

Visualising and evaluating learning/achievement consistency in introductory statistics

Taryn Axelsen, Rachel King & Elizabeth Curtis

To cite this article: Taryn Axelsen, Rachel King & Elizabeth Curtis (2025) Visualising and evaluating learning/achievement consistency in introductory statistics, Cogent Education, 12:1, 2492727, DOI: [10.1080/2331186X.2025.2492727](https://doi.org/10.1080/2331186X.2025.2492727)

To link to this article: <https://doi.org/10.1080/2331186X.2025.2492727>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 18 Apr 2025.



[Submit your article to this journal](#)



[View related articles](#)



[View Crossmark data](#)

Visualising and evaluating learning/achievement consistency in introductory statistics

Taryn Axelsen^a , Rachel King^a  and Elizabeth Curtis^b 

^aSchool of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Queensland, Australia;

^bSchool of Education, University of Southern Queensland, Toowoomba, Queensland, Australia

ABSTRACT

In tertiary education, assessment plays a critical role in shaping student engagement and measuring learning outcomes. In introductory statistics courses, understanding earlier material is essential for later topics, necessitating consistent engagement to avoid fragmented learning. Assessment influences motivation and the depth of conceptual understanding upon course completion. Traditional methods such as cumulative grading and learning analytics often fail to capture the complexity of student knowledge. This research employed a multi-layered approach, including innovative 'consistency of learning', 'combination analysis' and 'heatmap' techniques, to examine performance across 11 learning modules. Results showed that Pass-grade (50–64%) students often did not complete key modules adequately, resulting in fragmented understanding. The study highlighted the limitations of traditional evaluation methods in capturing the complexity and variability of student knowledge. It further emphasized the importance of thoughtful assessment design to ensure that students developed a cohesive understanding of the material regardless of the grade level they achieve. Given the increasing importance of statistical literacy in today's data-centric society, it is vital to equip students with the knowledge to make informed data decisions. By integrating these novel evaluation methods, educators can better understand and support student achievement and improve learning outcomes in introductory statistics.

ARTICLE HISTORY

Received 20 November 2024
Revised 2 March 2025
Accepted 8 April 2025

KEYWORDS



Assessment; statistics education; consistency of learning; combination analysis

SUBJECTS

Statistics & Probability;
Assessment; Higher Education

Introduction

Tertiary introductory statistics courses (units) play a vital role in developing quantitative reasoning and statistical literacy skills across a spectrum of disciplines, including psychology, biology, medicine, computer science, physics, chemistry, agriculture, accounting, commerce, and education (Gould, 2017). The necessity for graduates in these fields to possess at least an introductory level proficiency in statistics underscores the imperative for educators to ensure that students acquire a cohesive body of knowledge in these courses, thereby providing a robust foundation for statistical understanding. Inadequate statistical literacy can have considerable societal and professional consequences, particularly given the increasing reliance on data-driven decision-making across various fields. As data becomes increasingly integral to various aspects of life, the ability to understand and use statistical information is imperative for informed citizenship and professional competence (Wolff et al., 2016). Despite its importance, statistics often faces a negative perception among those students not engaged specifically with a mathematics and statistics program (i.e. non-specialists), with the abstract nature of its concepts and need for at least some mathematical working, posing a significant challenge and inducing high anxiety for many (Bromage et al., 2022; Gordon & Nicholas, 2010; Sutter et al., 2024).

CONTACT Taryn Axelsen  Taryn.axelsen@unisq.edu.au  School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, Queensland, Australia

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

There have been many changes in tertiary statistical education in the last 20 years, partly driven by reforms in mathematics education at the high school level (Shimizu & Vithal, 2023) and partly as a result of follow-on reforms flowing into tertiary teaching (Legacy et al., 2024). However, as stated by Shimizu and Vithal (2023), this is still a very under-researched topic in the literature. Recently, Galligan et al. (2020) explored the evolution of tertiary statistics education within Australia from 2016 to 2019, finding that many researchers in this period explored software and technology within statistical education (Dunn et al., 2017; Muir et al., 2020). The evolving landscape in tertiary statistical education has contributed to the challenging task of designing courses and assessments that promote students' engagement with and learning of statistics. Additionally, the rapidly changing global events of the 2020–2022 COVID-19 pandemic precipitated rapid shifts in how universities conduct assessments, necessitating a significant move toward online assessment (Sato et al., 2023). Adapting to these new demands while upholding academic standards requires a critical review and revitalization of assessment approaches moving forward.

Subsequently, assessment is a central feature of developing an engaging and cohesive curriculum (Raffaghelli & Grion, 2023). It can be used as a powerful method for motivating and assisting student learning and is one of the most significant influences on students' experience in the tertiary sector (Nieminen, 2025; Vaesse et al., 2017). Furthermore, it can assess if the learning goals of a course are met and can be used to help inform student success and retention (Mitra, 2023).

Recent existing literature in higher education assessment evaluation focuses on the evaluation and adaptability of assessments in the rapidly changing global landscape (Zeng, 2025), relationship with student learning and how it influences students welfare (Fischer et al., 2024), using data to implement assessment transformation (Kaspi & Venkatraman, 2023; Shivshankar & Acharya, 2024) and more recently on artificial intelligence (AI) (Crompton & Burke, 2023). The rapid increase of AI has created AI-driven techniques to promote educational assessments (Sánchez-Prieto et al., 2020), decrease educators workloads, promote adaptive learning (Ouyang & Jiao, 2021), provide personalized feedback to students (Vashishth et al., 2024) and enhance the learning journey for students while transforming the education system (Ouyang et al., 2023).

