DESIGNING ASSESSMENT TASKS TO PREVENT CHEATING IN A LARGE FIRST-YEAR STATISTICS UNIT

Ayse Aysin Bilgin and Huan Lin
Macquarie University, Department of Mathematics and Statistics, Australia
ayse.bilgin@mq.edu.au

The year of 2020 has witnessed a drastic change in education sector due to the COVID-19 pandemic. There has been a surge of online, non-invigilated assessments which required rethinking to ensure academic integrity due to e-cheating. We redesigned and implemented learning materials/activities constructively to transform student learning from surface to deep learning, even though teacher-student and student-student interactions were reduced. Assessments were redesigned at higher levels of Bloom's taxonomy (e.g. evaluate) to provide opportunities for students to express their understanding and minimize academic dishonesty. The assessments became online, non-invigilated and open book. A comparison of students' examination performances before and during the COVID-19 pandemic of a large first-year statistics unit shows that students' grades were not inflated or deflated due to the new assessments. The newly designed assessments were as good as or even better than the pre-COVID-19 assessments to quantify students learning while upholding academic integrity.

INTRODUCTION

The rise of the massification of the higher education sector can be traced back to the 1920s. The USA was the first country to introduce the concept of making higher education more accessible (Selyutin et al. 2017). According to UNESCO Institute for Statistics, Australia ranked 23 in countries with greater than one million higher education enrolments in 2016 (Calderon, 2018). The increase in the number of higher education students in Australia has two aspects. On the one hand, since 2008, enrolments of undergraduate local students with disadvantaged backgrounds rose significantly, for Indigenous students, students with disability, students from low socio-economic backgrounds, and regional and remote areas (Universities Australia, 2020). On the other hand, since 2001, International student enrolments have more than doubled (Universities Australia, 2020). The increased number of students created large first year classes with more than 1,000s of students, especially for disciplines required by other disciplines as foundation learning such as mathematics and statistics.

Over the years, academics around the world equipped themselves with methods to deal with large first year classes such as using learning management systems (LMS), Internet-based resources and lecture recordings (Baik et al., 2015). In terms of assessing student learning, many Australian universities have adopted online computer-based assessments. The implementation of online assessments assisted academics by reducing the large volumes of marking and assessment-related administration. Online assessments also enabled diversifying the assessment tasks, created opportunities for frequent assessments to provide timely and formative feedback on student progress. However, academics were caught unprepared to deal with instant move to fully online teaching and assessments due to the COVID-19 pandemic in 2020. During the first year of the pandemic, Australia along with New Zealand, had much better conditions than many countries in the world, that might be because Australia used strict lock-downs and border closures. The extent and period of closures within the country have varied significantly. Due to border closures, international students were unable to come to Australia, and interstate students were unable to come to the university. Therefore, it became essential to utilise hybrid learning and teaching methods, even when the restrictions were relaxed to allow face-to-face classes. The changes also required the use of non-invigilated online assessments due to having students in different countries with different COVID-19 restrictions. It is not unusual to have some proportion of non-invigilated assessments, however, having all of the assessments as non-invigilated is relatively new phenomenon. Educators are urged to consider designing assessments that safeguard academic integrity while ensuring constructive feedback can be provided to the students on a timely manner and students' learning can be assessed fairly.

We turned the crisis created by COVID-19 into an opportunity to modify learning materials to implicitly implement Bloom's taxonomy where the learning is measured by not remembering (a mere rote-learning strategy), but by evaluating and analysing (Bloom, 1956). Although the large class sizes make online assessment more challenging than small-sized second- or third-year units, through

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. ©2021 ISI/IASE iase-web.org/Conference_Proceedings.php

constructively aligned learning materials and creating many different versions of the questions in the assessments, it is possible to deter students from cheating, especially e-cheating. In this paper, we provide examples of open-ended questions which were designed for online non-invigilated final exam of a large first-year introductory statistics unit with more than one thousand students in each semester.

DATA SOURCES AND BACKGROUND

Two cohorts of students under comparison were from a large first-year statistics unit offered by a major metropolitan university in Sydney, Australia who were studying towards a Bachelor of Commerce degree. The topics covered in the units included descriptive statistics, graphical displays, two sample t-test, paired t-test, simple linear regression and chi-square test. The benchmark group consisted of students from 2019 who completed this course on-campus with 90% invigilated assessments where 50% was the sit-down, paper and pen final exam. Students from the 2020 cohort completed the unit mostly off-campus (online) with 100% non-invigilated assessments, including 50% final exam.

