], work, level, good, system, stress, Pass/Fail, require, feedback, really). Using a pre-trained ABSA model from the Huggingface library[2], a sentiment score was calculated for each of these words in regards to each response (i.e the sentiment of each student in regards to each aspect respectively). The outcome is demonstrated in the visualisation in Fig. 4 below.

Notably, ‘pass/fail’ and ‘system’ aspects co-occur in a number of responses, and generally have strong positive sentiment associated. Multiple students mention the aspects ‘first’ and ‘year’ (also referred to as Stage 1) within the same response with positive sentiment, indicating general positivity around the application of the Pass/Fail model in Stage 1. In addition, the entities ‘work’ and ‘level’ are mentioned frequently together with high degrees of positive sentiment, indicating students consider the work-levels associated with pass-fail to be manageable.

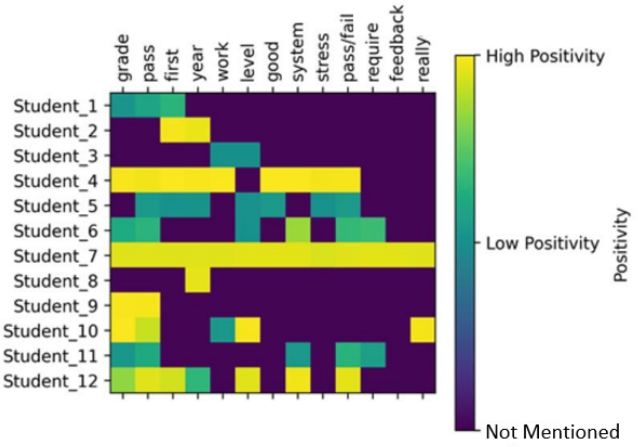

Fig. 4. Aspect-Based Sentiment Analysis (ABSA) visualised as a heat-map

It can also be seen that a handful of students mention the aspect ‘stress’ in a positive context (indicating the reduction of stress, as mentioned in the earlier analysis). Although the aspects ‘grade’ and ‘pass’ are most frequently mentioned in a positive context, this might be due to other factors, but it could be taken as a sign of confidence in the ability to pass, when faced with a binary model such as Pass/Fail. Finally, the aspect of ‘feedback’ is not as impactful as expected (i.e. it is mentioned by only a single student), but does appear in a positive context in that response.

It is worth highlighting that Student 7 used all of the most frequent words in their response. This response was the largest provided, at 477 words (significantly greater than the mean length of response for this question, which was 87 words). Furthermore, the response contained a reasonably comprehensive reflection of the student’s experiences of Pass-Fail grading (hence the coverage of most areas of the analysis, and indeed some aspects which were very specific to the student).

The above analysis was applied to all respondents, as well as respondents sliced by Stage 1 and 2. However, it was found that the split of respondents made it difficult to draw conclusions (Stage 1: n=8, and Stage 2: n=4). Furthermore, the overall distribution of sentiment was representative of the above analysis, and did not offer any significant differences. Therefore joint results are reported here, but it is acknowledged that further analysis could be conducted on this aspect in future.

[1] http://spacy.io

[2] https://huggingface.co/docs/hub/models-libraries

V. Conclusions and Future Work

From the analyses, the following conclusions were drawn. Students generally appreciate the application of the Pass/Fail model, although for most, the benefits become more apparent as they move beyond Stage 1, when they are exposed to alternative grading models. While Pass/Fail helped to clarify expectations surrounding assessment, students found it more difficult to contextualise their own performance.

A. Recommendations

In this section, we will provide recommendations for any instructors who wish to implement similar models within their curricula, reflect on the limitations that we experienced within this study, and close the paper with recommendations for future work.

Recommendation 1: We have seen a notable difference in the experience of Stage 1 and Stage 2 students, suggesting that whilst Pass/Fail is beneficial, the impact of these benefits does not become clear until students have experience of other grading models. In this case, it might be better to hold induction sessions with students to explain the rationale behind Pass/Fail, pointing to responses and experiences of previous students.

Recommendation 2: Efforts need to be made by academic teams to ensure that the amount of stress experienced by students due to the Pass/Fail marking scheme is mitigated against. Whilst this might be something that resolves itself as the Pass/Fail marking scheme embeds itself more deeply in the culture of the School, it is something that needs to be carefully monitored.

Recommendation 3: Adequate care should be taken to personalise feedback to each individual submission, and not make it seem generated, or "copied and pasted". A benefit of the Pass/Fail model is that students will look beyond the grade - the feedback that accompanies the grade therefore needs to be contextual, and useful.

B. Limitations

The survey was completed by a self-selecting subset of students in Stage 1 and Stage 2, so the responses may not be representative of the larger population. Furthermore, as demographic data was not collected, it was not possible to make inferences based on the diversity of the study sample. Future studies will aim to mitigate against both these items.