Nevertheless, existing assessment evaluation methods in higher education tiered grading systems often fail to capture learning inconsistencies among students, particularly those achieving a pass grade (gaining 50–64% overall), where a student can achieve a passing grade by demonstrating as little as 50% of the course material being assessed. Traditional assessments within higher education tend to focus on cumulative scores rather than a detailed analysis of understanding across different topics (Cain et al., 2022) and often fail to account for the complexity and variability of student understanding. For instance, Bloom's taxonomy (Bloom, 1956), while useful, may not fully capture the dynamic and personalized nature of modern learning environments (Alafnan, 2024).

Although there is growing literature towards evaluating authentic assessment (Ajjawi et al., 2024; Hamel & Lee, 2024; Schultz et al., 2022) and evolving technologies in assessment (artificial intelligence, learning analytics and extended reality technologies), for both design and analysis (Sembey et al., 2024), these methods often do not focus on capturing learning inconsistencies. Sembey et al. (2024) further highlighted a lack of research into the evaluation of the effectiveness of technology use within assessment, with Matcha et al. (2019) emphasizing the importance of research considering not only overall assessment scores but includes qualitative analysis. This gap in the evaluation of assessment practices means that students may pass a course without a comprehensive grasp of the core statistical principles.

The research gap is evident in the limitations of the traditional grading systems in capturing knowledge consistency. This study investigated the distribution of marks in a first-year tertiary statistics course with a focus on those students achieving a Pass-grade (50–64% overall) to explore how achievement is accumulated across core topics. Furthermore, we assessed the reliability of using overall grades and assessment analytics as a measure of student understanding, particularly examining whether a Pass-grade genuinely reflects a comprehensive grasp of fundamental statistical concepts upon course completion. Our innovative consistency of learning/achievement approach using combination analysis and heatmaps for visualisation, has enabled the identification of gaps in student learning and ensured a more robust evaluation of student competencies across all topic areas. Through a case study in a tertiary introductory statistics course, we demonstrate this approach and suggest that students who achieve a

Pass-grade may not necessarily have a thorough understanding of all fundamental statistical principles if the course structure allows for varied paths to accumulate partial marks. This novel approach aims to provide deeper insights into student learning and achievement consistency.

Re-evaluating assessment in tertiary statistics education

Undergraduate university students enrolled in an introductory statistics unit are generally asked to explore a rich and diverse landscape of ideas, many of which require a strong understanding of core concepts before more complicated theories and models are attempted. Assessment can play an integral role in helping students gain a genuine understanding of these core concepts (Henning & Roberts, 2023; Hodgson & Pang, 2012), enabling their success and continued progression through the unit's curriculum. Research has found that the way assessment is designed can support or hinder student understanding by impacting student attitudes (Harsy, 2020; Nieminen, 2025), their study habits (Collins et al., 2019), how they learn and retain information (Murphy et al., 2023), and the way they engage with the learning materials (Holmes, 2018). Many of these topics fall under the broader scope of formative assessment theory, which emphasizes assessment as a tool to support student learning and development (Cizek, 2010; Rutherford et al., 2025; Wiliam, 2010; Winstone & Boud, 2022). Formative assessments provide continuous feedback, allowing students to monitor their progress and adjust their strategies for success. By contrast, summative assessment, which is the main exploration of this study, traditionally focuses on evaluating final learning outcomes, sometimes lacking the feedback loop.

While assessment is a necessary part of tertiary education, being used to assign grades to indicate the quality of student achievements, provide evidence or certification to external partners (Winstone & Boud, 2022), and support student learning (Carless, 2015), the way in which that assessment is structured can also greatly impact student outcomes (Collins et al., 2019). The ways in which students are evaluated are among of the most salient classroom factors that can affect their motivation (Ames, 1992; Ferland et al., 2024); indeed assessment has been identified as perhaps the most important factor for student attitudes (Harsy, 2020), motivation and engagement in learning (Hansen & Ringdal, 2018). A particular challenge for educators is, therefore, finding an assessment design that facilitates engagement (Ahshan, 2021; Holmes, 2018) and the long-term retention of learning (Murphy et al., 2023), while at the same time improving and promoting genuine student understanding of core course concepts, and communicating feedback to students on their understanding of these same course concepts (Morris et al., 2021).

Shifting the focus of statistics assessment from merely controlling standards and certification to fostering genuine understanding and learning involves re-examining how assessments themselves are evaluated. This requires a movement beyond traditional testing to methodologies that promote sustainable student learning through continuous analysis and improvement of the assessment process.

Despite overwhelming research that assessment design plays an integral role in supporting student understanding of statistics, many assessment structures traditionally used in mathematics-based courses, including statistics, often do not effectively scaffold learning. This can lead to fragmented understanding, where students grasp isolated concepts but struggle to integrate them into a coherent framework of statistical knowledge, resulting in a lack of genuine comprehension of fundamental ideas. Tallman et al. (2016) found, in a review of university mathematics-based assessment design, that little has changed in the previous twenty-five years regarding assessment in statistics courses, with few structures requiring a genuine conceptual understanding and most still requiring students to recall and apply a rehearsed procedure. Traditional summative assessment methods, involving a cumulative of marks, often assess students on discrete topics in isolation rather than evaluating their ability to synthesize concepts across multiple modules (Maki, 2023). Consequently, students may demonstrate competency in specific areas but lack consistency in understanding across the entire curriculum. While traditional assessment methods such as this still exist there is a growing interest in more innovative, inclusive and accessible approaches (Biehler et al., 2024).

