Examining large student cohorts - a question of questions

Roger P. West[1] and Michael A. Wride[2]
[1]Department of Civil, Structural and Environmental Engineering, Trinity College, Dublin, Ireland
[2]Centre for Academic Practice and Learning, Trinity College, Dublin, Ireland

Emails: rwest@tcd.ie, wridem@tcd.ie

ABSTRACT: With pressure on academic courses worldwide to increase student numbers, the trends in exam marks distribution for subject modules become more meaningful, with more distinct pattern characteristics reflecting student choice, topic and exam question difficulty and lecturer marking severity, refinement and consistency. The practice of representing the overall exam results for a module through histograms enables Normality, skewness and randomness to be identified, interrogated and understood better. However, when dealing with large numbers of exam candidates (of the order of 1000 or more), an investigation of the averages and histograms for individual exam questions can further reveal refined explanations for unusual student performance. This paper investigates the outcomes of 1st and 2nd year examinations for modules on an engineering degree course in another jurisdiction, with class sizes of circa 2400 and 1700 students, respectively, in order to develop a deeper understanding of exam dynamics amongst students and academics setting and marking those papers. It can involve many tens of thousands of items of data in a histogram for just one module, in which trends are not random and have potential causes, intended or accidental. It emerges that at least four different question mark patterns may exist from a range of modules types investigated. These indicate the importance of examiners having a better appreciation, when delivering lectures, planning exam papers, question structure and marking scripts, of the different factors which give rise to unusual examination trend outcomes.

KEY WORDS: Histograms; Marks distributions; Modules.

1 INTRODUCTION

Reliability in assessment can be considered to be the extent to which test measurements reflect the properties of those individuals being measured [1]. Berkowitz et al. [2] define reliability in this context as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker". Reliability can also be considered to be how consistent or error-free the test measurements are [3]. When random error is minimal, scores will be accurate, reproducible and generalisable to other test occasions and similar tests.

There have been long-standing concerns about the reliability of marking and errors in engineering education. For example, McVey [4] showed that the range of marks awarded to a script was often disturbingly large. Furthermore, it is very important in STEM disciplines to provide well-written problems and exam questions [5]. Such problems should enable students to show how much they understand the process of problem-solving, through being given credit primarily for showing the process rather than finding the correct answer.

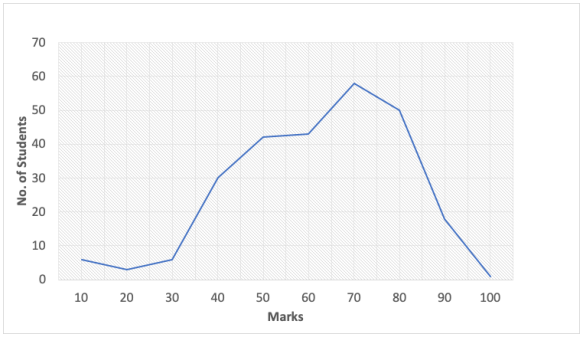
In order to improve marking consistency between multiple markers in engineering examinations at third level, marker training exercises can be implemented to improve marking reliability and consistency [6]. Using these approaches, there was a significant reduction in the spread of the marker means, indicating an improvement in consistency. This fits into the wider current focus on assessment literacy, which is designed to enhance the capability of staff and students to make sense of assessments [7].

This papers considers a selection of module exam outcomes (from some 35 modules) of first and second year engineering students at just one sitting where class sizes are very large indeed, with the advantage of being able to infer extra significance to the trends due to the enormity of the available data. The university from which the data emanates is anonymised but is largely irrelevant as many university courses will have experienced such trends but rarely with the benefit of access to such rich data, which gives strong credence to the observations made.

