Assessing Psychology Student Applied Knowledge of Statistics via Open-book Multiple Choice Online Exams

RESEARCH

SARVEN SAVIA MCLINTON
SHARON ELIZABETH WELLS
*Author affiliations can be found in the back matter of this article

ABSTRACT

Real-world applications of statistics are rarely 'off the top of your head'; however, statistics and research methods courses default to closed-book exams that only test rote learning. Trending research supports open-book exams testing the application of student knowledge rather than memory, however statistics courses in psychology are lagging amidst fears of cheating in online open-book multiple-choice exams. The aim of this study was twofold; first, to develop an online open-book multiple-choice exam that tests the application of psychology statistics and research methods knowledge, and second, to demonstrate that it is just as reliable a source of final grades as traditional closed-book exams. We compared results from a new Applied Exam (N = 104 undergraduate third-year psychology statistics students) with the previous year's Traditional Exam (N = 81), correlating these with Research Report grades (the best course-assessment indicator of real-world performance). Similarly strong positive correlations were observed between the written assessments and the Traditional Exam (.59**) or Applied Exam (.54**), and both exams display comparable bell curves for grade differentiation, suggesting we can depend on the new Applied Exam for final course grade data. It also reflects a better alignment with course objectives and graduate qualities for effective problem solving in novel situations. Automated assessment of applied knowledge benefits psychology instructors and organisations in reducing administration, and psychology students by alleviating the anxiety in closed-book invigilated exams. Together this presents an opportunity to improve student outcomes by encouraging the development of real-world skills, preparing them for competitive job markets that value critical thinking.

CORRESPONDING AUTHOR:
Sarven Savia McLinton
University of South Australia, AU
sarven.mclinton@unisa.edu.au

KEYWORDS:
assessment; evaluation; testing; online learning; scholarship of teaching and learning; psychology statistics; open-book exams

TO CITE THIS ARTICLE:
McLinton, S. S., & Wells, S. E. (2023). Assessing Psychology Student Applied Knowledge of Statistics via Open-book Multiple Choice Online Exams. *Designs for Learning,* 15(1), 58–69. DOI: https://doi.org/10.16993/dfl.211

INTRODUCTION

The traditional education system, especially in Science, Technology, Engineering and Mathematics (STEM), revolves around assessing rote-learned information via final exams; this is at odds with both the demand for online education, and the applied skills desirable in the job market. Our study has two objectives: first, to explore whether psychology statistics courses which are traditionally 'rote-learned' can instead move toward online open-book multiple-choice examinations that better test students' application of skills (rather than memorisation), and; second, to identify whether that new applied exam can still be a reliable source of final grade data. Whilst other subjects have quickly adapted to the recent drive toward digital open-book exams, statistics courses for psychology have lagged, likely due to the outdated expectation that questions should be heavily mathematical in nature, instead of seeking to test the application of research methodology skills. Thankfully, new content is continually emerging such as the Open Learning Initiative (OLI; Carnegie Mellon University, 2023), but we have yet to establish via direct comparison whether these scenario-based online exercises can also be used as a reliable source of final grade data in statistics courses (in the Psychology discipline in particular). In a recent study, Goedl and Malla (2020) call for researchers to pay attention to the design of digital non-proctored testing, such that grade equivalency can still be established with traditional invigilated exams. If successful, it would provide advantages for; 1) psychology statistics instructors, due to automated marking rather than time-consuming assessment of written responses; 2) psychology faculties, via reduced invigilation and administration costs, and; 3) psychology students, in reducing their anxiety toward being tested on mathematical-style course content in their otherwise heavily social sciences field.

Unlike STEM, the 'soft sciences' see a diversity in teaching of concepts, conceptualisation and research methodologies (Vo et al., 2017). Although psychology is a non-STEM discipline, statistical method courses are commonly required in the first year of study. This presents a challenge for educators, as the demographic of psychology students rarely includes those who enter the degree wanting to 'crunch numbers' and learn statistics (Laiu et al., 2014). Up to 39% of psychology students experience a negative attitude towards learning statistics and commonly question the relevance for their future careers (Griffith et al., 2012). Not only is there an inherent challenge in teaching statistics and research methods to psychology students, but this is further complicated by the drive to move education online (Forsey et al., 2013). Although the shift to Massive Open Online Courses (MOOC) holds some favourable views such as the opportunity for students to share their understanding and be involved in the process of changing or adding to the course (Richardson, 2003; Forsey et al., 2013), students enrolled in statistics courses are at risk of diminished engagement due to perceived anxiety and attitudes towards statistics (Onwuegbuzie & Wilson, 2003).

Studies on statistics courses that use the flipped classroom approach (i.e., learning course material online before applying knowledge in face-to-face classes; Forsey et al., 2013) have indicated that online testing equates to better retention of learnt knowledge 21 months later (Winquist & Carlson, 2014). The flipped classroom improves exam and quiz performance even when controlling for maths anxiety (Nielsen et al., 2018), but this raises questions about the future for examinations as MOOCs set a precedent for more Universities to push their content delivery online. Psychology students have suggested that there are consequences of online learning and subsequent testing, in that they examine only a restricted band of information learnt (Jensen, 2011). To begin, conventional exams test students' ability to rote learn material via closed-book examinations, which whether multiple-choice or otherwise, appeals to the lowest orders in the taxonomy of cognition ('Remembering' and 'Understanding') because they are a test of student memory and general comprehension of the question (Anderson & Krathwohl, 2001). Examinations tend to default to closed-book, conducted in foreign environments with an artificial testing setting and invigilators to monitor academic integrity (Agarwal et al., 2008), which is another example of how STEM models are a poor match for psychology students who are attempting to learn statistics in an otherwise non-STEM program. Statistics is conventionally tested via rote learning which involves memorising names, numbers, theories and correct methodology. However, this approach to learning statistics is unrealistic as the real-world application of statistics is rarely 'off the top of your head', rather requiring more higher order functions like 'Applying', 'Analysing', and 'Evaluating' (Anderson & Krathwohl, 2001) in the process of accessing resources, exploring, and trial and error, which are important in identifying the right approaches and interpretation of statistics (Tu & Snyder, 2017). This raises the question of whether the typical closed-book approach to exams in statistics courses are even aligned to the course objectives (Biggs, 2003); students may be trapped learning skills that do not actually reflect the intended learning outcomes that make them a desirable psychology statistics graduate. By exploring the open-book design space for statistics exams we better set up undergraduate psychology students for success and begin to tackle the challenges experienced by educators in coordinating STEM-style courses within non-STEM disciplines. However, when there are objectively correct and incorrect answers (e.g., numbers or calculations) it becomes a challenge to assess statistics knowledge; a

conventional exam would force unrealistic memorisation, whereas an open-book exam provides the opportunity to achieve 100% with access to resources. Open-book exams certainly support the new flipped classroom and blended learning approaches, but can an open-book format be explored for psychology statistics?