There is a significant amount of research focusing on evaluating the quality and complexity of assessment, generally on one type of assessment only (usually examinations) (Garfield et al., 2011; Marriott et al., 2009) or on the overall assessment framework (Huber et al., 2024) and structure (Dierker et al.,

2018). For example, Dunham et al. (2015) explored the use of Bloom's taxonomy (Bloom, 1956) as a tool to evaluate the complexity of assessment tasks in statistics examinations in comparison with other subjects and as an attempt to align assessment tasks on the taxonomy's scale. Dierker et al. (2018) looked at comparing different assessment structures from a student learning and student experience perspective, finding that project-based courses in statistics had higher percentages of students demonstrating critical understanding of higher-order statistical concepts.

The use of analytics to evaluate assessment, and using these to guide higher education assessment design, is becoming increasingly important and has been used to address the need for increased transparency (Raković et al., 2023). This is particularly important in the university sector which faced major changes to delivery of assessment, often with fewer resources, during the COVID-19 global pandemic. One consequence was the move from face-to-face invigilated assessment to online assessment. Due to the high demands of all aspects of change that occurred during this time, many universities did not change their assessment structure when moving from face-to-face to online examinations (Gamage et al., 2020; Slade et al., 2022); only the mode of delivery changed. With the even more recent explosion of Artificial Intelligence (AI) tools there is a recent focus on online assessment, academic integrity, and the evolving landscape of digital learning environments (Kolade et al., 2024; Perkins, 2023; Sato et al., 2023).

Even where assessment models might be structurally sound, both in quality and complexity, this does not necessarily mean that students who obtain a pass grade (usually 50%–64%) have a cohesive set of statistical knowledge on exit, nor an appreciation for all forms of fundamental ideas presented within the course. The phenomenon of fragmented understanding means that students may succeed in isolated assessments without truly integrating knowledge across the course. Of importance is being able to effectively evaluate assessments to satisfactorily identify this fragmented learning. Students who acquire a cohesive set of knowledge are more likely to be able to identify what it is they have learnt in the course and how they have achieved a pass grade (Pullen et al., 2018). Students who complete a course with a clear sense of what they have learnt and achieved are more likely to develop a positive attitude towards statistics, reduce negative perceptions, and potentially contribute to higher overall statistical literacy in the population (Noraidah et al., 2011). Understanding why some students do not achieve a cohesive level of understanding within a course can help in designing instructional and assessment strategies that prevent fragmented understanding and make learning statistics more enjoyable.

Research questions

This study commenced with concerns regarding course assessment and the adequacy of student learning, specifically regarding core topics that may not be comprehensively mastered by all students. This led to the two main objectives of this study to show that traditional assessment structures may not give students a cohesive set of fundamental introductory statistical knowledge on course exit.

1. How do traditional assessment evaluation methods, such as grade distribution, item analysis, and learning analytics, represent diversity in student knowledge, particularly for students achieving a pass grade?
2. How can an alternative evaluation approach provide a more comprehensive understanding of student performance across diverse knowledge levels?

Materials and methods

Study data

This study centred around a large first-year tertiary statistics course offered at an Australian university in three teaching periods annually (TP1, TP2, and TP3). It was taught predominantly online and had a 50% final assessment as an in-person supervised examination. Students enrolled in this course came from a variety of different undergraduate programs in the fields of Business/Commerce, Science, Education, Psychology, Health, and Information Technology (IT). The course curriculum aligned with typical content

covered in introductory tertiary statistics courses at many universities and was organized into 11 modules (M1 to M11), covering topics such as basic concepts of statistics, exploratory data analysis, probability distributions, regression, confidence intervals, and hypothesis testing. Prior to 2020, the course comprised four main summative assessment components: three assignments (with the first being a low-weighted activity to encourage early engagement) and a final invigilated examination, which included 20 multiple-choice questions and five to seven short-answer questions (S1 Table). In 2020, in response to the global pandemic, the invigilated examination was replaced with an online, non-invigilated version of the examination.

Each summative assessment item's weighted scores were aggregated to calculate a final overall score out of 100. Students were subsequently assigned a final grade based on predetermined cut-offs: High Distinction (HD) for scores ranging from 85–100%; Distinction (D) for scores between 75 and 84.4%; Credit (C) for scores between 65 and 74.4%; and Pass (P) for scores between 50 and 64.4%. Grades below 50% were designated as Fail (F).

For the sake of consistency, data was collected from TP2 spanning three years: 2018, 2019, and 2020. TP2 data for 2018 included 279 students who successfully graduated from the course with complete assessment data, with 224 and 253 students, respectively, in 2019 and 2020. These specific years were chosen as this comprehensive assessment of the data was used to guide a new curriculum design, including a new assessment design, which was due to be implemented in 2022 (with trials in 2021). This data collection would then allow student performance before and after the rollout to be compared. TP2 was selected due to its larger cohort size, ensuring a more robust analysis and reliable insights.

Approval to conduct this study was gained from the University of Southern Queensland Research Ethics Committee (ETH2022-0065). The collection of retrospective assessment data was de-identified at the point of collection.

Descriptive analysis

This research was a repeated cross-sectional design (Almond & Sinharay, 2012; Wang & Cheng, 2020) to analyse assessment data. The goal was to identify and explore patterns of student performance and knowledge diversity. For the first research objective, traditional methods were applied to analyse the assessment data. With the rise of large educational datasets, using this data for informed decision-making has become essential. Recently, most analysis methods have focused on learning analytics and traditional statistical techniques for exploratory analysis (Caspari-Sadeghi, 2023).