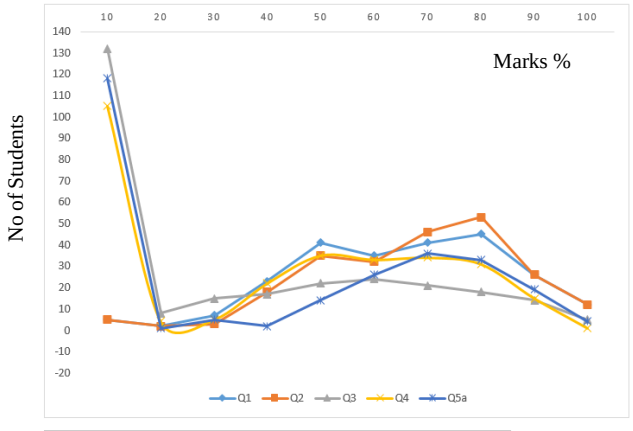
2 MARKS ANALYSIS

2.1 Sample Marks Histogram in Fresher years in Engineering at Trinity College Dublin

There has been an expansion in recent years in the numbers of entrants into the first year of the integrated 5-year engineering science degree at Trinity College, which now stands at a cohort of about 240 students. Evidence of high or low standard deviations in exam marks distributions are not uncommon, and occasionally more unusual trends such as extreme skewness or bi-modality may be observed and must be acted upon. Seldom are distributions in individual exam questions investigated and would be of little value given the small class sizes in sophister years. Considering only the first two common years, when class sizes are relatively large, conventional summative examinations reveal typical first year module exam results as shown in Figure 1(a), with a pass mark of 40% and an average exam mark of 64%. In this particular exam, there were 6 questions to be answered in three sections, with one question to be answered from each section, where each section is delivered and marked by a different academic. It follows that about one third of the class did not attempt one question in each section (with the exception of question 2, which had been perceived as easier) and, subsequently, it may be observed that the distribution of marks per question (Figure 1(b)) for those who did attempt those questions, is broadly Normally distributed, with a slight skew. The data is also not perfectly bell shaped due to the relatively low number of students and the



(a)



(b)

Figure 1. (a) Examination marks distribution for a typical module for 240 first year engineering students at TCD and (b) typical marks distribution by question.

granularity of marking (to at best 1 mark out of 20 in most cases). This graph broadly indicates that, with the exception perhaps of question (Q) 3 (which has a flatter curve), each question has a reasonable spread of marks and the questions and marking are suitably gradated so that only the best students do very well but most have a good chance of doing reasonably well. This is the norm and what one has come to expect of typical exam mark distributions in classes of mixed ability.

In some jurisdictions outside Ireland, very large class sizes are not unusual. In examining these students' performance, one may expect that, again assuming mixed ability but very capable students in a class, the exam marks distribution in modules and in questions within modules will be broadly similarly Normally distributed but with more refined and meaningful trends due to the high numbers involved

In the university under study, the overall pass mark in first year is 35%, which comprises three component marks: continuous assessments (called sessionals), a formal mid semester test/exam (MST) and an end-of-semester test/exam (EST), combined into a gross test result of 100% (GT). It is not uncommon for the EST to be weighted between 30 and 40% of the GT, so there is a heavier reliance on summative continuous assessment than would be the norm in Ireland. But there is an additional pass/fail criterion where every candidate must obtain a minimum of 25% in the MST/EST exam components combined.

In first year engineering in 2020, there were over 2400 students (ten times that in Trinity) with somewhat less (about 1700) in second year engineering due to intake expansion this year. Given the number of modules involved in the two years combined (some 35), this amounts to over 26,000 (14,270 and 11,940 in first and second years respectively) end-of-semester exam scripts being examined in one semester. An analysis of the wealth of marks data available on examining these students can give new insights into many facets of an examining system: the teaching quality (where often 10 academics teach one module, all covering the entire syllabus in parallel lectures), the degree of difficulty of exam questions and the length of model answers, the granularity of marking and potential unintentional marking bias. The universal module exam rules specify that all students must answer all questions (which can be from 4 to 10 in number, depending on the module) and usually each lecturer marks just the entirety of one question on each paper, which helps to minimise marking bias, which could easily arise due to a single lecturer marking, perhaps, as many as 2440 scripts on the one exam question.