Other STEM fields have caught up with modern teaching (Johanns et al., 2017), and are evolving away from closed-book exams which necessitate students rote-memorise material. Open-book exams encourage higher-level thinking skills (Agarwal et al., 2008), increase learning by up to 57% (Myyry & Joutsenvirta, 2015), and may promote more comprehensive understanding as this represents situations closer to real world expectations (Stowell, 2015). In general, students either agreed (27%) or strongly agreed (73%) that open-book open web exams were preferable to closed-book (Williams & Wong, 2009). In fact, one study found students are more likely to use deeper critical thinking skills, synthesize and evaluate information, experience less anxiety, and be organised when prepping for open-book exams (Theophilides & Koutselini, 2000), but report that the exam is not necessarily easier. Given the challenge in examining statistics knowledge that would conventionally be measured via rote-learned information, the question becomes whether it is possible to create a valid test that still taps into the construct of statistics skills (Messick, 1989) but via an open-book format, which is our first research objective. Our second objective is to elucidate whether that new test still differentiates students for the purposes of generating final grades that accurately reflect individual student knowledge of course material.

The immediate assumption is to create an exam that requires students to write open-ended responses to express their knowledge about a given question, e.g., short-answer responses to questions about statistics. Despite the benefit of seeing a students' freeform understanding of a topic, there are major challenges such disadvantaging English as Second Language students, and long marking times. Since the Global Financial Crisis in 2007 and the lasting impact of COVID-19, university budgets underline the need for cost-effective solutions. Multiple-choice question (MCQ) exams allow for greater sampling of testing, meaning higher reproducibility and substantial reduction in potential for examiner bias (Hift, 2014). In fact, MCQs have been shown to assess higher-order cognition as effectively as written prose (Hift, 2014), with 4–8 multiple-choice questions potentially providing the same amount of information as one essay (Lukhele et al., 1994).

Once made open-book, a good MCQ cannot simply test memory recall or information that can be answered in a search engine. In order to differentiate between students who actively understand the content, online MCQ open-book exams must be novel and require the student to interpret contextual information, and possibly apply a suite of skills in concert. Hift (2014) draws upon Conceptual Change Theory which suggests that knowledge and concepts are linked as mental representations, and empirical evidence indicates multiple-choice exam formats can predict proficiency as they require students to internally represent situations which involve problem-solving. Specifically, applied knowledge vignettes (i.e., 'scenario-based') with a context-rich multiple-choice format require students to internalise the information provided, form an accurate mental representation, which in turn interacts with the relevant mental model, resulting in the student selecting the appropriate solution (Hift, 2014). This type of scenario-based approach has been pioneered by online teaching platforms such as the aforementioned OLI, which embeds continuous learning quiz questions that revolve around hypothetical research or statistics scenarios (for a relevant example, see the Statistical Reasoning [Open+Free] Syllabus; Carnegie Mellon University, 2023). This format involves active engagement and deeper understanding of the course content, and without the ease of locating a fast definitive answer in a simple online search it also safeguards against cheating. Whilst such courses see students apply their new knowledge through these embedded scenario-based MCQs, the primary drivers behind their development are continual learning and progress checks; as yet there has been no empirical demonstration that applied knowledge MCQs are a reliable source of final grade data in psychology statistics courses via direct comparison with their traditional invigilated exam equivalent.

Therefore, the aim of this study was twofold; first, to develop an online open-book multiple-choice exam that tests the application of psychology statistics and research methods knowledge; and second, to demonstrate that it is just as reliable a source of final grades as traditional closed-book exams, which in comparison require conventional invigilation and test only rote knowledge recall. However, this is the bare minimum goal, i.e., establishing alternative forms of testing that are at least similarly reliable sources of final grade data. In the spirit of continual improvement we propose that an open-book exam that tests the application of knowledge rather than rote-recall actually better appeals to the higher order cognitive skills that Bloom's Taxonomy espouses as educational ideals (Anderson & Krathwohl, 2001). Beyond the principles around testing approach and content, the ideal new exam would also account for logistical limitations for instructors and organisations, meaning that open-book testing for psychology statistics courses may be delivered online and use multiple-choice response. Should this new format be just as reliable as invigilated exams, then there are extensive side benefits to the instructors in reduced administration time, the students in reducing public exam hall anxiety, not to mention financial benefits to the organisation.

Therefore, the overarching theme behind our two aims is the research question:

> **RQ:** 'Can an online open-book multiple-choice examination be as reliable as traditional invigilated exams for psychology statistics?'

METHOD

The primary author had been coordinating the third-year undergraduate psychology statistics course 'Advanced Research Methods' since 2018 and wanted to redesign a new final exam which would be guided by best practice informed by the literature. Final exam data from 2019 would be the benchmark for comparison: a traditional closed-book 'rote-learning' style multiple-choice question (MCQ) format.

Here we will detail how the first research objective was achieved: we aimed to develop an online, open-book multiple-choice exam that tests the application of psychology statistics and research methods knowledge (objective two, evaluating the exam's effectiveness will be presented in the Results section). To begin, our literature search identified evidence to suggest that MCQ tests and progress quizzes (but not explicitly exams) could be designed to promote the application of student knowledge to novel scenarios about research methods and statistics. A good example is the Statistical Reasoning module of the Open Learning Initiative (Carnegie Mellon University, 2023). OLI exercises begin with a Learning Objective, such as to identify whether a study is a 'controlled experiment' or an 'observational study', and subsequently students read the topic content and then complete a progress-check style question at the end. In the aforementioned OLI example, "Which of these is a matched pairs design?" is followed by two possible responses, one assigning half of the participants to an active condition (a localized drug) and half to a placebo, whereas the other response suggests they test both conditions on the same participant with one arm being the local drug and the other arm the placebo (Unit 3 Module 7 p.83; Carnegie Mellon University, 2023). The course is very well designed for its purposes; however, these embedded quiz questions are clearly for the goal of continual learning, i.e., progress checks that encourage students to apply recently learnt skills and provide extrinsic reinforcement upon correct answers. In most instances the answer is relatively easily identified if the student has at least a general understanding of the previous passage; exercises at the end of each segment are often entitled "Did I get this?".