The data included the overall percentage score, grade category, mark out of 20 for the multiple-choice section of the examination, all part-marks from the exam short answer questions and the questions in the three assignments. Course markers were responsible for assigning marks, using a detailed marking guide, for each question within assignments and exam short-answer questions, and all marks were rigorously moderated. Overall scores and grades for each student receiving an HD, D, C, or P were collated, as well as a detailed breakdown of the part-marks allocated for each of their individual assessment items. The availability of part-marks for all assessment items allowed for a comprehensive view of the specific aspects of the course achieved by students at different grade levels, extending beyond their overall grade results. These part-marks were then mapped to each module (M1–M11) so that each student's achievement by module could be assessed. All statistical analysis was performed using the R statistical software (Posit Team, 2023; R Core Team, 2023).

The initial assessment evaluation focused on item analysis (Kehoe, 1994; Mukherjee & Lahiri, 2015), of the multiple-choice questions from the examination within each teaching period, following established indicators such as the difficulty/facility index (FI), which assessed the proportion of students answering an item correctly; the discrimination index (DI), which gauged the item's effectiveness in distinguishing between high and low scorers; and the discrimination efficiency (DE), which measured the percentage of attempts needed to estimate DI relative to question difficulty. While this method provided initial insights into examination-specific performance, it evaluated only one aspect of assessment data. A more comprehensive evaluation that included all types of assessment, not just examination data and final grades, would offer a more complete picture of student learning and performance.

To gain a broader understanding of student achievement and to highlight the inconsistent fundamental statistical knowledge, descriptive statistics were calculated for all teaching periods and modules (topic areas) to examine mark distributions for each student and to identify how many modules were not successfully passed. To ensure that patterns identified were not due to significant differences related to specific teaching periods or teaching teams, a one-way ANOVA and a chi-square test of independence were conducted to identify any significant differences between teaching periods in terms of assessment items, grades, and the number of students passing each module. To further the characteristics of Pass-grade students, Fisher's exact test of independence and regression analysis were used to examine the landscape of Pass-grade students, allowing for a more nuanced understanding of factors influencing student success. The combination of these statistical methods provided an initial framework for evaluating assessment outcomes, to try to capture diversity in student knowledge performance.

However, these traditional methods alone did not provide sufficient granularity in identifying fragmented knowledge, particularly among students achieving a Pass-grade. While overall grades and aggregated scores reflected broad performance trends, they obscured finer details about student strength and weaknesses in specific modules. This limitation necessitated the development of an alternative evaluation approach that mapped part-marks to individual modules, enabling a more detailed examination of student learning.

Combination analysis, consistency of learning, heatmaps

Subsequently, this study introduced a novel approach to examine student achievement of all assessment data (not just examination results) through a method the researchers termed *Consistency of Learning*. This approach utilized a *combination analysis* to gain deeper insights into student performance across various modules for all combined assessment data, an analysis made possible by the comprehensive breakdown of all assessments items (not only examination results), including partial marks.

For each student, an 11-digit binary code was constructed, where each digit represents one of the 11 modules in order from 1 to 11. A digit was coded as '1' if the student sufficiently completed the module (gained over 45%) or a '0' if they did not. This binary sequence effectively encapsulated the student's achievement profile across all modules. For example, a binary code of 1111111110 indicated that a student sufficiently completed modules 1 to 10 (gaining over 45%) but did not gain sufficient knowledge in module 11. By analysing these binary codes, the degree of coherence in the knowledge obtained by each student could be fully assessed.

For each module we considered an achievement of over 45% as 'sufficient completion' to allow for some variance in grading. A mark of 45% or above was used based on several factors. A common benchmark when using criterion-referenced assessment in Australia is that student must achieve an overall mark of at least 50% (Sadler, 2005) to pass the course., A score marginally below this is often considered a borderline or minimally competent pass (Shulruf et al., 2015), sometimes called a conceded pass (45–50%) (Sadler, 2005). McKinley and Norcini (2014) found that there is no best method to determine a passing score for performance-based tests and suggested selecting a method within the resources available. Thus, setting the conservative cut off score of 45% and above to constitute sufficient knowledge per module allowed for some leniency in different marking styles, highlighted the complexities of grading practices and accounted for the perceived difficulty of some modules compared others, while ensuring that students have grasped the essential concepts.

This combination analysis provided a distribution of student achievement by module, identifying the extent to which students achieved a grade based on their performance across different aspects of the course. This methodology provided a nuanced perspective on student learning outcomes, allowing educators to ascertain the degree to which a Pass-grade reflected a cohesive set of knowledge for the cohort of Pass-grade students. The aim for the combination analysis was to enhance the understanding of student achievement patterns and improve assessment processes within educational settings. Integrating module-based evaluation of assessment with existing statistical methods, allowed a richer understanding of student knowledge diversity and could be used to reveal patterns that traditional methods could not easily detect.

The term *consistency of learning* referred to a metric derived from combination analysis to measure the extent to which students consistently achieved grades based on specific combinations of passed modules. This metric was calculated by examining the frequency with which students passed identical sets of modules. A higher frequency of students passing the same combination of modules indicated a higher consistency of learning among each cohort.

