EMPOWERING NON-SPECIALISTS TO INTERPRET AND DISSEMINATE STATISTICS THROUGH STRUCTURED ASSESSMENTS

Thomas R. Honnor, Niloufar Abourashchi, and Matina J. Rassias
Department of Statistical Science, University College London, United Kingdom
t.honnor@ucl.ac.uk

Assessment is an important tool for quantifying each student's relative ability but when carefully designed can also be used to educate and empower students. Because the importance of statistical understanding is becoming increasingly recognised, there is a subsequent growth in non-specialist students taking statistics service courses. The assessments for such courses can define students' personal course aims and level of engagement and set the tone for their future interactions with the subject. We have designed and implemented a modernised assessment pattern, with assessments structured to build upon each other and lead students from the foundations of probability to interpretation and communication of authentic statistical analyses. We discuss our experiences having implemented this new assessment pattern across three courses, totaling more than 600 students.

INTRODUCTION

Assessment is an important part of any course. Although a main aim of assessment is to "provide a method for assigning numerical scores to determine letter grades" (Garfield, 1994, abstract), carefully structured assessments can also perform a secondary role of educating and empowering learners. It is natural that students will structure their approach to a course around the associated assessments and therefore all the more important that those assessments are well-designed.

The statistical sciences, with their roots in mathematics, have traditionally been assessed via timed examinations including numerical calculations and proofs. Students in programmes with major mathematical and statistical components have ample opportunity to develop the technical skills needed to succeed on such examinations; however, the increasing prevalence of statistical analyses in academia and across society has brought to the forefront the need for suitable statistics training for growing numbers of non-specialist students. Thought must be given to whether traditional examination formats are the best route to assess statistical understanding, and, if they are not, then which alternatives would be most beneficial.

A number of educators considered the same question over multiple decades. Garfield (1994) drew attention to the fact that "traditional forms of assessment … rarely reveal information about how students actually understand and can reason with statistical ideas or apply their knowledge to solving statistical problems" (abstract). Particularly relevant for non-specialist students, Chance (1997) stressed that "assessment should mirror the skills students will need in order to be effective communicators and evaluators of statistical information" (para. 1). Onwuegbuzie and Leech (2003) highlighted that students' approaches to learning are inherently influenced by assessment patterns and that when assigned a task "that necessitates problem solving or integrating knowledge, students will strive to understand and to apply information" (p. 123). Despite these recognitions of the importance of thoughtful assessment design, academia is often slow to change, with Garfield et al. (2011) reiterating that "it is time to assess our own assessment practices and instruments and to critically evaluate their relationship to the important, desired student outcomes" (p. 2). More recently, one of the six recommendations of the Guidelines for Assessment and Instruction in Statistics Education (GAISE College Report ASA Revision Committee, 2016) stressed the need to "use assessments to improve and evaluate student learning" (p. 3).

We have experience leading three statistics service courses, open to first- and second-year undergraduates. Content for all of the courses begins with the fundamentals of probability and concludes with multiple linear regression. The courses differ in student eligibility, but course content and assessment patterns are essentially identical. The courses are taken by a total of around 600 students a year, varying across programmes from Bachelor of Science in Chemistry with Mathematics to Bachelor of Arts in Linguistics. The prerequisite for all courses is a standard pass in mathematics assessments taken at age 16.

Accepting the need to reconsider our approach to assessments for these three service courses, this paper outlines the decision-making processes behind the design of a modernised assessment

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

pattern, structured specifically to educate and empower non-specialist students, and the experiences of educators and students following its implementation.

HISTORICAL ASSESSMENT PATTERN

In the academic year 2018/19, and for at least the five years prior to this, assessments for the department's service courses were made up of two components: (a) an in-class examination with a time limit of 40 minutes that contributed 10% of the course mark and (b) a final examination with a time limit of two and a half hours that contributed 90% of the course mark. Typically, questions on these examinations required students to manipulate probabilities, conduct hypothesis tests, and perform other procedures, all of which typically produced numerical results. Students were allowed to use a standard scientific calculator and were provided with statistical tables, with which they could determine probabilities and quantiles for necessary distributions. In an attempt to reduce the focus on memorisation, assessments were conducted open book.