The questions that we designed use the same approach to formatting in that we also ask an MCQ in relation to a hypothetical scenario. However, the fundamental goal behind our questions was to critically evaluate high level reasoning and lateral thinking (without hints or leading) such that students can be accurately differentiated into final grades for the course. This ties back to the second research objective of our study: we did not only set out to test application of statistics knowledge, but further we aimed to create a test that separates students as reliably as a traditional closed book final exam. Therefore, the methodology and format of questions may be similar, but the critical underlying design is different. Results data from other courses that use this scenario-based applied knowledge format are unavailable, but based on the framing and nature of questions we posit that the mean rate of correct answers would be higher (and have a 'flatter' distribution) due to the continual learning progress-check nature of their design.

Therefore, we sought to use the same scenario-based applied MCQ approach, but model our new test directly on our previous year's traditional (closed book and invigilated) final exam so as to directly evaluate whether the new applied exam would be a reliable source of final grade data for an undergraduate psychology statistics course. Development of the new exam was guided by the overarching aim set out in the Advanced Research Methods course information, which was for students to "attain advanced skills in the design, conduct, analysis and reporting of psychological research". More pointedly, the primary Course Objective (CO1) stipulated students must "Apply psychology knowledge to select appropriate research designs and the appropriate methods for collecting and analysing data", which was linked in the course information guide as fostering the third Graduate Quality (GQ3), wherein a student will become "an effective problem solver, capable of applying logical, critical, and creative thinking to a range of problems." Therefore, following the 'alignment' aspect of Constructive Alignment (Biggs, 2003), the new final exam for 2020 was designed with CO1 and GQ3 at the forefront to increase the likelihood of desired learning outcomes, as well as testing higher order cognitive skills on Bloom's revised Taxonomy (Anderson & Krathwohl, 2001). We tallied up the number of questions on each topic in the traditional closed-book invigilated exam and developed the same proportion of parallel versions that used a hypothetical scenario to test similar underlying knowledge, i.e., both final exams tested an equal proportion of required course content. This process took 7 weeks and was conducted by the course coordinator, after which the draft was shared with another psychology statistics academic for review. Some questions underwent major revisions (11.1%) considering feedback, and most of the remainder received minor changes in grammar or wording (75.5%). The revised exam was then pilot tested with three junior academics: two current and past tutors on the course to evaluate content itself; one unrelated to statistics whose role was proofing wording and readability. Track-change comments were received independently, then

a meeting was held to discuss perceptions, including face validity and the effectiveness of 'distractors', i.e., erroneous responses specifically designed to catch students who had a surface-level familiarity with the topic but insufficient understanding of the topic to recognise that it was incorrect. A unanimous decision was made to increase the presence of distractors (up to a total of 66.6% of questions) to improve the likelihood of success for research objective two, i.e., to reliably differentiate student performance. This revision and pilot testing phase ran for 5 weeks, bringing the total development time for the exam to 3 months. It was concluded that the final version reflected the expectations of student knowledge of statistics and research methods that could be delivered in an online open-book MCQ format, satisfying research objective one. Further detail on the exam itself and examples of the questions are presented in the Measures, and we then continue on to evaluating research objective two in the Results, i.e., testing whether a new open-book 'applied' style exam is as reliable as a traditional invigilated exam for deriving final grade data. Approval to access and publish on de-identified archival data was received from the University of South Australia Human Research Ethics Committee (Protocol #203505).

PARTICIPANTS

Students were enrolled in Advanced Research Methods and were completing undergraduate degrees in Psychology, Psychological Science, or Cognitive Neuroscience at the University of South Australia. The 2019 cohort consisted of N = 85 students (n = 81 once Fail results removed), whereas the 2020 cohort initially comprised N = 111 (n = 104, Fail removed). Removing 'Fail' results was an important step in comparing these two cohorts. Whilst traditional logic suggests that including students of all grade thresholds is relevant for analysis, in reality there are a number of reasons to focus only on students who score a final 50% Pass grade or higher. First, the reasons for course Fail results are varied and extend beyond the course content and structure with which this study is interested, including; late withdrawal, personal complications, placing study on hold, or completely disengaging from the course regardless of course content and quality of the exam. However, removing Fail results is most important for 2020, wherein COVID-19 led to university courses being moved online partway through the semester. This history effect is likely the reason for the slightly higher proportion of Fail results in this course (6.3% compared with 4.7% in 2019), with many students adversely impacted by the rapidly changing conditions surrounding the pandemic. By removing this small proportion of data in both cohorts, we more likely tap into the actual differences in digital course content and exam structure between 2019 and 2020.

The gender breakdown was roughly even between both cohorts, with a high female representation as seen in most psychology courses. The sample for analysis from the 2019 cohort consisted of 76.5% female (n = 62) and 23.5% male (n = 19) students, whilst the 2020 cohort comprised 71.2% female (n = 74) and 28.8% male (n = 30) students. Age was also similar across both years, with the mean age for 2019 cohort being 21.7 years (SD = 3.90) and the mean age for 2020 was 21.9 years (SD = 4.30).

MEASURES

Traditional Exam (2019)

The final exam for the course is worth 50% of a student's overall grade. This is a conventional exam sat in-person in an invigilated exam hall, and as always, traditional exam performance favours those who have memorised material. Questions often focus on rote-recall, for example "Which of the following is not one of the 4 stages in the PRISMA framework for systematic reviews?" a) *Screening*; b) *Eligibility*; c) *Identification*; d) *Sorting* (Correct). Therefore, the existing closed-book MCQ sought to test largely the first half of GQ3, i.e., when rote-recall fails, students apply logical processes to problem solving such as the process of elimination. All responses are multiple-choice with 4 possible responses, only one of which is correct. The Traditional Exam contains 135 such questions, to be completed within 3 hours, and the grade for the final exam is the percentage of correct responses (with no penalty for an incorrect or missing selection). It is important to note that as a closed-book exam, instructors were afforded the liberty of relying on vast test banks of questions, which whilst convenient were not actually tailored to satisfy CO1 (only a small portion dealt with selecting appropriate research designs, in lieu of rote-recall of key facts).