B. Future Work

Whilst this study presents an important snapshot of findings and recommendations based on this pedagogical change, it is important to grow the dataset, to ensure that any conclusions are representative of the larger population. To that end, another study is planned for the start of the upcoming semester, where data might be captured at various points in the Stage 1 timeline. This would allow us to pinpoint whether the perception of the Pass/Fail model changes - and how - at various milestones throughout the semester. Expanding the scope of the sentiment analysis would provide a more comprehensive dataset, enabling us to draw more meaningful conclusions. Furthermore, students might be invited to focus groups, to get a better understanding of their perceptions of this grading model, beyond a simple survey.

We have shown that many students report working to the minimum requirement, finding motivation to be particularly tricky when faced with little guidance. We may consider having a third set of "could have" requirements which are optional and would have no bearing on the grade. Further study might consider the impact of these requirements, particularly on student motivation and their desire to excel.

A companion study where the module coordinators are surveyed is also planned. Whilst the student perception is important to get right, and vital when considering the larger student experience, it is also important to archive lessons learnt by academics when planning and grading assessments using the Pass/Fail model.

Acknowledgments

The authors would like to thank the participants for giving their time and effort, as well as the original Pass/Fail curriculum development team for giving their time at the start of the project to shape the direction the study would take. Whilst there are too many people to name, the authors would like to thank Dr Pam Johnston for her invaluable commentary and advice.

References

[1] Biggs, J., 1996. Enhancing teaching through constructive alignment. Higher education, 32(3), pp.347-364

[2] Elton, L. and Johnston, B., 2002. Assessment in universities: A critical review of research.

[3] Broadfoot, P. and Black, P., 2004. Redefining assessment? The first ten years of assessment in education. Assessment in Education: Principles, Policy & Practice, 11(1), pp.7-26.

[4] Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T. and Welsh, M.E., 2016. A century of grading research: Meaning and value in the most common educational measure. Review of Educational Research, 86(4), pp.803-848.

[5] Anderson, L.W., 2018. A Critique of Grading: Policies, Practices, and Technical Matters. education policy analysis archives, 26(49).

[6] Iamarino, D.L., 2014. The benefits of standards-based grading: A critical evaluation of modern grading practices. Current Issues in Education, 17(2).

[7] York, T.T., Gibson, C. and Rankin, S., 2015. Defining and measuring academic success. Practical assessment, research, and evaluation, 20(1), p.5.

[8] Michaels, J.W., 1977. Classroom reward structures and academic performance. Review of Educational Research, 47(1), pp.87-98.

[9] Boud, D., 2017. Standards-based assessment for an era of increasing transparency. Scaling up assessment for learning in higher education, pp.19-31.

[10] Flores, M.A., Veiga Simão, A.M., Barros, A. and Pereira, D., 2015. Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. Studies in Higher Education, 40(9), pp.1523-1534.

[11] Tai, J., Ajjawi, R., Bearman, M., Boud, D., Dawson, P. and Jorre de St Jorre, T., 2023. Assessment for inclusion: rethinking contemporary strategies in assessment design. Higher Education Research & Development, 42(2), pp.483-497.

[12] Marginson, S. 2016. The worldwide trend to high participation higher education: dynamics of social stratification in inclusive systems. High Educ 72, 413–434. https://doi.org/10.1007/s10734-016-0016-x

[13] Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Basterra, E. Trumbull, E., & G. Solano-Flores (Eds.), Cultural validity in assessment: Addressing linguistic and cultural diversity (pp. 3–21). New York, NY: Routledge.

[14] Winstone, N.E. and Boud, D., 2022. The need to disentangle assessment and feedback in higher education. Studies in higher education, 47(3), pp.656-667.

[15] Li, J. and De Luca, R., 2014. Review of assessment feedback. Studies in higher education, 39(2), pp.378-393.

[16] Wiliam, D., 2011. What is assessment for learning?. Studies in educational evaluation, 37(1), pp.3-14.

[17] Astin, A.W., 2012. Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education. Rowman & Littlefield Publishers.

[18] Linn, R.L., 2000. Assessments and accountability. Educational researcher, 29(2), pp.4-16.

[19] McDonnell, L.M., 2005. Assessment and accountability from the policymaker's perspective. Teachers College Record, 107(14), pp.35-54.

[20] Biesta, G., 2015. What is education for? On good education, teacher judgement, and educational professionalism. European Journal of education, 50(1), pp.75-87.

[21] Pattison, E., Grodsky, E. and Muller, C., 2013. Is the sky falling? Grade inflation and the signaling power of grades. Educational Researcher, 42(5), pp.259-265.

[22] Atkinson, R.C. and Geiser, S., 2009. Reflections on a century of college admissions tests. Educational Researcher, 38(9), pp.665-676.