3 CASE STUDIES

Investigating several examples of modules in first and second year, which have been anonymised, will show some interesting and insightful trends in exam marks.

3.1 Module 1: 2nd Year Engineering Module

In comparison with Figure 1, a typical exam results histogram in the university under study can be seen in Figure 2 in which the average MST, EST and GT marks are 50%, 64% and 56% respectively, noting that 1223 students took this module. Here the Normality of the curves and the better performance in the EST than MST may be observed. Either the EST exam was easier or when the MST results were published after mid-term, students ascertained they had to perform better, as the EST results show they did, where only 79 students failed the GT overall – a distinct advantage of having MST exams. The degree of Normality of marks in this case is typical of the results distributions in the 35 modules considered (as is expected) with just a few exceptions as described presently.

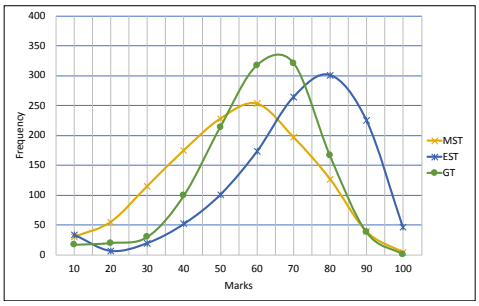


Figure 2. Typical module MST/EST and GT marks for Module 1 with 1223 students

3.2 Module 2: 1st Year Sciences Module

In contrast, an exception is the trend in exam marks distribution for a sciences module in first year (Figure 3(a)) which suggests that the opposite is occurring. With MST/EST/GT averages of 59, 43 and 67% respectively (with 69 overall failures), one may speculate that the exams were easier in the MST and some students work less hard before the EST as they were more confident of passing overall.

An examination of the marks breakdown in individual questions for the entire cohort (of 1300) was possible due to the

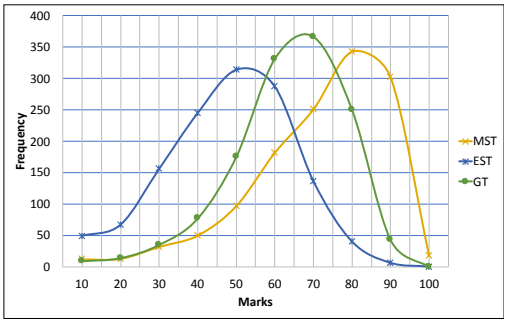
diligence of the faculty in compiling these statistics (Figure 3(b) and Table 1) and this reveals, perhaps, an alternative reason for the poor performance in the EST.

Table 1. Marks statistics by question for a science module in 1st year with 1300 scripts.

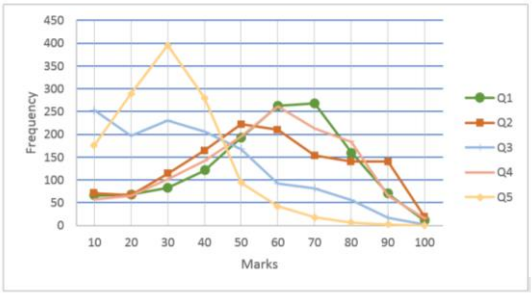
	Q1	Q2	Q3	Q4	Q5
Ex marks	20	20	20	20	20
Average	10.5	10.6	6.5	10.5	5.3
St Dev	4.2	4.6	4.4	4.2	2.8
% mark	53%	53%	32%	53%	27%

It may be observed that two questions, Q3 (32% average) and Q5 (27%), have very low averages which has special statistical significance because of the large cohort of students taking this exam – there must be a systematic reason for the averages being so low. The histogram of individual questions in Figure 3(b) reveals that Q1, 2 and 4 have regular Normal distributions with slight skews. It should be noted that about 6,500 individual marks were compiled to derive this figure. The histogram for Q5, however, has considerably higher kurtosis, showing clearly a low standard deviation and few obtained good marks in this question. The histogram for Q3 indicates a quite different pattern of marks, where, again, few candidates obtained more than 10 out of 20 for the question, but here a downward sloping distribution exists. This could represent either an unusually difficult question or a more severe marking regime, given that all 1300 scripts in that question were marked by one person.