Applied Exam (2020)

The final exam for the course is worth 50% of a student's overall grade. This is an open-book online exam that favours 'application' of skills and therefore tests critical understanding of content as well as the ability to source information and trouble-shoot creatively when complex problems arise. The process of elimination is still used like the 2019 Traditional Exam, however GQ3 is actually better tested in its entirety due to the focus moving away from rote-recall and more toward the application of deeper understanding in novel scenarios. All responses are multiple-choice with 4 possible responses, only one of which is correct. The Applied Exam is also to be completed within 3 hours, however unlike the Traditional Exam students may elect to commence this online at a time of their choosing within a two-day period. Due to the reduced need for rote-recall, each question features a vignette describing a specific scenario, and since questions are longer and more involved there are only 55 in total. The grade for this final exam is the percentage of correct responses, with no penalty for an incorrect or missing selection. Without invigilation (e.g., an online proctored exam) it was not possible to rely on a

test bank, and therefore all questions had to be designed bespoke. This process resulted in better addressing CO1 because all 55 questions required students to apply their knowledge to select the best fit design, methodology, or approach to collecting (or analysing) data. Presented in Figures 1 and 2 are two examples of questions created for the Applied Exam, the first of which is considered an 'easy' level based on the course content.

The easy question (Figure 1) saw 70.3% of students answer correctly, and the majority of the remaining students (20.8%) selecting the 'distractor' option designed to test whether the student fundamentally understands the underlying principles. second example (Figure 2) is a demonstration of a harder question, wherein only 36.6% of students answer correctly, and the remaining responses are roughly equally distributed between the other options (potentially at random). This type of question serves the role of identifying diligent students with a critical understanding of research principles. Answering multiples of these difficult questions are required in order earn a High Distinction and therefore clearly demonstrating CO1 and GQ3.

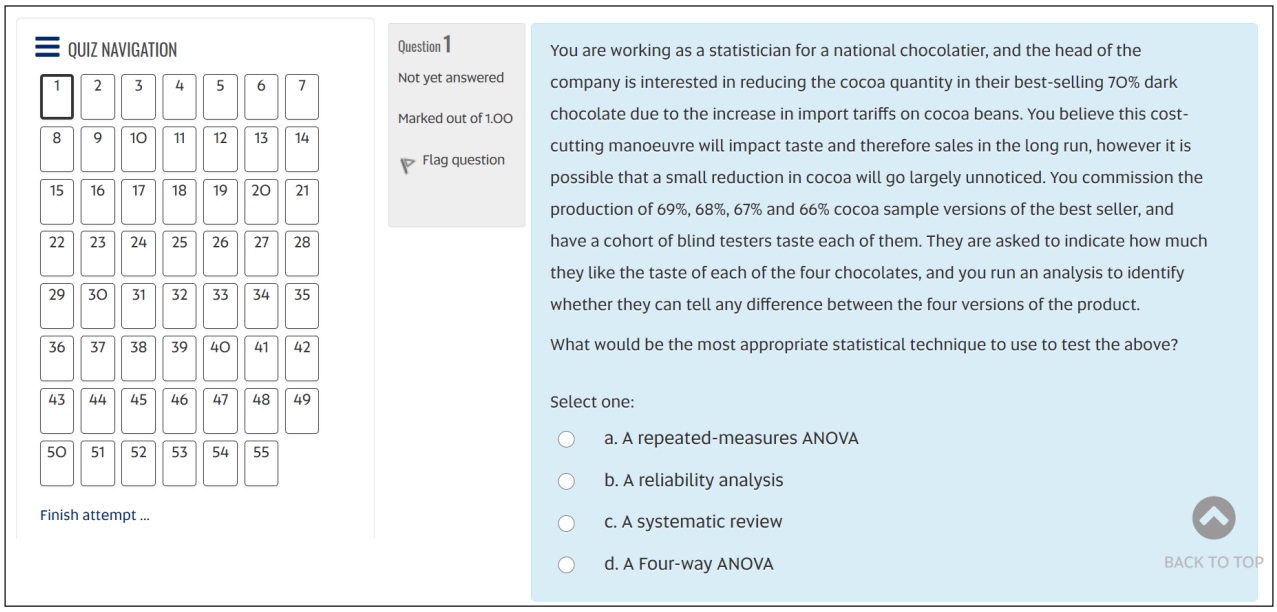

Figure 1 Example Applied Exam Question (level: easy).

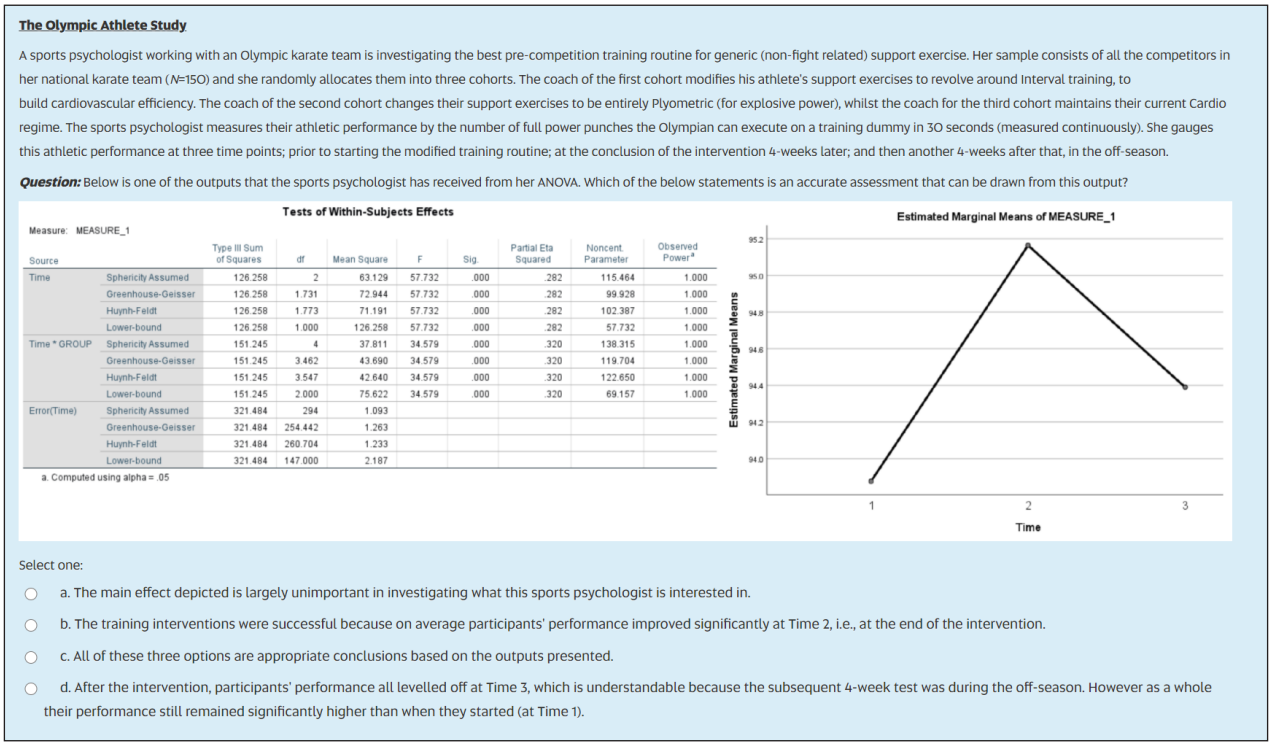

Figure 2 Example Applied Exam Question (level: difficult).