Heatmaps were also used to provide a visual representation of the combination analysis using colour gradients. This visualization technique was invaluable for identifying patterns, trends, and relationships in the data that may not be immediately apparent through numerical analysis alone (Gu, 2022). By representing data points with varying intensities of colour, heatmaps enabled a quick and intuitive understanding of complex data sets. The creation of these heatmaps was facilitated using the `heatmap.2` function from the `gplots` package (Warnes et al., 2022) and the `ggplot2` package (Wickham, 2016) within R (R Core Team, 2023). This function allowed the generation of heatmaps with dendrograms, which cluster points based on similarity. This clustering helped in identifying groups of students who exhibited similar patterns of module pass/fail combinations, thereby revealing deeper insights into student learning and module interrelationships. The dendrograms added another layer of information by showing hierarchical relationships among the data points. In this analysis these allowed us to identify which modules were frequently passed together and which combinations were less common, and allowed understanding of how certain modules may be interdependent in terms of student performance. In this analysis, these dendrograms helped in identifying which modules were frequently passed together and which combinations were less common, allowing for an understanding of how certain modules may have been interdependent in terms of student performance.

A detailed stepwise explanation of data collection, coding, and analysis process was provided in [S2 Table \(Supplementary Material\)](#) to enhance reproducibility.

Results

In TP2 2018, 87 students achieved a Pass-grade (52 students in TP2 2019 and 76 students in TP2 2020). A chi-square test of independence showed no association between the number of students in grade categories (P, C, D and HD) and year (2018, 2019 and 2020) ($\chi^2(6) = 5.12, p = 0.53$), indicating there was no significant difference in student performance among years, even given the change in the final assessment item in 2020 due to the global pandemic. ANOVA analysis also showed that there was no significant difference between years for the overall grade (out of 100) for all students obtaining a Pass-grade or higher ($N = 756$) ($F(2,753) = 1.98, p = 0.14$) and the short-answer question section of the examination ($F(2,753) = 2.14, p = 0.12$). The multiple-choice examination mark ($F(2,753) = 34.02, p < 0.0001$) was significantly different between years, with 2020 (online and non-invigilated examination) having a higher average multiple-choice question examination mark than previous years, however the overall average examination mark was not different between years ($F(2,753) = 2.96, p = 0.052$), suggesting that the multiple-choice component may have influenced scores in 2020 but did not translate into overall grade inflation.

Item analysis evaluation of the multiple-choice question in the examination (Table 1) determined that there did not appear to be any concerns for the quality of the questions. However, several questions obtained a poor DI, indicating poor discrimination between some options (each question had 5 options). This analysis also determined that the year 2020 (impacted by COVID-19 and the move to online examinations) had a significantly higher FI value than 2018 and 2019 ($F(2,57) = 3.24, p = 0.047$), however the DI ($F(2,57) = 0.13, p = 0.88$) and DE ($F(2,57) = 0.16, p = 0.86$) were consistent between years.

Table 1. Item analysis for the multiple-choice examination questions.

Year	FI %	Mean (SD)	
		DI %	DE %
2018	56.88 (19.83)	32.20 (11.70)	42.74 (15.72)
2019	58.94 (15.81)	33.98 (7.97)	43.75 (10.50)
2020	70.34 (18.21)	33.01 (12.75)	45.19 (14.98)

Even though different assessment questions were designed for each teaching period, all had to align with the course specifications which outlined the weighting contribution of each module. Assessment for each teaching period was within $\pm 2\%$ of this outline (M1: 8%, M2: 6%, M3: 8%, M4: 12%, M5: 12%, M6: 10%, M7: 6%, M8: 14%, M9: 8%, M10: 8%, M11: 8%).

Descriptive analysis

The exploration of descriptive statistics provided valuable insights into the performance of students across different modules. Table 2 showed the average scores (and standard deviations), for each of the 11 modules (M1 to M11) for those students obtaining a Pass-grade (P) and gave a breakdown of the percentage (%) of Pass-grade students who scored more than 45% in each module. This threshold ($>45\%$) is considered indicative of 'sufficient knowledge' (sufficiently completed) for that specific module. While a high proportion of Pass-grade students had sufficient knowledge in Module 1 (97.7%), this knowledge declined significantly for Module 10 (40.9%) and Module 11 (32.1%). This pattern suggests that while the initial five modules are generally completed to a satisfactory standard, there is a significant decline in performance in the subsequent modules for Pass-grade students, possibly due to increasing content complexity or cumulative learning challenges.

To provide a visual perspective Figure 1 illustrated the number of 'insufficiently completed' modules (achieving less than 45%) relative to the overall mark for each grade level. Due to module weightings, students were able to excel in one module while encountering difficulties in several others and still achieve a Pass-grade. This lack of attainment of a cohesive set of knowledge across the course was particularly evident among Pass students, with the possibility of insufficiently completing up to 7 modules while still passing the course [HD ($M = 0.6$ ($SD = 0.24$)), D ($M = 0.42$ ($SD = 0.58$)), C ($M = 1.36$ ($SD = 1.02$)), P ($M = 3.42$ ($SD = 1.31$))].

Table 3 revealed further insights into the performance of passing students compared with students in other grade categories. As expected, the number of students insufficiently completing each module (achieving less than 45% marks) was notably higher for Pass-grade students compared with their peers who earned higher final grades. This difference was statistically significant for each module ($p < 0.05$).

Percentage and number of Pass students 'insufficiently completing' each module in comparison with students achieving all other grades (Other Students).