We extracted data for both groups to facilitate comparisons both for their abilities (e.g. weighted average mark (WAM), whether they failed the unit previously) and their assessment marks for the unit. At the time of the data extraction, WAM included students grades for the unit investigated here. We also recorded whether they were local or international students. The assessment marks included five hurdle quizzes (10%), two class tests (40%) and final exam (50%) as well as their grade for the unit (e.g. from Fail to High Distinction). Unique to this unit is the hurdle quiz component. Students must pass all five hurdle quizzes to pass this unit. Hurdle quizzes serve both as the formative and summative assessment to help students identify learning deficiencies and make students accountable for learning progress. They were online, non-invigilated with unlimited time and 10 attempts for both cohorts. There were 69 students who failed the unit due to being absence for the final exam in 2019 while the number was 106 in the 2020 student cohort. We excluded these students who did not sit the final exam from our analysis. In 2019, 2 students failed the unit due to failing the hurdle requirement, they were kept in the data. The final number of students included in the analysis were 1,119 in 2019, and 935 in 2020.

Majority (84%) of the students in both cohorts were taking the unit first time while 15% of them were repeating the unit from an earlier failure. Only 1% of them failed the unit more than once. The proportion of international students declined from 16.4% in 2019 to 11.7% in 2020. A slight decrease in the number of international students could be due to border closures during the COVID-19 (Tomazin, Millar, Carey 2021). Moreover, there was a fall in the total enrolment numbers, which could also be attributed to the COVID-19 pandemic.

EXAMPLES OF ASSESSMENT QUESTIONS

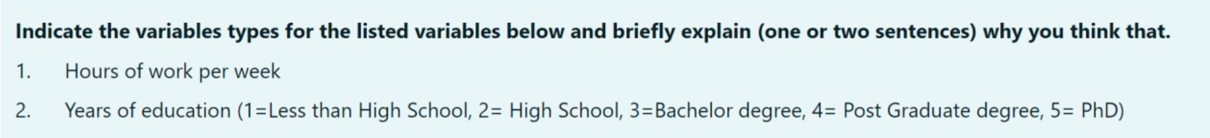
Most of the assessment questions designed to be marked automatically online, while some questions required manual marking. To prevent cheating on the automatically marked questions, we created different scenarios and used R-exam package (Zeileis et al. 2014) to create many different versions of questions for each scenario. For such questions, students downloaded a data set, analysed the data and entered their answers such as a numeric value (e.g. absolute value of test statistics) or chose an option from a drop-down menu (e.g. suitable conclusion for the results). Automatic marking ensured that assessment outcomes and feedback could be delivered to students promptly, enabling students to identify learning deficiencies and take remedial actions rapidly. Having many different versions and different scenarios safeguarded academic integrity as much as possible.

Student voice is not audible with automatically marked questions, regardless of how cleverly designed the questions are. It is possible to measure student learning at the higher levels of Bloom’s taxonomy with carefully designed multiple choice questions but student voice is lost and student thinking is hidden. To be able to hear student thinking, we created open ended questions where students required to type their answers. These questions focused on assessing students on higher levels of Bloom's taxonomy, such as focusing on evaluating and creating. Emphasis of these questions is assessing students’ conceptual understanding instead of focusing on procedural understanding. In this section, examples of newly designed open-ended assessment questions from the 2020 final exam are presented along with good and bad student answers.

Example of Justifying Their Answer - Variables Types Question

Identifying variables types is very important and crucial skill for statistics students. It could even be considered a *threshold concept* in statistics. With the correct identification of variables types, it is clear which graphics would be suitable for individual variables and for investigating the relationship between the variables. If a variable is numeric, histogram should be the graphical display, while for categorical variable, bar chart is the best. Investigating the relationship between a categorical and a numerical variable requires the use of boxplots. It is clear that *identification of variables types* is like finding the key to drive a car. Although necessary, not sufficient to move the car, but without the key, nothing could be done.

At face value, the variables types questions seem easy. However, experience show that students struggle to differentiate between different types of variables. If they fail to identify the variable type, then it is more than likely that they will not be able to choose suitable statistical technique(s) for their analysis. An example for variables types question is given in Figure 1 which was marked by giving 2 marks for the correct identification of variables types and 3 marks for the reasoning/justification of why.



Indicate the variables types for the listed variables below and briefly explain (one or two sentences) why you think that.