In developing the new exam to be better in line with CO1 we encountered an issue in that whilst very comparable, it was not possible to design the two exams to have a direct question-for-question comparison. For example, correcting for exam length, it was not the case that one question on PRISMA guidelines in the Applied Exam would parallel a commensurate 2–3 questions on the same topic in the Traditional Exam. This issue was due to both the relative ease with which some topics can be tested via rote-recall as opposed to application in novel scenarios, as well as the active desire to design an exam that was better grounded in CO1 and GQ3. However, we addressed this by ensuring that there were sufficient questions on the same topic. Taking the previous example, we replaced questions on the PRISMA itself with scenarios involving meta-analyses or systematic reviews in general, since the steps of a PRISMA guideline could be obtained via an online search engine much easier than the procedural skills and factors surrounding it. This further underlined the importance of testing whether both exams produced similar bell curve distributions for final grades (rather than item-for-item comparisons), as well as correlated similarly with other touchstones in the course such as research reports and progress quizzes (see Analysis Strategy below).

Research Report

This assessment is a 2,250 word scientific report which is marked by experienced course academic staff and worth 40% of a student's overall grade. Students identify a potential research question, conduct analyses on mock data, and then prepare their hypothetical report for submission to peer review in a fictitious academic journal. This research report is a good indication of real-world skills, because conceptualising research questions, conducting analyses, writing to a brief, and troubleshooting all use a variety of skills such as resource retrieval, information sorting, and a critical understanding of underlying principles and application thereof. Therefore, this assessment piece was deemed the best metric by which to evaluate the validity of the new Applied Exam (see Analysis Strategy below).

Progress Quiz

The progress quiz is an activity worth 10% of a student's grade in the course. It was originally developed in 2018 and is a precursor to the format of the Applied Exam (2020). The key differences are that questions are not weighted equally, and use some non-MCQ response formats such as 'drag and drop' words. For example, one such question is:

> You are reporting the results of a 2 × 2 ANOVA which was conducted to evaluate the effects of sex and age on adolescents' duration of sleep on school nights. How would you report the main effect of sex from the SPSS output in a single sentence in your results section using the options provided below?

A screenshot of an SPSS output is presented, based on which students then drag-and-drop the relevant numbers into the vacant boxes in the answer sentence:

> A significant main effect of Sex was found, with males exhibiting a longer sleep duration on school nights than females, F(___,___) = ___, p < ___, η^2 = ___.

Course Evaluation Instrument (CEI)

At the end of every course delivered by the University of South Australia, students are asked a series of questions that comprise a feedback mechanism to instructors to assist with redevelopment year-to-year. The response rates (2019 = 25.9%; 2020 = 24.3%) are reflective of the typical participation rate in this optional feedback instrument. Only a small amount of these data pertain to the final assessment piece and therefore no qualitative analysis will take place; instead a handful of quotes are used as demonstrative examples in the Discussion as a form of data enrichment for this otherwise quantitative study.

ANALYSIS STRATEGY

Data were extracted from the Advanced Research Methods course websites for 2019 and 2020, and imported into IBM SPSS 25 (Statistical Package for the Social Sciences) and screened for normality (skew and kurtosis). Due to the nature of the progress quiz assessment being un-timed and open for 7-days, scores tended to cluster at the top of the range, and as expected there were violations of skew and kurtosis for both years. Pearson Correlation was run to examine the relationship between students Exam and Research Report results, whilst Kendall Rank Correlation was used as a non-parametric option for comparing these to the progress quiz. Next, both 2020 and 2019 Exams were plotted side by side in histograms to identify whether there were any issues with distribution (i.e., to suggest disparity in difficulty/ease, or potential cheating). Last, we also extracted any feedback in which students mentioned the exam in the text-based comments offered at the end of the course, via the Course Evaluation Instrument.

RESULTS

Research Objective two was to demonstrate that the new exam is just as reliable a source of final grades as traditional closed-book exams. To begin, results showed both traditional exam (.59***) and applied open-book exam (.54***) display similarly strong positive correlations with the best indicator of real-world performance, i.e., correlation with Research Report. A correlation matrix of the three assessment pieces for each cohort is presented in Table 1.

ASSESSMENT		2019			2020		
		RESEARCH REPORT	PROGRESS QUIZ	TRADITIONAL EXAM	RESEARCH REPORT	PROGRESS QUIZ	APPLIED EXAM
2019	Research Report	1	.30***	.59***	.	.	.
	Progress Quiz	.30***	1	.41***	.	.	.
	Traditional Exam	.59***	.41***	1	.	.	.
2020	Research Report	.	.	.	1	.35***	.54***
	Progress Quiz	.	.	.	.35***	1	.39***
	Applied Exam	.	.	.	.54***	.39***	1

Table 1 Correlations for 2019 and 2020 Psychology Statistics Assessments.

Note: ***$p < .01$; $N = 104$(2020); $N = 81$(2019); Spearman correlation for Exam & Report; Kendall's Tau for correlations involving Progress Quiz.