The two low question average marks suggest that, in fact, the candidates in this exam, who had to attempt all questions, were not being assessed fully on 5 questions (that is, 100% is not attainable) and thus the average EST mark for candidates is, not surprisingly, low at 43%.



(a)



(b)

Figure 3. (a) Typical module MST/EST and GT marks for Module 2 with 1300 students (b) Histogram of distribution of marks for individual questions.

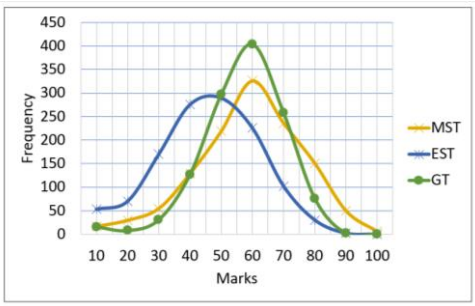
3.3 Module 3: 2nd Year Engineering Module

The histogram of overall marks for the 1219 candidates who took the third module under consideration (a non-numerate second year engineering module) in Figure 4(a) has MST/EST and GT average marks of 54, 41 and 52% respectively (with 77 GT failures) and attracted the attention of the authors because, despite a poor MST performance, the EST performance is worse and the marks spread is much higher than the MST (Table 2 and Figure 4(b)).

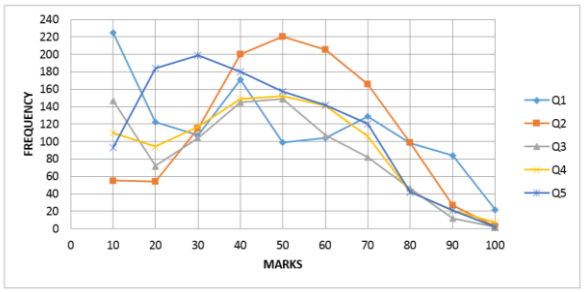
Table 2. Marks statistics by question for an engineering module in 2nd year with 1219 scripts.

	Q1	Q2	Q3	Q4	Q5
Ex marks	18	24	20	20	18
Av Mark	7.2	11.1	7.7	8.2	6.8
Percentage	41%	47%	39%	41%	38%

While the marks show consistency in the question averages for a large cohort of students, albeit with low percentages (a maximum of 47% in Q2), the histogram of the marks per question in Figure 4(b) shows that Q5 has a strong skew in the distribution and Q1 has a downward sloping distribution similar to that seen in Module 2 above. This unusual distribution goes a long way to explaining the large number of students (circa 230) who did poorly in Q1 in particular.



(a)



(b)

Figure 4. (a) Typical module MST/EST and GT marks for Module 3 with 1219 students (b) Histogram of marks on individual questions.

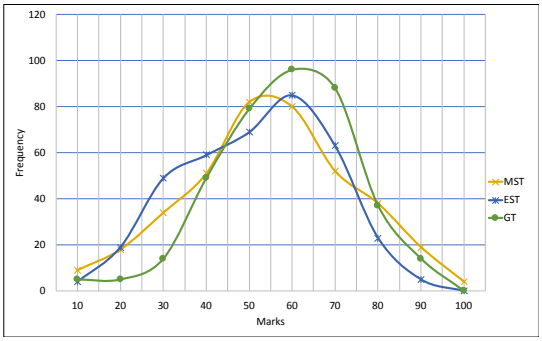
3.4 Module 4: 2nd Year Engineering Module