Assessments of this format were not without their strengths. The absolution of a final answer that is correct or incorrect made it possible to develop a rigorous marking scheme. Confidence in a marking scheme breeds confidence in the usefulness of the assessment to identify a student's ability from their performance. There was additionally a degree of comfort in the familiarity of this assessment format—assessors took confidence from similarities between these assessments and those which they set for specialist statistics students.

Examination-focused assessments such as these also come with drawbacks. The use of scientific calculators and statistical tables was not reflective of the ways in which statistical techniques are implemented in the modern era. All of the department's service courses included a computing element (Stata or SPSS in recent years), but students viewed these as being less important because they did not feature in the assessments. Similarly, the focus on questions with numerical results led students to neglect the skills necessary to communicate the results of an analysis. The restrictions on time and resources available to students during an examination implicitly put limitations on the variety of questions that could be asked. Datasets would often be limited to single-digit sample sizes and linked to contrived, inauthentic scenarios. A result of these limitations was that "even if students leave these classes able to perform routine procedures and tests, they do not have the big picture of the statistical process that will allow them to solve unfamiliar problems and to articulate and apply their understanding" (Garfield et al., 2012, p. 885). A focus on numerical questions also did little to accommodate non-specialists with lesser and weaker mathematical backgrounds, if anything only reinforcing mathematics anxiety (Dowker et al., 2016) amongst affected students.

MODERNISED ASSESSMENT PATTERN

The new structured assessment pattern includes three different components.

Online Quizzes

The historical approach of assessments via examinations had some strengths, as outlined above. It was therefore decided to include two online quizzes in the new assessment pattern, which combine to contribute 25% of the total course mark.

Quiz questions were prepared using the 'exams' package within the R statistical computing programming language (Grun & Zeileis, 2009; Zeileis et al., 2014) before being uploaded to the Virtual Learning Environment (VLE, Moodle in this instance). This approach to question generation allowed large numbers of question variants to be produced, each requiring identical thought processes and calculation steps to solve but involving different numerical values, at essentially zero additional cost to the assessor. The assessment question generation mechanism was expanded to the generation of effectively unlimited numbers of practice questions, allowing those students who learn best by repetition the facility to do so, again at essentially zero additional cost to the course organiser. Although the time taken to hone the skills required to generate questions in this way is certainly non-negligible, the VLE's ability to automatically mark students' quizzes led to a net reduction in the workload burden on the assessor in comparison with the previous in-class tests.

By dividing this component of the assessment into two online quizzes, students were able to receive results and feedback from the first quiz before attempting the second quiz. Furthermore, it meant that the content examined by each of the quizzes was limited to only one half of the course

material. Both the intermediate feedback and reduced cognitive load improved the accessibility of the course to those with specific challenges around traditional mathematical examinations.

Comprehension of a Published Data Analysis, Via Group Coursework

Post-university exposure to statistics for many non-specialists will likely be through the interpretation of published statistical analyses, whether these are as formal as published papers in an academic setting or as general as reports on the news. This is the focus of the second assessment component, contributing 25% of the total course mark, for which students are provided with a journal article including a reasonable component of data analysis and are asked to comment on the following:

- The approach the authors took to data collection, with corresponding strengths and weaknesses.
- How they would apply techniques taught as part of the course to the data collected by the authors, to answer relevant and interesting questions.
- The application of a technique applied by the authors which is not taught as part of the course.
- The results of the author's analyses and any limitations of those results.

Examples of published analyses set for this assessment include a study investigating the impact of financial incentives on smoking cessation in pregnant women (Berlin et al., 2021) and a study investigating the effects of differing diets on the health and wellbeing of older adults in residential care (Iuliano et al., 2021). Clinical trials such as these have relatively straightforward designs and interventions that can be understood without deep application-specific knowledge, allowing the focus to be on the statistical components of the studies. Unlike historical examination questions, these studies and the issues they describe are real, bringing strong authenticity to discussions of study design and data collection.