The findings highlighted that Pass-grade students encountered substantial difficulties in maintaining adequate performance across a range of modules. These challenges were particularly pronounced in Modules 6, 7, 8, 9, 10, and 11, where Pass students had significantly higher insufficiently completed rates compared with other students. This pattern suggested that certain modules, likely characterized by increased complexity or a higher degree of cumulative knowledge, seem to pose greater hurdles for Pass-grade students.

Table 2. Percentage of Pass-grade students achieving $>45\%$ in each module.

	Percentage of pass-grade students achieving $>45\%$ [$N=(215)$]	Mean (%) mark (SD) pass-grade students achieved in each module
Module 1	97.7%	78.7% (13.5%)
Module 2	94.9%	71.2% (14.7%)
Module 3	75.3%	63.4 % (23.0%)
Module 4	80.5%	57.1% (13.5%)
Module 5	97.2%	72.9% (13.6%)
Module 6	53.5%	48.1% (17.4%)
Module 7	50.2%	48.7% (29.3%)
Module 8	54.0%	47.8% (16.6%)
Module 9	80.0%	59.9% (19.4%)
Module 10	40.9%	40.6% (18.8%)
Module 11	29.8%	34.6% (26.9%)

This table showed those % of Pass-grade students (overall grade between 50–64%), who gained more than 45% for each of the 11 modules and the mean and standard deviation (SD) for each module for this cohort of students. The bold grey cells emphasized those modules where the average of Pass-grade students for this module was below 50%.

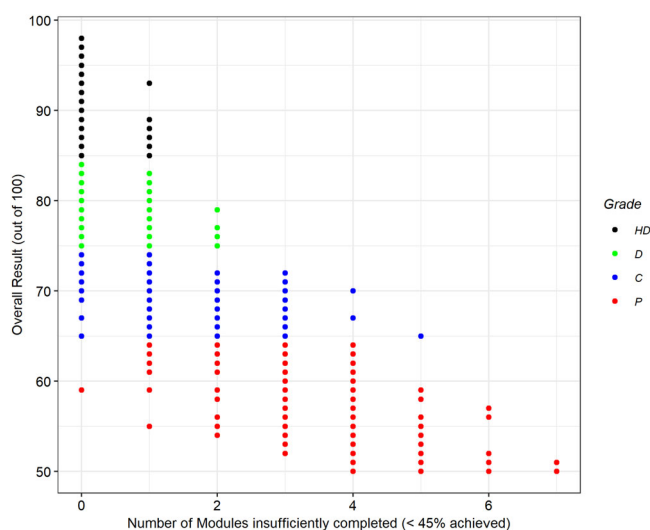


Figure 1. Number of modules 'insufficiently completed' (Achieved < 45%) by Grade Level (HD, D, C, P). The relationship between the number of modules insufficiently completed (where students achieved less than 45% for that module) and the overall final mark across different grade levels: High Distinction (HD), Distinction (D), Credit (C), Pass (P).

Table 3. Number (and %) of students 'insufficiently completing' each module.

	% Pass-grade Students achieving <45%	% Other Students achieving <45%	p-value
Module 1	2.3% (n = 5)	0.4% (n = 2)	0.023*
Module 2	5.1% (n = 11)	0.7% (n = 4)	<0.0001**
Module 3	24.5% (n = 53)	5.9% (n = 32)	<0.0001**
Module 4	19.4% (n = 42)	2.8% (n = 15)	<0.0001**
Module 5	2.8% (n = 6)	0.5% (n = 3)	0.019*
Module 6	46.3% (n = 100)	6.3% (n = 34)	<0.0001**
Module 7	49.5% (n = 107)	15.0% (n = 81)	<0.0001**
Module 8	46.3% (n = 100)	5.7% (n = 31)	<0.0001**
Module 9	19.9% (n = 43)	1.7% (n = 9)	<0.0001**
Module 10	58.8% (n = 127)	6.1% (n = 60)	<0.0001**
Module 11	67.6% (n = 146)	16.5% (n = 89)	<0.0001**

Combination analysis and consistency of learning

While Tables 2 and 3, along with Figure 1, shed light on the inconsistent knowledge achieved by Pass-grade (P) students across different modules, it was also valuable to further explore the combinations of modules that contributed to this issue. It is worth emphasizing, that while High Distinction (HD), Distinction (D), and Credit (C) students also exhibited some variation in knowledge across modules, the potential for variation in attainment was increasingly reduced as overall marks were successfully accumulated.

As shown in Figure 1, some students attained less than 45% in up to seven modules and still passed the course. Among the 203 students who achieved a HD, 191 students successfully completed all 11 modules, resulting in only 6 distinct combinations of binary codes being observed. For students who received a D grade (n = 147), there were 11 unique combinations of module results. As expected, students who received a C grade (n = 191) exhibited greater diversity, with 51 distinct combinations of module results. Notably, the most common combination among Credit students was 11111111111, which was achieved by 40 students, highlighting that a large percentage of Credit grade students sufficiently completed all modules. Of greatest interest were the 215 Pass-grade students, where there were 115 distinct combinations of modules sufficiently completed, emphasizing the wide variety of pathways students took through the assessment and accumulation of marks to attain a Pass-grade (Table 4). Remarkably, only one of these students successfully completed all 11 modules, reflecting the rarity of this achievement within the Pass-grade student cohort.

When more students passed the same combination of modules, the consistency of learning among students has improved. Table 4 showed the consistency of learning/achievement and demonstrated the improved consistency of learning in the higher-grade categories, with the much higher average number

of students per combination reflecting the reduced variability in the way in which students accumulate marks across modules.