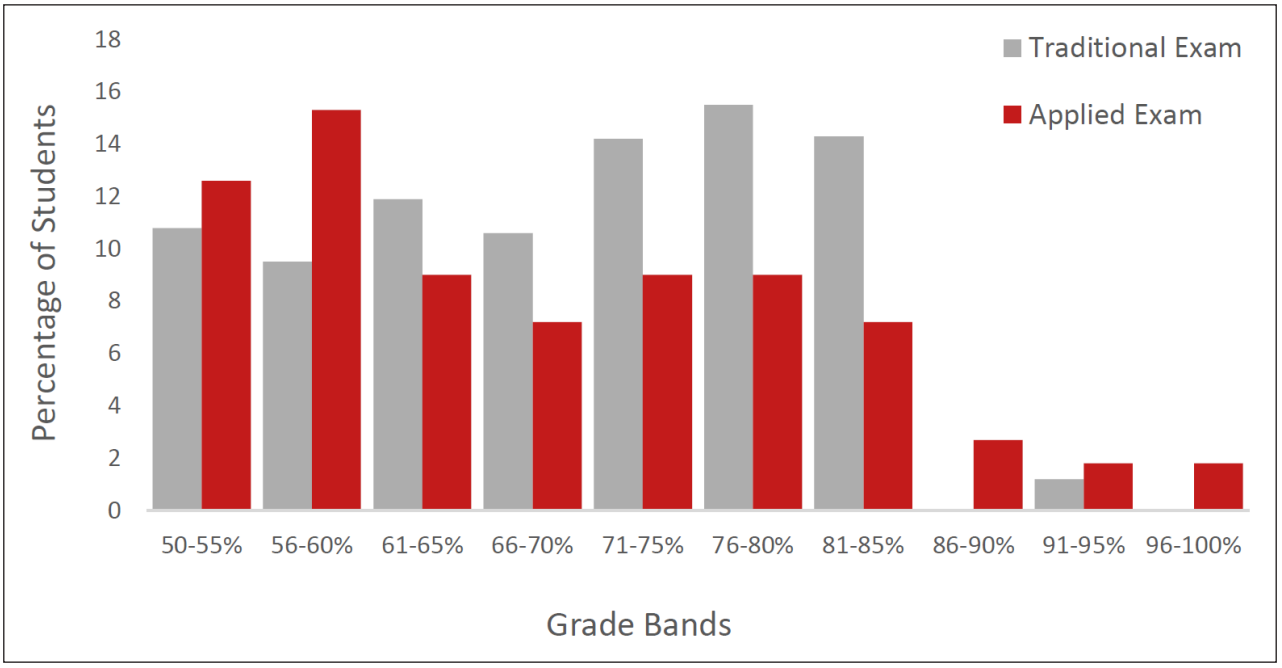

Figure 3 Exam grade percentage brackets for the 2019 Traditional vs. 2020 Applied Multiple-Choice Exams.

These metrics suggest that the new open-book approach is in line with other data points of the same student's performance in the course. First, the correlations with conventional measures of success in statistics like Research Report writing skills do follow a similar pattern regardless of year, which is the first step toward addressing our aim to identify whether a new Applied Exam could be a similarly reliable source of final grade data as a Traditional Exam. However, the histogram (Figure 3) suggests there are still some areas to improve upon.

A direct comparison of the histogram of grade distribution on both exams (Figure 3) suggests that the bulk of data matches student performance between 2019 and 2020, meaning that the open-book scenario-based applied exam is just as good as the conventional closed-book invigilated exam at deriving student final grade data. However, the 2020 Applied Exam has proportionally more scores in the Pass range (50% to 65% brackets) whilst the 2019 Traditional Exam has more scores in the Credit and Distinction Range (66% to 85%). Contrary to anticipated, the open-book nature of the exam did not lead to a highly skewed 'easier' exam, and instead appeared more challenging than a traditional closed-book invigilated exam overall.

DISCUSSION

Our results demonstrated that online open-book multiple-choice examinations can be designed to be a similarly reliable source of final grades as conventional exams, when evaluating psychology student knowledge of statistics. This is despite differences in the volume of

content on specific topics, some of which lend themselves to rote-recall and therefore were overrepresented in the Traditional Exam. Given that the course objectives and graduate qualities guided its development, the bespoke nature of the Applied Exam placed it better in line with the aims of the course itself, satisfying the alignment aspect of Brigg's (2003) Constructive Alignment principle. Our results suggest that psychology statistics courses can move toward open-book exams that better fit the type of cognitive learning skills we seek to foster in Advanced Research Methods (Anderson & Krathwohl, 2001), and moving away from the largely outdated concepts of testing only rote knowledge recall like conventional invigilated exams often do. The new exam better taps into the construct that psychology statistics lecturers want to see, namely using effective problem solving and creative thinking (GQ3) to identify the best approaches to design, methodology, and analysis (CO1). Although the results address our RQ because the Applied Exam can be just as reliable a source of final grade data for instructors to base student performance on, a far more important consequence has been providing the students with a final exam that better reflects the skills we wish to engender in them. With a new assessment model, our course prepares them for a far more realistic scenario when applying their skills in the workforce, an environment that cares more about the application of skills in novel scenarios rather than memorisation and recall of facts.

Beyond the improved quality and alignment of the Applied Exam, the new assessment format will lead to numerous advantages for instructors (reduced administration time), students (alleviating public exam hall anxiety), and organisations (reducing financial costs). The last point is of particular importance because in 2020, the COVID-19 pandemic impacted university budgets further as international students were grounded overseas, resulting in a $4.6 billion revenue gap across all Australian universities, leading to further job cuts (Duffy & Sas, 2020). Therefore, short-answer exam papers which require more time and markers are simply less feasible in the modern age due to the cost. Likewise, with the education sector having to cut budgets and rationalise staff, even the expense of invigilation in exams, collection of papers and associated administrative costs all add up.

From an administrative perspective, there are also a number of other side benefits for teaching staff using a digital non-proctored exam. There are major advantages in streamlining the logistics for creating, editing, cycling through a question pool, and rolling-out exams each semester. Likewise, after responses are collected the grade collation is instantaneous. However, beyond this administrative logistic there are also benefits in question-bank systems themselves, such as providing immediate calculations for discriminant validity and other metrics that allow poorly worded (or mis-coded) questions to be removed from assessment and revised for following years. Some systems even offer metrics such as the average length of time a student spends on a given screen, allowing the administrator to identify questions that may take proportionally longer, and even weight them differently if needed.

Whilst there were initial concerns that cheating would invalidate the use of an open-book exam for statistics in psychology, in our results, the histogram suggested that the exam was not 'easy' and there is little evidence to suggest that systematic cheating was possible or beneficial. The extant fears that online exams facilitate cheating is probably far less relevant when the questions are applied in nature, and when there is a time limit, randomised questions, and randomised responses. If anything, other studies have found that the open-book nature of an exam can in effect reduce the motivation to cheat as students know that they are allowed access to their material, and some research suggests little grade difference between open-book and closed-book exams (Stowell, 2015). Previous studies have explored online proctored exams by directly comparing the same exam paper under open versus closed-book settings, with the assumption that invigilation is a necessity to prevent cheating (e.g., see Daffin & Jones, 2018; Harmon & Lambrinos, 2010). However we specifically designed the Applied Exam bespoke to CO1 in such a way that proctoring was irrelevant, testing student comprehension because the answers would not be readily available in text books or search engines. Instead, they would have to creatively problem solve and apply their knowledge in novel situations (GQ3). Further, the use of proctored exams is not a strict upside, given that popular systems such as Proctor-U have a cost, including the burden placed on students and unsatisfactory student experiences (e.g., see Milone, Cortese, Balestrieri, & Pittenger, 2017).