The assessment is completed by groups of up to four students. Running the assessment in groups allows students to develop communication skills that might be stifled by collusion regulations were the assessment instead run on an individual basis. Students must first develop comprehension of the published paper for themselves, then be able to communicate that comprehension to their peers, before the group finally communicates their understanding to the assessor via a written submission, on the basis that "the use of writing assessments in statistics courses can be beneficial to students" (Woodard et al., 2020, p. 41). Students are not directly rewarded for how well they understand the concepts but for how well they are able to communicate that understanding to the assessor. This "ability to synthesize the components of a statistical study and to communicate the results in a clear manner" (Mustafa, 1996, abstract) was previously highlighted as one of the three specific competencies for the ability to use statistics in the real world.

A single course is limited in the number of techniques that can be taught, and statistics is an ever-developing field. The requirement of this assessment for students to extend their knowledge to a technique that isn't directly taught to them aims to empower them for the future, giving them the confidence that through self-directed research and discussion with their peers, they can understand new techniques—techniques whose importance is clearly highlighted by their inclusion in a published data analysis, thus helping to avoid the problem where "tools that are used to answer artificial questions will seem artificial too" (Smith, 1998, para. 5).

The element of this assessment component that requires students to discuss how they would analyse the data collected by the authors aims to prime them for the final component of the assessment pattern. Grades and feedback for this assessment are provided prior to the final component of the assessment pattern, identifying for students their misunderstandings and miscommunications to give them the opportunity for reflection and improvement before the final assessment component.

Investigation of a Dataset, Via Individual Coursework

The increasing ease with which data can be collected, and moves towards open data more generally, make it likely that the non-specialist students in these courses will at some point in the future analyse data for themselves. This is the focus of the final assessment component, contributing 50% of the total course mark, for which students are provided with a dataset and are asked to discuss:

- The results of their exploratory analyses.
- The most appropriate simple and multiple linear regression models for a specified output variable.
- Their application and interpretation of results for an extension technique not taught in the course.

Examples of datasets assigned for this assessment include the relationship between COVID-19 case rates, vaccination rates, and other demographic variables for subdivisions of London, and the relationship between crime rates, regulations on carrying a concealed weapon, and other demographic variables for U.S. states. As with the group coursework task, the intention is that the application is authentic but that the degree of application-specific knowledge required is limited. The use of real data is important, with research indicating "that actual data would be better able to demonstrate the real-life relevance of statistics" and "engender more interest, motivation, and engagement in students" (Neumann et al., 2013, p. 67).

When presented with a large dataset, students have little choice other than to take advantage of computational statistics software. With the powerful tools available in such software, there can be a strong temptation for students to apply great numbers of analyses and report endless numerical outputs, without understanding the methodologies or how the output should be interpreted. This assessment stresses the importance of communicating understanding and interpretations, hoping to further ingrain the importance of these aspects within students beyond their time on our courses. Assigning marks for these interpretations, in contrast to the purely numerical results, encourages students to draw contextual conclusions appropriate for a broader audience.

The linear regression section of the assessment requires students to explore different simple and multiple linear regression models to arrive at a final, "best" model. In doing so, students are exposed to the more artistic side of statistical modelling. Decisions are made on the basis of qualitative as well as quantitative results (for example, model fit diagnostic plots), and analyses from two individuals can disagree with neither being necessarily incorrect. This can come as a surprise to both students in science and in arts degree programmes.

The final part of this assessment is again designed to encourage students to venture outside of their comfort zone by researching, implementing, and discussing the results of a technique that is not taught as part of the course. The aim of this element is to take advantage of learning through assessment to empower students with the confidence that they can understand and communicate the results and limitations of more advanced statistical concepts.