This analysis underscored the diversity in students' module completion patterns across different grade levels, shedding light on the range of academic performances within each grade category. Most importantly, it confirmed the inconsistent knowledge demonstrated by Pass-grade students with approximately 53% of students passing with different combinations of module understanding compared with other students in their final grade category.

Heatmaps

To visualise the patterns of inconsistent student knowledge within various grade categories, heatmaps (Figure 2) were produced to illustrate the landscape of inconsistencies. The heatmaps in this study provided a visual representation of how students exhibited disparities in their grasp of knowledge across different modules. They used a calculated Euclidean distance hierarchically clustering matrix (Gu, 2022)

Table 4. Combination analysis and consistency of learning for each grade category.

Grade	Number of distinct binary combinations (<i>n</i>)	Consistency of learning/achievement
HD	6 (<i>n</i> = 203)	34
D	11 (<i>n</i> = 147)	13
C	51 (<i>n</i> = 191)	4
P	115 (<i>n</i> = 215)	2

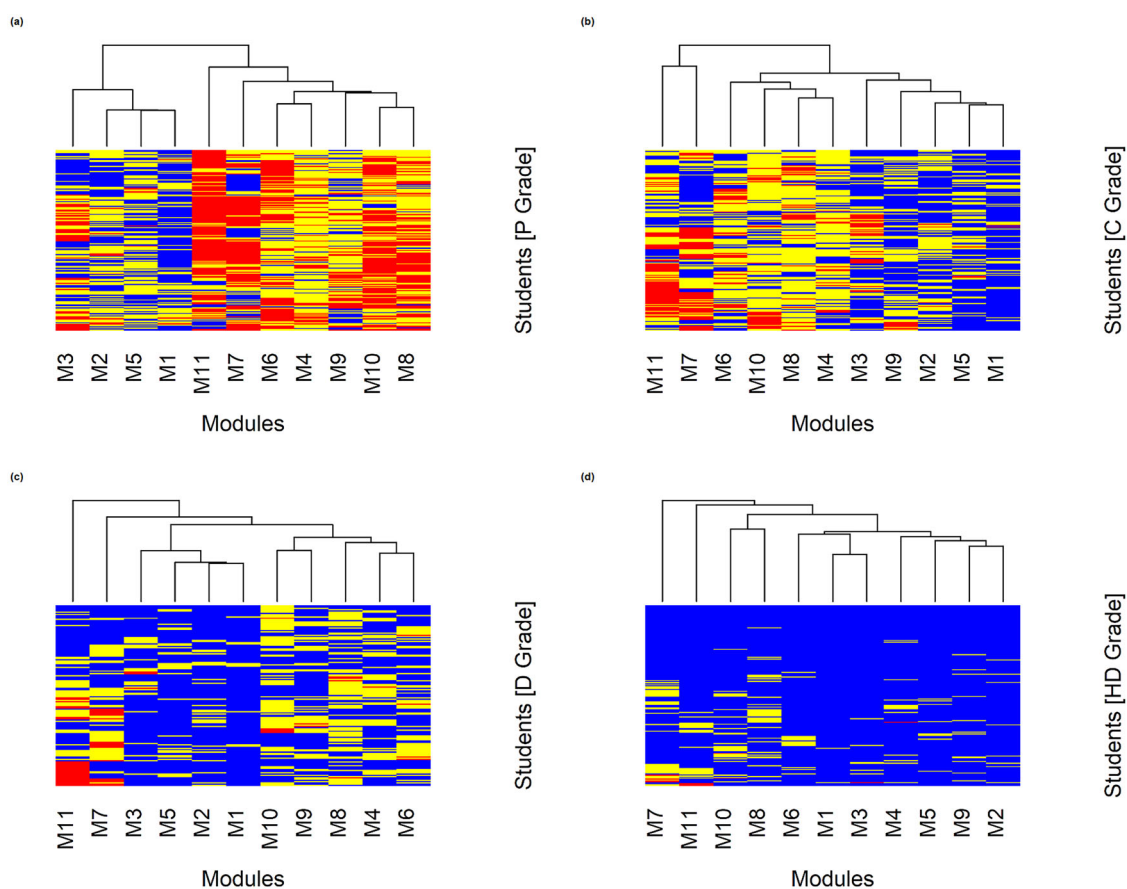


Figure 2. (a-d) Heatmaps showing student inconsistencies per grade. (a) Heatmaps for P Grade, (b) C Grade, (c) D Grade and (d) HD Grade showing the students who failed each module (red), gained a moderate result (yellow) or a high result (blue).

to not only show the knowledge of students but to also show the associations between modules for different grade levels.

The heatmaps were structured with modules on one axis and students on the other. Each cell in the heatmap represented the knowledge level of an individual student in a particular module. The colour gradient within each cell signified the proficiency level: red signified that a student had insufficient knowledge (<45%) in that module; yellow that they gained a moderate result (45%–75%); or blue indicating a high result (>75%). The modules were ordered based on the similarity in scores, indicated by the dendrogram on the top of each heatmap. These dendrograms helped in the understanding of how modules clustered together based on student performance, revealing which modules had similar patterns of knowledge acquisition. This clustering allowed us to see which modules tended to be learned together and how this varied across different grade categories.