Although the new online open-book multiple choice exam presents numerous benefits for our psychology statistics course, the notion that all exams should move to the new model is not necessarily clear-cut. The decision to use the Applied or Traditional Exam is accompanied by numerous pros and cons that need to be taken into consideration, especially in relation to which best meets the course objectives. For example, it is possible that the open-book nature of the Applied Exam in the present study served to reduce student anxiety (Schmidt et al., 2009), but conversely caused them to spend less time studying for that exam. Even if the content covered may be similar, the perceived difference in format (i.e., the non-proctored nature of the Applied Exam) may in itself explain why there is a slightly lower number of students earning a 'Distinction' (71–85%), and slightly higher representation in the 'Pass' (50–60%) and 'Fail' brackets (below 50%) and compared with the previous year. Given that other studies have suggested that one-third of students spend less time preparing for non-proctored

online exams (Myyry & Joutsenvirta, 2015), concerns are raised regarding learner autonomy and motivation (Nielsen et al., 2018). Therefore, we suggest that it may also be beneficial to build into the course a stratified learning activity that educates students on what an open-book exam entails and dismisses the assumption that answers for applied knowledge questions can simply be located within textbooks or search engines. Due to time limitations and the short turnaround required to develop and run an applied online exam in the COVID-19 period, it was not possible to present exact (or similar) exam questions for students to practice (e.g. past papers). This was a critique that arose in the Course Evaluation Instrument, with students suggesting: 'Maybe include more practice questions for exams', and 'A set of MCQ questions to practice with [in addition to the 10 students did as part of the Progress Quiz assessment] to use as a revision exercise would have been great.' It is therefore plausible to have integrated multiple-choice quizzes throughout the course with the same structure as the open-book exam, which provide feedback (even if the final exam used for grades does not). Multiple-choice quizzes with feedback could aid the development of applied knowledge skills and allow students to experience the type of questions posed in open-book exams. Students build on their knowledge throughout the course and apply those skills in the open-book exam; applied knowledge questions typically require multiple step processes, which tests students' ability to apply the knowledge base learnt in the statistics course to complex real-world scenarios. Conversely, open-book exams to test students' applied knowledge skills may not be appropriate for introductory statistics courses as it is unlikely that students have developed the skills necessary to apply learnt knowledge to complex scenarios.

The main limitation of open-book exams is the large investment of time taken to develop reliable applied knowledge questions, some of which should be new for each year. Nevertheless, students clearly valued the quality of the questions in the applied exam, with one citing in their feedback:

> The exam questions and assignment were very good at testing what we had learnt and the skills we had gathered. Their biggest strengths were the originality of the studies and the obvious thought and care that went into crafting the assignment and questions.

It is evident that the content is appreciated, if staff have the time in their workload to develop such questions. However, in general, students do prefer feedback on performance, and unfortunately given the sheer time taken in developing high quality reliable exam items, it is unlikely that feedback on individual question performance can be made public. Furthermore, the short timeframe for rolling-out the 2020 open-book multiple-choice exam meant that a pilot study was not possible. Therefore, the slight differences between 2019 and 2020 distributions may reflect challenges students faced by not having access to applied knowledge practice questions beforehand. It is possible that the exam questions were too difficult for some students, particularly those who expected rote memory type questions, and therefore, the results presented are limited to this study. Regardless, it is good in principle to vary types of assessment so that students have opportunities to perform effectively via different platforms and test skills in different ways, and there is still an undeniable advantage for students in presenting exam conditions that causes less distress.

Given the short timeframe to develop, refine, and roll-out this new applied exam for psychology statistics during the COVID-19 period, it was not possible to conduct a proper pilot study. However, in many respects our findings can be used as a pilot, in order to inform future developments, refinement, and subsequent research. Future studies could incorporate qualitative measures, similar to the anonymous feedback that is often provided via Course Evaluation Instruments in order to tap into the student experience of their applied exam. For example, in our study one student provided feedback in the Course Evaluation Instrument: 'Also, it was very difficult to switch between questions on different studies in different orders in the exam, it was quite overwhelming and confusing.' Whilst randomisation of item presentation was a safeguard against cheating, it is possible that future studies could investigate whether there is any measurable impact on student grades when questions are delivered at random. A middle-ground would be a pseudo-random sequence whereby similar studies or analysis approaches are presented in blocks in order to help structure the student's thought processes but without losing the random nature of presentation. In addition, future studies could expand on this by exploring the impact that applied open-book exams have on anxiety levels in statistics courses, given that psychology students in particular are a cohort that experience a high level of statistics-anxiety. Last, the relationship between student motivation to study and applied open-book exams could be further investigated. Previous research has indicated there is less impetus to revise under open-book conditions, but this has as yet not been explored for exams that apply student knowledge to hypothetical scenarios.

CONCLUSION

The COVID-19 pandemic presented a significant challenge for tertiary education around the world, but also provided the impetus to explore new, more streamlined approaches to exams. Although we initially set out to identify whether our new exam format would be just as reliable a source of final grade data

as traditional invigilated exams, what we actually learned was that using a robust and theoretically driven approach to design actually resulted in a bespoke new exam that actively addressed our course objectives and builds a better psychology statistics graduate. Whilst multiple-choice exams have historically been common in psychology research methods courses (due to the nature of a clearly delineated 'correct' answer in statistics), the initial fears regarding the design of open-book exams with this response format were unfounded. Our study suggests an open-book multiple-choice exam that evaluates applied knowledge is just as reliable in providing student final grade data as closed-book invigilated exams. With the shift toward providing increased online learning platforms, combined with the pressure to reduce administrative financial costs, the applied exam offers numerous logistic benefits for a psychology subject that has historically only been able to test rote learning. Restructuring learning and assessment toward applied knowledge is particularly crucial for students who need to demonstrate critical thinking skills which are highly valued in their future job market. Therefore it is imperative for psychology lecturers to bring their statistics and research methods courses up to the standard that other disciplines are already using, and by embracing multiple-choice exams that test applied knowledge we leverage trending digital pedagogical practices that actually foster deeper and longer-lasting learning. By creating a reliable applied knowledge exam in the digital non-proctored space, our findings benefit organisations that run psychology programs, in reducing administration costs; psychology instructors, by way of automated marking; and psychology students themselves, by reducing the anxiety caused by closed-book invigilated exams.