1. Hours of work per week
2. Years of education (1=Less than High School, 2= High School, 3=Bachelor degree, 4= Post Graduate degree, 5= PhD)

Figure 1. An example of variables types question

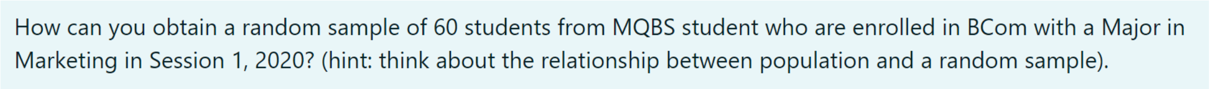
A good answer: *1. numerical, continuous. hours of work per week would not be expressed in an integer value and it is not discrete as you are able to work for example 38.5 hours per week not just 38 hours. 2. categorical, ordinal. this is because years of education are categories based in somewhat of an ordered fashion as they increase instead of in random order, for instance, you can't reach number 2 (high school) without completing number 1 (less than high school).*

A bad answer: *1. Discrete. This is because discrete are variables that can take integers values that are distinct like number of students in a class. This means hours of work per week is a similar so taking discrete as a variable type is right. 2. Continuous. This is because it takes variables within a certain range. Year of education is within a certain range from 1 to 5.*

Questions relating to variable identifications can be structured as multiple-choice questions or choosing answers from a drop-down menu, which is convenient but inappropriate for the online, non-invigilated assessment if the aim is to fairly assess student learning and prevent academic dishonesty. With open ended question, students required to consider what they know about variables types, analyse the given information and formulate their answers to the question. Allowing students to provide justifications for their answers provided an opportunity for students to elaborate their ideas and articulate their thought processes. This question was structured to determine whether students attained the ability to operate at *applying or analysing* level (Bloom, 1956). The bad answer with zero marks is an example of student's inability to understand the variables types.

Example of Applying Their Knowledge to a Context to Obtain a Random Sample

Students struggle to identify *population* for a given problem which creates difficulties for them to explain whether a sample is a *representative sample* and/or whether a sample is a *random sample*. They might memorise or take (allowable) notes with them to assessments to answer such questions. To be able to assess students thinking, we assessed such concepts by giving a context (Figure 2). We expected students to write about a) need for population frame to enable equal chance for each element in the population to obtain a random sample and b) a mechanism for selecting randomly.



How can you obtain a random sample of 60 students from MQBS student who are enrolled in BCom with a Major in Marketing in Session 1, 2020? (hint: think about the relationship between population and a random sample).

Figure 2. An example of obtaining random sample question

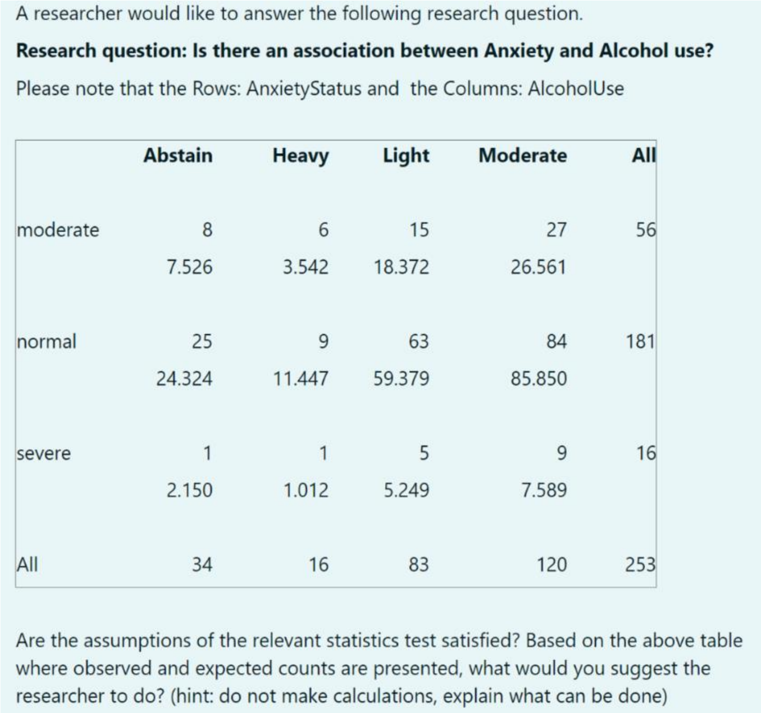
A good answer: *We need to take into account all the students in the BCom class with a major of Marketing for Session 1, 2020 as this population will be N in the test conducted. A sample of 60 from N is needed so we must assign a random number to each student that is part of the population. Hypothetically if the population is 5500 then we assign from 0001 to 5500. The students selected will be conducted from a random number generator or table until you have 60 numbers selected for the random sample.*