[23] Thorsen, C. and Cliffordson, C., 2012. Teachers' grade assignment and the predictive validity of criterion-referenced grades. Educational Research and Evaluation, 18(2), pp.153-172.

[24] Sawyer, R., 2013. Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. Applied measurement in education, 26(2), pp.89-112.

[25] De Castella, K., Byrne, D. and Covington, M., 2013. Unmotivated or motivated to fail? A cross-cultural study of achievement motivation, fear of failure, and student disengagement. Journal of educational psychology, 105(3), p.861.

[26] Glaser, R., 1963. Instructional technology and the measurement of learning outcomes: Some questions. American psychologist, 18(8), p.519.

[27] Lok, B., McNaught, C. and Young, K., 2016. Criterion-referenced and norm-referenced assessments: compatibility and complementarity. Assessment & Evaluation in Higher Education, 41(3), pp.450-465.

[28] Baron, H., 1996. Strengths and limitations of ipsative measurement. Journal of Occupational and Organizational Psychology, 69(1), pp.49-56.

[29] Hughes, G., 2011. Towards a personal best: A case for introducing ipsative assessment in higher education. Studies in Higher Education, 36(3), pp.353-367

[30] Seery, N., Delahunty, T., Canty, D. and Buckley, J., 2017. Illustrating educational development through ipsative performance in design based education. PATT2017: Technology & Engineering Education-Fostering the Creativity of Youth around the Globe.

[31] Hughes, G., 2014. Ipsative assessment: Motivation through marking progress. Springer.

[32] Hay, P.J. and Macdonald, D., 2008. (Mis) appropriations of criteria and standards‐referenced assessment in a performance‐based subject. Assessment in Education: Principles, Policy & Practice, 15(2), pp.153-168.

[33] Echauz, J.R. and Vachtsevanos, G.J., 1995. Fuzzy grading system. IEEE Transactions on Education, 38(2), pp.158-165.

[34] Wolming, S. and Wikström, C., 2010. The concept of validity in theory and practice. Assessment in Education: Principles, Policy & Practice, 17(2), pp.117-132.

[35] Durm, M.W., 1993, September. An A is not an A is not an A: A history of grading. In The educational forum (Vol. 57, No. 3, pp. 294-297). Taylor & Francis Group.

[36] “Guide to the code of assessment 2022-23”, The University of Glasgow, https://www.gla.ac.uk/media/Media_275332_smxx.pdf (accessed May 12, 2023).

[37] “The UoA Common Grading Scale (CGS)”, University of Aberdeen, https://www.abdn.ac.uk/students/academic-life/common-grading-scale.php (accessed May 12, 2023).

[38] Sadler, D.R., 2008. Transforming holistic assessment and grading into a vehicle for complex learning. In Assessment, learning and judgement in higher education (pp. 1-19). Dordrecht: Springer Netherlands.

[39] Weiss, R.S., 2017. Issues in holistic research. In Institutions and the Person (pp. 342-350). Routledge.

[40] Wats, M. and Wats, R.K., 2009. Developing soft skills in students. International Journal of Learning, 15(12).

[41] De Grez, L., Valcke, M. and Roozen, I., 2012. How effective are self-and peer assessment of oral presentation skills compared with teachers’ assessments?. Active Learning in Higher Education, 13(2), pp.129-142.

[42] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," IEEE Trans. Affective Computing, vol. 13, no. 2, pp. 845-863, 2022, doi: 10.1109/TAFFC.2020.2970399.

Appendix 1: Pass/Fail Survey

1. What is your current Year of study?
    - First Year
    - Second Year
2. What degree route are you on?
    - Computer Science
    - Computing and Creative Design
    - Cyber Security
3. In terms of effort, when working on an assessment with a Pass/Fail Grade do you:
    - Put in your best effort
    - Put in good effort but not necessarily your best
    - Put in minimal effort to achieve a pass
4. If you put in more effort than required, what is your motivation for going beyond the requirements? (free text response)
5. When you received a Pass/Fail grade, you also got feedback about this grade. How did you use the feedback provided? (free-text response)
6. From second year onwards, you are graded using letter grades (A-F). How do you feel about transitioning to letter grades instead of Pass/Fail? (free text response)
7. When comparing Pass/Fail to letter grades (A-F), do you think that Pass/Fail:
    - Adds to my stress levels
    - Reduces my stress levels
    - Makes no difference to my stress levels
8. When comparing Pass/Fail to letter grades (A-F), do you think that Pass/Fail:
    - Presents a good balance in terms of preparing you for future grade expectations
    - Feels too vague regarding future grade expectations
    - Undecided
9. Is there any other feedback you would like to give us about the Pass/Fail grading system used in first year? (free text response)