RESULTS

The most reliable data is available for the largest of the three service courses. As a result, data from that course is used to draw comparisons between the historical and modernised assessment patterns. Three years of data are available for the historical assessment pattern, 2016–2019, and two years of data are available for the modernised assessment pattern, 2020–2022. Data for the intervening year of 2019–2020 is excluded from discussion because of the impact of the COVID-19 pandemic.

Numbers of students attending the course increased 22% from an average of 320 students to an average of 391 students following introduction of the modernised assessment pattern. The course is compulsory for some students and is available as an elective to essentially the entire undergraduate population of the university. Considering these groups separately, the average number of compulsory students increased 1% from 181 to 183, whereas the average number of optional students increased 49% from 139 to 208. This reflects that course expansion is being driven by increased student demand that, in the context of largely unchanging course material and lecturing, is consistent with the modernised assessment pattern being more attractive to students. Students in the course may alternatively be divided based upon whether they are in a science or arts degree programme. The number of sciences students increased 6% from an average of 280 to an average of 298, whereas the number of arts students increased 176% from an average of 33 to an average of 90. This is consistent with the modernised assessment pattern being more accessible to students without a strong background in traditional mathematical or scientific examinations.

All assessments under the historical and modernised assessment patterns were criterion-referenced as opposed to norm-referenced, using the same criterion. As a result, comparison of marks between the two assessment patterns can provide useful insights. Overall, course marks fell by 3.03 (95% confidence interval (CI) [1.46, 4.60]) percentage points following introduction of the modernised assessment pattern. This result could be confounded by the previously described differences in balance between compulsory and optional students. Focusing on the compulsory students alone, overall course marks fell by 0.20 (95% CI [-1.93, 2.34]) percentage points. This indicates that the increased number

and variety of assessments hasn't negatively impacted compulsory students, but there is work to be done to raise the level of achievement amongst optional students.

Under the historical assessment pattern, correlation between marks of the two assessment components was 0.56. In contrast, correlations for the modernised assessment pattern are 0.38 between the online quizzes and group coursework, 0.49 between the group coursework and individual coursework, and 0.60 between the online quizzes and individual coursework. These generally weaker correlations are consistent with the different components of the modernised assessment pattern requiring the application of different skills, providing evidence that the new assessments allow different avenues to success for students with different strengths and weaknesses.

Student feedback, collected in an end of course questionnaire, reflected positive opinions of the modernised assessment pattern:

- "Assessments 2 [group coursework] and 4 [individual coursework] are challenging but interesting, which allows me to apply the statistical methods in lectures to the data from real life and analyse the problems in new aspects."
- "I think in comparison with most years, this online Stata version was far more helpful than sitting for an in-person exam where most of the hard work goes into memorising formulas and life is much more stressful. I really like this new format and feel it can actually help me more in the real-world application of skills I learnt in third year projects etc."

Many course organisers dread the question, "Will this be on the exam?", used by a subset of students to identify what they then interpret to be the only important parts of the course and to justify their subsequent lack of engagement with other aspects of the course. This question loses its potency under the increased variety of the modernised assessment pattern. One side effect of such a question is "teaching to the test," whereby course organisers design and present material with the main aim of preparing students for the examination. The move towards a broader, more authentic approach to assessment has increased the flexibility with which teaching material can be designed and presented. Rather than introducing contrived examples to justify the application of a particular statistical technique, the focus can instead be reversed to the more realistic scenario of determining which technique should be applied to a particular set of data to answer a specific question.

A major limitation of the inferences that can be drawn between the comparison of the two assessments approaches arises from the fact that no single student completed both assessment patterns, and an external measure of statistical competency was not applied. The success of future modifications to course or assessment pattern design could be more effectively quantified through the use of an additional assessment that is held constant, such as the Comprehensive Assessment of Outcomes in a first Statistics course test (Garfield & delMas, 2010), the Basic Literacy in Statistics assessment (Ziegler & Garfield, 2018), or the Statistical Anxiety Rating Scale (Hanna et al., 2008).