A bad answer: *To ensure that the population obtained is completely random, you must take from the population (p) the sample (p hat) without taking anything else into consideration. This is done by firstly defining the population and identifying a sample size smaller than said population. After this, one must assign numerical values to each member of the population as to avoid bias and select randomly based off the numerical values alone.*

Instead of directly asking students *the procedure to obtain a random sample*, which could be easily looked up from external resources, especially during online non-invigilated assessments (e-cheating), the question was structured as an open-ended question that required students to operate at the highest level of Bloom's taxonomy, *create*. To demonstrate their understanding necessary for creating, students need to consolidate all of their knowledge and assemble relevant facts to formulate a strategy to answer the question.

Example of Applying Their Procedural Knowledge to a Given Problem – Chi-square Question

Last two weeks of the semester, students learn about Chi-square tests (e.g. goodness-of-fit test, test of independence), assumptions for Chi-square test, steps for the calculations and what kind of research questions can be answered by utilizing them. They are expected to be able to use software to solve such problems as well as being able to make calculations without software. Usually, they are good at following steps to reach a numerical solution, just like following a recipe for cooking. In the online assessment, we assessed their procedural knowledge where they were required to enter their numerical solutions and chose from a drop-down menu the best explanation for the given solution. With an open-ended question, we also wanted to assess whether students really understood the assumptions for Chi-square and how they can deal with problems when the assumptions are not satisfied (Figure 3).



A researcher would like to answer the following research question.

Research question: Is there an association between Anxiety and Alcohol use?

Please note that the Rows: AnxietyStatus and the Columns: AlcoholUse

	Abstain	Heavy	Light	Moderate	All
moderate	8	6	15	27	56
	7.526	3.542	18.372	26.561	
normal	25	9	63	84	181
	24.324	11.447	59.379	85.850	
severe	1	1	5	9	16
	2.150	1.012	5.249	7.589	
All	34	16	83	120	253

Are the assumptions of the relevant statistics test satisfied? Based on the above table where observed and expected counts are presented, what would you suggest the researcher to do? (hint: do not make calculations, explain what can be done)

Figure 3. An example of Chi-square question

We expected the students to notice that some expected counts were less than 5, therefore the assumptions for Chi-square test is not satisfied. A way forward could have been suggestions to collapse some categories, such as moderate and heavy alcohol consumption and drop some categories such as the severe category for anxiety or increasing the sample size to see whether assumptions could be satisfied with increased sample size. To obtain full mark for this question, students needed to identify the problem in the given table and suggest a solution.

A good answer: *The expected counts for both abstinent and heavy alcohol drinkers with severe anxiety are under 5, hence the assumptions for a chi-square test are not satisfied. Normally you can combine categories, however abstinent and heavy consumption are polar opposites so cannot be combined into one category, and the total would still not be equal to 5. It is recommended that the sample is retaken with a higher number of observations, strategically measured to include more abstinent and heavy drinkers.*

A bad answer: *To determine if there is an association between the two variables the researcher could run a chi-squared test between the two categorical variables and if variable X increase and variable Y increases at the same rate the correlation would be positive however if they were both decreasing at the same rate then their association would be negative. To determine this the researcher could work out the correlation by using the formula or putting the data into excel and finding the correlation through there.*

This is a typical Chi-square question examining students' understanding of the assumption underlying a valid Chi-square test. Structuring this question as a multiple-choice question or a short-answer question would lead to easy googling or looking up lecture slides by students during the online non-invigilated assessment. We constructed it to be an open-ended question, which required student thinking in the context and evaluated whether the pre-requisite for the chosen hypothesis test was satisfied or not. Besides identifying whether the condition was met or not, more importantly, the question asked students to come up with a solution if the condition failed. As a result, students must demonstrate the ability of evaluation (i.e. justify their decision) to gain full marks.

RESULTS

Students abilities in these two cohorts is measured with the comparison of WAM where on average the 2020 cohort (mean=65.76, std=11.49) was slightly better than the 2019 cohort (mean=63.53, std=11.11). Although the differences of average WAMs were statistically significant between the two cohorts (t (2046) = -4.431, p<0.001), the practical significance is questionable. The descriptive statistics for each assessment task for both cohorts is provided in Table 1. Pairwise comparisons of each assessment task using two sample t-tests provided evidence against the null hypothesis that there is no difference in the average marks between two student cohorts (p<0.0001 for each test).