Another engineering module in 2nd year was of interest because the MST and EST spreads were more or less identical with slight evidence of bi-modality in the EST. With average MST/ES/GT marks of 50, 47 and 53% respectively (Figure 5(a), with 32 fails in 387 candidates), it was instructive to view the marks per question distribution (as shown in Table 3 and Figure 5(b)): Questions 6 and 7 appear to be poorly answered, while the other averages are typically randomly varying, as one

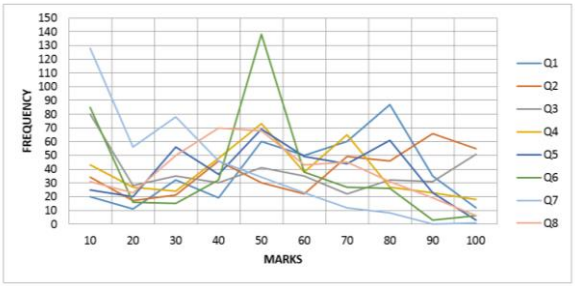
might have expected. However, the histogram of the distribution of marks for each of the 8 questions (Figure 5(b)) shows that Q3, 6 and 7 had high numbers of students who could not start the question (with scores of less than 10% in those questions), which partly explains the low averages in these questions. Q7 broadly sloped downwards in contrast to Q2 which broadly sloped upwards, explaining the higher average in this question. It is also interesting to note that the distributions in almost every question is more random and less "Normal" than in, for example, the majority of questions in Figure 2(b), which may possibly be attributed to a less refined marks allocation. This more random distribution of patterns is the third type of pattern which will be discussed presently.

Table 3. Marks statistics by question for engineering module in 2nd year with 387 scripts.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Ex	12	13	12	13	12	12	12	14
Avge	6.8	7.6	5.5	6.1	5.8	4.6	2.8	6.2
%	56%	58%	46%	47%	49%	38%	23%	44%



(a)



(b)

Figure 5. (a) Typical module MST/EST and GT marks for Module 4 with 387 students (b) Histogram of marks for individual questions.

3.5 Module 5: 2nd Year Engineering Module

This highly numerate module was separated into two distinct parts, where the year was split in two, each half taking one or other part in the two semesters, with some overlap between the two sets of four questions in each part. The histogram of the MST and EST marks (Figure 6(a)) shows a strong skew (MST/EST/GT averages are 33, 33 and 43%) and the results show there are 155 candidates of 453 (35%) who failed by obtaining less than 25% in the combined MST+EST score and 109 (24% in the GT) failed overall. To attempt to understand why the overall exam performance is so poor, a review of the EST question marks distribution in some 232 of the 453 scripts shows the individual question breakdown as shown in Table 4

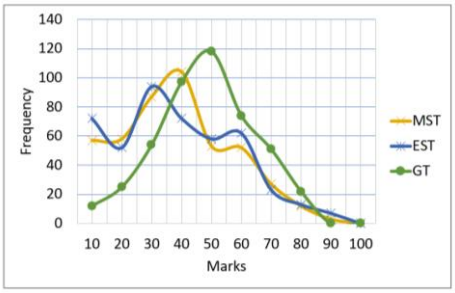
Table 4. Marks statistics by question for an engineering science module in 2nd year from sample of 232 scripts.

	Part A				Part B			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Ex	30	25	25	20	20	30	30	20
Avg	5.5	9.2	10.4	2.7	9.9	11.7	10.5	5.2
%	18	37	42	13	49	39	35	26

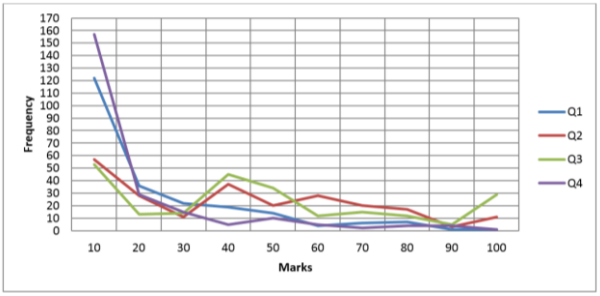
(and Figures 6(b) and (c)), noting that Part A and Part B have exam average percentages of 31% (that is, below the overall pass mark) and 37% respectively.