The number one drawback of the modernised assessment pattern is the increase in workload for course organisers. We have found that the positives of the new assessment pattern outweigh this drawback, but this is certainly something that should be considered by any course organiser considering changing the assessment pattern of their course.

CONCLUSIONS

Across three introductory statistical service courses incorporating over 600 students per year, we have introduced a modernised pattern of structured assessments, designed to empower students to interpret and produce analyses of data to solve authentic problems for real-world scenarios. Students are graded both on their level of understanding of course concepts and also on their ability to communicate that understanding, in recognition that improperly reported statistical analyses can have catastrophic direct effects and indirectly erode trust in the field of statistical sciences as a whole.

Following introduction of the modernised assessment pattern, more students are opting to take the courses; average marks have remained constant; and correlation between assessment component marks has reduced. Broadening the scope of assessment has unlocked the scope of teaching on the courses, leaving open future changes to course material. These benefits may come at the cost of increased workload for the course organiser, but thoughtful assessment design and technical solutions can ensure that this cost is sufficiently outweighed by the benefits.

We recommend that all course organisers consider the pattern of assessments for their course. Do the assessments serve to improve each student's ability to meet the aims and objectives of the

course, or are they a relic of a potentially outdated historical system used solely to organise students into different grades of achievement? In the latter case, course organisers may take confidence and inspiration for the introduction of a modernised assessment pattern from this case study.

REFERENCES

Berlin, I., Berlin, N., Malecot, M., Breton, M., Jusot, F., & Goldzahl, L. (2021). Financial incentives for smoking cessation in pregnancy: Multicentre randomised controlled trial. *BMJ*, *375,* Article e065217. https://doi.org/10.1136/bmj-2021-065217

Chance, B. L. (1997). Experiences with authentic assessment techniques in an introductory statistics course. *Journal of Statistics Education*, *5*(3). https://doi.org/10.1080/10691898.1997.11910596

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*, Article 508. https://doi.org/10.3389/fpsyg.2016.00508

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/gaisecollege_full.pdf

Garfield, J. (1994). Beyond testing and grading: Using assessment to improve student learning. *Journal of Statistics Education*, *2*(1). https://doi.org/10.1080/10691898.1994.11910462

Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2–7. https://doi.org/10.1111/j.1467-9639.2009.00373.x

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G. W., Chance, B. L., & Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, *65*(1), 1–10. https://doi.org/10.1198/tast.2011.08241

Grun, B., & Zeileis, A. (2009). Automatic generation of exams in R. *Journal of Statistical Software, 29*(10)*,* 1–14. https://doi.org/10.18637/jss.v029.i10

Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences*, *45*(1), 68–74. https://doi.org/10.1016/j.paid.2008.02.021

Iuliano, S., Poon, S., Robbins, J., Bui, M., Wang, X., De Groot, L., Van Loan, M., Ghasem Zadeh, A., Nguyen, T., & Seeman, E. (2021). Effect of dietary sources of calcium and protein on hip fractures and falls in older adults in residential care: Cluster randomised controlled trial. *BMJ*, *375*(2364). https://doi.org/10.1136/bmj.n2364

Mustafa, R. Y. (1996). The challenge of teaching statistics to non-specialists. *Journal of Statistics Education*, *4*(1). https://doi.org/10.1080/10691898.1996.11910504

Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, *12*(2), 59–70. https://doi.org/10.52041/serj.v12i2.304

Onwuegbuzie, A. J., & Leech, N. L. (2003). Assessment in statistics courses: More than a tool for evaluation. *Assessment & Evaluation in Higher Education*, *28*(2), 115–127. https://doi.org/10.1080/02602930301670

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education*, *6*(3). https://doi.org/10.1080/10691898.1998.11910623

Woodard, V., Lee, H., & Woodard, R. (2020). Writing assignments to assess statistical thinking. *Journal of Statistics Education*, *28*(1), 32–44. https://doi.org/10.1080/10691898.2019.1696257

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. https://doi.org/10.18637/jss.v058.i01

Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, *17*(2), 161–178. https://doi.org/10.52041/serj.v17i2.164