DATA ACCESSIBILITY STATEMENT

The dataset used in this research has been provided and is available for repository hosting.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

First Author conceived and designed the research project, collected these data, conducted and interpreted the analyses. Second Author conducted the literature review and drafted the prose for the overall manuscript. Both authors revised the manuscript for critical intellectual content.

AUTHOR AFFILIATIONS

Sarven Savia McLinton orcid.org/0000-0001-9125-3860
University of South Australia, AU
Sharon Elizabeth Wells orcid.org/0000-0003-0979-716X
University of South Australia, AU

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & **McDermott, K. B.** (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. DOI: https://doi.org/10.1002/acp.1391

Anderson, L. W., & **Krathwohl, D. R.** (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Biggs, J. B. (2003). *Teaching for quality learning at university*. Buckingham: Open University Press/Society for Research into Higher Education. (Second edition).

Carnegie Mellon University. (2023). *Open Learning Initiative*. Carnegie Mellon. https://oli.cmu.edu

Daffin, L. W., & **Jones, A. A.** (2018). Comparing student performance on proctored and non-proctored exams in online psychology courses. *Online Learning*, *22*(1), 131–145.

Duffy, C., & **Sas, N.** (2020). Australia's university sector hit by job losses, fall in international students and Federal Government reform — so what's next? *ABC News*. https://amp.abc.net.au/article/12654828

Forsey, M., Low, M., & **Glance, D.** (2013). Flipping the sociology classroom: Towards a practice of online pedagogy. *Journal of Sociology (Melbourne, Vic.)*, *49*(4), 471–485. DOI: https://doi.org/10.1177/1440783313504059

Goedl, P. A., & **Malla, G. B.** (2020). A Study of Grade Equivalency between Proctored and Unproctored Exams in Distance Education. *American Journal of Distance Education*, *34*(4), 280–289. DOI: https://doi.org/10.1080/08923647.2020.1796376

Griffith, J. D., Adams, L. T., Gu, L. L., Hart, C. L., & **Nichols-Whitehead, P.** (2012). Students' attitudes toward statistics across the disciplines: A mixed-methods approach. *Statistics Education Research Journal*, *11*(2), 45. DOI: https://doi.org/10.52041/serj.v11i2.328

Harmon, O. R., & **Lambrinos, J.** (2010). Are online exams an invitation to cheat? *The Journal of Economic Education*, *39*(2), 116–125. DOI: https://doi.org/10.3200/JECE.39.2.116-125

Hift, R. J. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education*, *14*(1), 249–249. DOI: https://doi.org/10.1186/s12909-014-0249-2

Jensen, S. A. (2011). In-class versus online video lectures. *Teaching of Psychology*, *38*(4), 298–302. DOI: https://doi.org/10.1177/0098628311421336

Johanns, B., Dinkens, A., & **Moore, J.** (2017). A systematic review comparing open-book and closed-book examinations: Evaluating effects on development of critical thinking skills. *Nurse Education in Practice*, *27*, 89–94. DOI: https://doi.org/10.1016/j.nepr.2017.08.018

Lukhele, R., Thissen, D., & **Wainer, H.** (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. DOI: https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. DOI: https://doi.org/10.3102/0013189X018002005

Milone, A. S., Cortese, A. M., Balestrieri, R. L., & **Pittenger, A. L.** (2017). The impact of proctored online exams on the educational experience. *Currents in Pharmacy Teaching and Learning*, *9*(1), 108–114. DOI: https://doi.org/10.1016/j.cptl.2016.08.037

Myyry, L., & **Joutsenvirta, T.** (2015). Open-book, open-web online examinations: Developing examination practices to support university students' learning and self-efficacy. *Active Learning in Higher Education*, *16*(2), 119–132. DOI: https://doi.org/10.1177/1469787415574053

Nielsen, P. L., Bean, N. W., & **Larsen, R. A. A.** (2018). The impact of a flipped classroom model of learning on a large undergraduate statistics class. *Statistics Education Research Journal*, *17*(1), 121. https://link.gale.com/apps/doc/A541103811/AONE?u=unisa&sid=AONE&xid=90bb761e. DOI: https://doi.org/10.52041/serj.v17i1.179

Onwuegbuzie, A. J., & **Wilson, V. A.** (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195–209. DOI: https://doi.org/10.1080/1356251032000052447

Richardson, V. (2003). Constructivist Pedagogy. *Teachers College Record*, *105*(9), 1623–40. DOI: https://doi.org/10.1046/j.1467-9620.2003.00303.x

Schmidt, S. M. P., Ralph, D. L., & **Buskirk, B.** (2009). Utilizing online exams: A case study. *Journal of College Teaching and Learning*, *6*(8), 1–8. DOI: https://doi.org/10.19030/tlc.v6i8.1108

Stowell, J. R. (2015). Online open-book testing in face-to-face classes. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 7–13. DOI: https://doi.org/10.1037/stl0000014

Tu, W., & **Snyder, M.** (2017). Developing conceptual understanding in a statistics course: Merrill's First Principles and real data at work. *Educational Technology Research and Development*, *65*, 579–595. http://www.jstor.org/stable/45018569. DOI: https://doi.org/10.1007/s11423-016-9482-1

Theophilides, C., & **Koutselini, M.** (2000). Study Behavior in the closed-book and the open-book examination: A comparative analysis. *Educational Research and Evaluation*, *6*(4), 379–393. DOI: https://doi.org/10.1076/edre.6.4.379.6932

Vo, H. M., Zhu, C., & **Diep, N. A.** (2017). The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies in Educational Evaluation*, *53*, 17–28. DOI: https://doi.org/10.1016/j.stueduc.2017.01.002

Williams, J. B., & **Wong, A.** (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology*, *40*(2), 227–236. DOI: https://doi.org/10.1111/j.1467-8535.2008.00929.x

Winquist, J. R., & **Carlson, K. A.** (2014). Flipped statistics class results: Better performance than lecture over one year later. *Journal of Statistics Education*, *22*(3), 1–10. DOI: https://doi.org/10.1080/10691898.2014.11889717

TO CITE THIS ARTICLE:
McLinton, S. S., & Wells, S. E. (2023). Assessing Psychology Student Applied Knowledge of Statistics via Open-book Multiple Choice Online Exams. *Designs for Learning*, 15(1), 58–69. DOI: https://doi.org/10.16993/dfl.211

Submitted: 01 April 2022 **Accepted:** 03 August 2023 **Published:** 16 August 2023