In Part A, the average percentage mark for questions 1 and 4 (18 and 13% respectively) are abnormally low while in Part B, question 4 has a very low average mark (at 26%) – and explain to a large degree the unusually high failure rates.

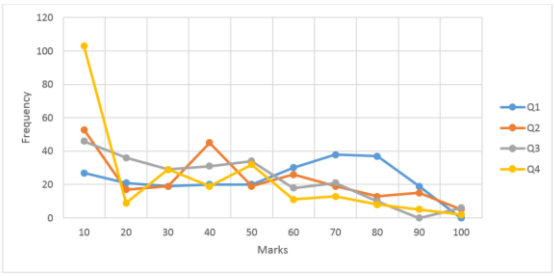
In addition, plots of the histograms of the entire cohort in Part A and Part B are shown in Figures 6(b) and (c) respectively. In some questions (Q1 and 4 in Part A and Q4 in Part B), many students did very poorly indeed (less than 10%), while in almost all questions of the eight there is no evidence of a Normal distribution in the marks awarded. In fact, there is some evidence of a downward or flat trend in numbers of students obtaining marks from 20% to 100% in most questions, indicating the degree of difficulty of these exam questions. These trends will be discussed later in the paper.



(a)



(b)



(c)

Figure 6. (a) Typical module MST/EST and GT marks for Module 5 with 453 students (b) Histogram of marks for individual questions on Part A (c) and Part B.

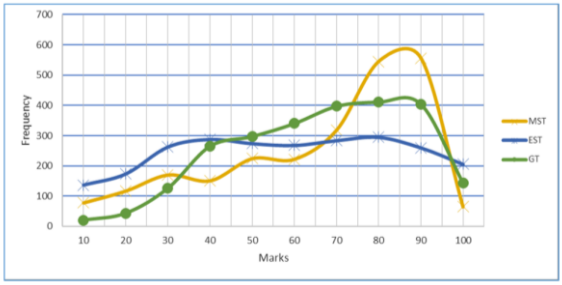
3.6 Module 6: 1st Year Engineering Module

This 1st year engineering module had more than 10% failures (242 in a cohort of 2446) despite an EST average of 53% and GT average of 61%. Both the MST and EST marks distribution (Figure 7(a)) are unusual in that they are not bell-shaped and have high standard deviations. In the EST marks, a complete set of individual question marks (over nine compulsory questions, about 22,000 marks in total) were evaluated for question averages (Table 5). From this, it is evident that none of the questions was very poorly scored, but given the very high number of candidates, it might be argued that this topic was not being marked out of 100% as a consequence of having to answer all 9 questions (where questions 5, 6 and 9 had averages less than 50%) and, with 2446 candidates, 597 (24%) failed to obtain at least 33% in the EST.

Table 5. Marks statistics by question for a engineering module in 1st year from 2446 scripts.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Ex	10	10	10	10	20	10	10	10	10
Avg	6.2	5.1	6.4	6.4	8.9	4.5	6.2	5.3	4.0
%	62	51	64	64	45	45	62	53	40

Furthermore, it was possible to inspect the graphs of student attainment on a question-by-question basis, as shown in Figure 7(b). In this case, one can observe that a clear cohort of students are doing very well in individual questions (> 9/10), but also many are also doing very poorly (< 1/10). What is particularly striking here is that a flat curve exists in between these extremes (that is, students have just as much chance of scoring 2 as 3 as 4 etc. marks out of 10), and that this marking trend exists for all 9 questions with only one minor exception (Q8). Possible reasons for this, the fourth and last trend type, will be discussed presently.