Table 1. The comparisons of students' performance

Assessment Items	2019 cohort (n=1,119)			2020 cohort (n=935)		
	min	mean(std)	max	min	mean(std)	max
Class Test 1 (max 15)	0.00	11.43 (2.72)	15.00	0.00	11.78 (3.58)	15.00
Class Test 2 (max 25)	0.00	16.93 (4.77)	25.00	0.00	16.16 (5.85)	25.00
Hurdle quizzes (max 10)	4.27	8.662 (1.01)	10.00	3.23	8.46 (1.02)	10.00
Final exam (max 50)	0.00	29.02 (8.92)	48.00	0.00	31.65 9.30)	50.00
Unit Mark (max (100)	20.00	66.53 (14.63)	97.00	17.00	68.54 (15.83)	100.00

Although the differences in assessment scores are statistically significant, it is worth noting that the magnitude of these differences is not large enough for us to claim practical significance, especially when the sample size is large for both student cohorts. In summary, based on the comparisons, it is reasonable to conclude that with the newly designed assessments, students enrolled during the COVID-19 period perform on par with those who studied in the same course before the COVID-19 pandemic.

In addition to individual assessment marks, we also compared the grade distributions (Table 2). Chi-square goodness of fit test, used to examine whether 2020's grade distribution was similar to the grade distribution of 2019, provided evidence against the null hypothesis ($\chi^2(4)=41.66$, $p<0.001$). Specifically, higher proportions of students obtained grades in Credit (CR), Distinction (D) and High Distinction (HD), while lower proportions of students obtained a Pass grade (P) or failed (F) in 2020 compared to 2019.

Table 2. The comparisons of students' grades

	HD	D	CR	P	F
2020	151	214	230	241	99
2019	140	225	238	393	123

CONCLUSION

The newly designed assessments introduced in a large first-year statistics unit enabled us to turn crisis into an opportunity and to strengthen academic integrity. Though there are many reflections on dealing with academic integrity during COVID-19, little has been discussed in the context of large first-year statistics units. This paper fills a much-needed gap in this regard. The combination of analysis (automatically marked) and open-ended (manually marked) questions enabled us to evaluate student learning remotely via online assessments without invigilation during the COVID-19 university campus lockdown. The comparison of the students' performances before (2019) and during the COVID-19 (2020) showed that the newly designed assessments are as good as the previous assessments to assess students' learning while providing opportunities for feedback, even for final exam.

More than a decade ago, Ward (2004) found no differences performance of students in a hybrid versus traditional unit. In this paper, we showed that the differences in performance of students before (traditional assessments) and during COVID-19 (fully online, non-invigilated assessments) were negligible. We conclude that fully online units with non-invigilated assessments do not disadvantage student performance. However, some students find it hard to study fully online due to bad time management skills and not being able to fully transition from high school to university life.

REFERENCES

Baik, C., Naylor, R., & Arkoudis, S. (2015). The first year experience in Australian universities: Findings from two decades, 1994-2014. *Melbourne Centre for the study of higher education*.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain*. New York: David McKay.

Calderon, A.J. (2018). *Massification of higher education revisited*. Melbourne: RMIT University.

Martin, L. (2020). Foundations for good practice: The student experience of online learning in Australian higher education during the COVID-19 pandemic. *Australian Government Tertiary Education Quality and Standards Agency*.

Selyutin, A.A., Kalashnikova, T.V., Danilova, N.E., Frolova, N.V. (2017). Massification of the Higher Education as a Way to Individual Subjective Wellbeing, in F. Casati, G.A. Barysheva, and W. Krieger (Eds.), *Proceedings of the III International Scientific Symposium on Lifelong Wellbeing in the World*, vol. 19, pp. 258–263, Tomsk Polytechnic University, Tomsk, Russia.

Tomazin, F., Millar R., Carey A. (2021). International student losses set to punch $18 billion hole in economy, *The Age*, 4 April 2021.

Universities Australia (2020). 2020 Higher Education Facts and Figures. Retrieved from https://www.universitiesaustralia.edu.au/wp-content/uploads/2020/11/200917-HE-Facts-and-Figures-2020.pdf

Ward, B. (2004). The best of both worlds: A hybrid statistics course. *Journal of Statistics Education, 12*(3). Retrieved from http://jse.amstat.org/v12n3/ward.html

Zeileis, A., Umlauf,N., Leisch, F. (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. *Journal of Statistical Software*, *58*(1), 1–36. doi: 10.18637/jss.v058.i01