(a)

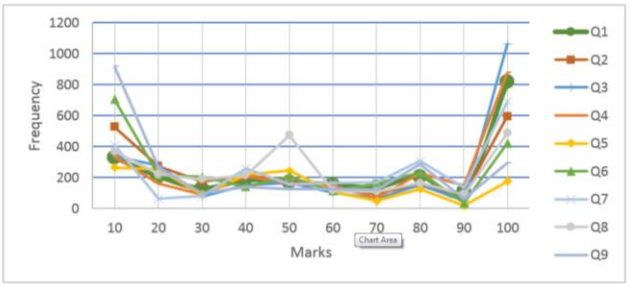


Figure 7. (a) Typical module MST/EST and GT marks for Module 6 with 2446 students (b) Histogram of distribution of marks for individual questions.

4 DISCUSSION

There is a number of potential issues highlighted by the shapes of the graphs presented here which, from time to time, may be universal in their application: the degree of difficulty and the length of model answers of exam questions could be too disparate; later parts of questions may rely on knowing the correct answer to earlier more difficult parts; overly 'granular' marking rubrics can result in cumulative errors; there could be potential unintentional marking bias between different academics marking the different questions. This latter should provide internal consistency within questions, but not necessarily between questions. Also, in some questions, there could be an over-emphasis placed on giving credit for the final answer (the product) rather than also giving due credit to the process, even if the final answer is incorrect [5]. There is also the fact that students are expected to answer all the questions on the exam paper, so there may be some issues around having sufficient time to answer the later questions (assuming they attempt to answer the questions in order) or abandoning some questions with no time to go back to complete them - and there may be some evidence for these effects in Tables 2-5 and in Figure 3 for Q5.

The graphs discussed in this paper fall into four general categories as follows: Normally distributed, left to right fall, peaky/granular and flat. Figures 1(b) and 2 represent a Normal distribution of marks which existed in the large majority of exam modules studied here. The bell curve indicates that there is a distinct beginning, middle and end to the questions, which enables reliable separation out of students of different abilities with respect to each other. Also, in TCD there is a choice of questions while in the other university there is not; with choice, the students have an opportunity to focus on questions in topics they feel more confident in, whilst excluding those on topics they are less comfortable/familiar with. This means that there is also an added layer of student self-assessment, that is, the students carry out metacognition on their own thinking and abilities during the examination itself in order to select their preferred questions to answer [8].

In some profiles, the mean score is skewed left to right, such as for Q3 in Figure 3. It seems that this question is simply too hard for the students to complete, with perhaps some parts of it being disproportionately hard compared to other parts (or a hard part in the middle section on which the latter part relies), since progressively smaller numbers of students achieve the higher marks. Given that all questions must be answered, a tactic many students might employ is when they get part way through a question and come across a difficulty, they move on to another question, but can't complete the difficult one in time before the exam ends. Thereby, the histogram for that question has a peak to the left, such as for Q1 and Q5 in Figure 4(b), Q7 in Figure 5(b) and almost all questions in Figure 6b. It could also be that the marking is too harsh, although one would feel that if this were the case for each part of the question, the marks would be uniformly low with a flatter curve.

This falling off profile also exhibits some aspects of the peaky/granular profile as exemplified by Figure 5b. It is characterized by little evidence of any Normal behaviour, which mitigates against any meaningful discernment of differences in student ability. Some questions exhibit a discrete high frequency to the left, indicating that students are unable to

get started on those questions. The stochastic nature of the profile indicates an almost random allocation of marks which could be explained in several ways. The questions (with some variability, where some students do better on some parts of the questions than others) could be uniformly hard. It could also be explained by the fact that the exam consists of mixed abilities of students in this topic, with an interest in diversifying into different disciplines in later year, and/or the criteria/rubrics for marking could be too granular (such as marking at discrete intervals 2 or 3 marks only). This brings into question the shared understanding of the nature of assessment by the markers in this module (assessment literacy – see [7]) and the way questions are structured to discern differences in student ability more seen, in absence, by its flat profile.

The problem of an absence of ability of a question to discern levels of achievement is exemplified by the flat profile seen in Figure 6(b) and 6(c). The flat profile is characterized by smoother and lower peaks, which fail to discriminate levels of achievement. While there are a high number of students who obtain very low marks, that is, they are unable to get started (see Questions 4 in both 6(a) and 6(c)), overall, there seems to be a random distribution of marks across the overall student cohort; for example, there is little difference in the frequency of student achievement for 20% or 80%. It is possible that the problem here is that students guess the answers to different parts of the questions, with varying and perhaps random success.

There is a more severe case of a flat profile in Figure 7b for individual questions in the 1st year engineering module. The module has a very unusual profile, since the flat trend in the centre is maintained for each question, but superimposed on this are two cohorts of students, one of which does extremely poorly, while the other does extremely well. The explanation for the extremities of this profile may be the presence of candidates who have (or have not) a particular aptitude in this subject. In this scenario, it would appear that the questions need to be adjusted to at least allow the vast majority of students to get started in these questions and so marking could reward more the process, context and application of the concept rather than, perhaps, being based solely on the final result. It seems like this module may benefit from consideration of Biggs' principle of constructive alignment [9], that is, aligning learning outcomes with appropriate assessments.

CONCLUSIONS

It should be recognised that this paper has limitations in that only exam results from two universities were examined. However, the trends discussed may well reflect those in exam results which could occur from time to time, often unnoticed, anywhere in the world. What is unique about the paper is the access which was allowed to very large quantities of data for the exam performance of student engineers, the compiling and analysis of which was only possible through the diligence and hard work of the academics and administrators involved.

This paper reflects on a series of unusual marks profiles in first and second year examinations in a jurisdiction where class sizes are extremely large, whereby trends in individual questions' marks can adopt particular significance. At least four different exam question marks profiles were identified and possible explanations were proposed as to why these marks distributions might arise, based on student attitudes or actions by academics in relation to question setting or marking. With this developed understanding of the causes of such profiles, suggestions are put forward as to how assessment literacy might assist many academics worldwide to avoid trends which lead to undesirable and unusual distributions of examination marks.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge with thanks the enormous efforts of the many academics and administrators in the main university under study who were involved in compiling the exams data for the modules discussed in this paper. The consent of the anonymous university to publish this data has been obtained, with sincere thanks.

REFERENCES

[1] Rudner, L. M., and Schafer, W.D., 200, *Reliability*, ERIC Digest.

[2] Berkowitz, D., Wolkowitz, B., Fitch, R. and Kopriva, R., 2000, *The use of tests as part of high-stakes decision-making for students: a resource guide for educators and policy makers*, Washington, DC: US Department for Education.

[3] Ebel, R. L. and Frisbie, D.A., 1991, *Essentials of Educational Measurement* (5th ed.) New Jersey: Prentice Hall

[4] McVey, P.J. 1975, *The Errors in Marking Examination Scripts in Electronic Engineering* 12(3): 203-216 https://journals.sagepub.com/doi/abs/10.1177/002072097501200305.

[5] Banta, T.W., Jones, E.A. and Black, K.E., 2009, *Designing Effective Assessment: Principles and Profiles of Good Practice*, John Wiley & Sons, San Francisco.

[6] Buskes, G, and Chan, H. Y., 2018, Implementation of marker training exercises to improve marking reliability and consistency [online]. In: *29th Australasian Association for Engineering Education Conference*. Hamilton, New Zealand, 92-98.

[7] Price, M., Rust, C., O'Donovan, B., Handley, K., and Bryant, R., 2012, *Assessment literacy: The foundation for improving student learning* Oxford Centre for Staff and Learning Development, Oxford Brookes University.

[8] Pintrich, P. R. (2002) The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing, Theory Into Practice, 41:4, 219-225, DOI: 10.1207/s15430421tip4104_3

[9] Biggs, J. (1999) Teaching for Quality Learning at University. Buckingham, UK: SRHE and Open University Press.