In interpreting these heatmaps, we observed distinct patterns of knowledge distribution across different grade categories. For higher grade categories (HD and D), there was a noticeable concentration of blue cells, indicating a higher level of knowledge proficiency across most modules. Conversely lower grade categories (P and C) exhibited more red and yellow cells, indicating areas where students struggled or had weak to moderate understanding. The dendrogram structures highlighted meaningful associations between modules and these also varied across grade categories. For Pass-grade students (Figure 2(a)) the modules clustered together where students performed well were Modules 1, 2, 3 and 5; all early modules before any statistical inference content were introduced. This suggested that these foundational modules, were areas of relative strength for Pass-grade students. However, as content complexity increased, weaker students exhibited more fragmented knowledge indicating difficulties in integrating later concepts. The shifting module associations in higher-achieving students implied that they formed broader connections across topics rather than relying on early module success to pass the course. These findings had pedagogical implications, emphasising the need for better scaffolding and assessment alignment to ensure that essential skills were reinforced and assessed adequately throughout the curriculum, particularly for students at risk of lower performance.

Discussion

In tertiary education, assessments often drive student motivation and their sense of achievement (Ismail et al., 2022; Schneider & Preckel, 2017). Final grades represent the culmination of a student's learning journey and an institutional benchmark for learning outcomes, and thus it is crucial to understand the complete distribution of achievement across course content. This is particularly pronounced in introductory statistics courses, where cumulative knowledge builds sequentially, and student attitudes towards statistics can influence engagement and persistence (Sutter et al., 2024). Research has shown that negative perceptions or early struggles with statistics can lead to students to disengage prematurely, particularly if they perform poorly in early assessments (Bromage et al., 2022).

This study aimed to investigate the knowledge attained by students who achieved a Pass-grade in a tertiary introductory statistics course. By mapping assessment questions to course modules, it was evident that Pass-grade students had a fragmented understanding of topics, lacking a cohesive grasp of entire modules. This fragmentation raises concerns about the extent to which passing students genuinely acquire foundational statistical knowledge, when only examining the final grade. Additionally, these students might leave the course unsure of their learning, unable to consolidate their knowledge, unaware of their gaps, and with an unchanged or diminished perception of statistics. Although Pass-grade students were deemed proficient based on passing the course overall, their understanding might not encompass the necessary breadth of topics.

While final cumulative marks provide a standardized metric, they often do not fully capture the depth of consistency of student understanding (Cain et al., 2022). Although alternative forms of grading exist, such as competency-based assessment (Katoue & Schwinghammer, 2020) and Pass/Fail assessment (Chan, 2023), higher education institutions still regularly use the traditional tiered grading systems (Cain et al., 2022). Traditional grading frameworks often assume that students within the same grade category possess comparable knowledge. However, this study highlighted the student knowledge profiles within the Pass-grade category varied significantly.

By introducing novel approaches (combination analysis, heatmaps and consistency of learning), educators could now visualize and quantify learning fragmentation. Combination analysis advances beyond traditional evaluation methods by shifting the focus from a single cumulative score to an analysis of how well students maintain competency across multiple topics. Heatmaps allow educators to easily visualize and detect patterns of partial mastery that are not readily apparent in traditional assessment evaluations. The consistency of learning uses this combination analysis to allow educators to compare different grade cohorts, or even different teaching periods or subjects. These new methods on assessment evaluation are imperative for educational stakeholders to move beyond static grading methods, fostering meaningful evaluations that align with the long-term goal of statistical education of creating learning environments that foster cohesive learning.

In our case study, the item analyses did not identify any key differences among years or grade categories. Overall, Pass-grade students on average obtained 86.6% in Assignment 1, 74.9% in Assignment 2, 56.6% in Assignment 3, and 52.7% in the multiple-choice section and 45.7% in the short answers section of the final examination. Although these results suggested that Pass-grade students simply did more poorly in the later modules, it did not show which specific modules they struggled with, nor if similarly graded students struggled (or achieved) in the same areas. Some specific examples of Pass-grade students in our case study who demonstrated different aspects of scattered knowledge included:

- Student 1 – Just passed M1 (54%) and M5 (51%), above 80% in M3–M5 and 75% in M11. However, M7 (23%), M8 (11%), M9 (0%) and M1 (11%) were very poor. This indicated that they had little to no knowledge of hypothesis testing and confidence intervals and only minimal knowledge in statistical basics (M1).
- Student 2 – Performed poorly (less than 35%) on M1–M4 and M7. This indicates that some of the basic information was not grasped (graphing, regression, types of variables), which was fundamental to understanding statistics.
- Student 3 – This student scored less than 40% in 7 modules, gaining a pass by scoring highly in M1 (85%) and M5 (83%). They obtained less than 30% in M6, M7, M10 and M11 and less than 40% in M3, M8 and M9. This student does not appear to have a grasp of many topics in introductory statistics.

While exploring the performance of individual students offer valuable insights for small student cohorts, they become impractical for large-scale cohorts. Thus, developing scalable assessment analytics is critical for identifying and addressing learning gaps in diverse student populations.

The second aim of this study was to explore additional forms of analysis for assessment marks. An innovative multifaceted approach was used that included combination analysis, a consistency of learning measure and heatmaps based on assessment-to-module mapping. These methods provided an expanded lens for evaluating learning consistency, demonstrating that performance variability within a grade category was more pronounced than traditional assessment evaluating methods suggested. These results highlighted the heterogeneity and complexity of student knowledge profiles and helped to provide an understanding of student performance that was not clear from standard grade evaluation approaches. Combination analysis quantified this diversity and established an innovative approach to assessing assessments and understanding student knowledge. This research contributes to the broader goal of refining assessment practices to foster deeper, more cohesive understanding among students, ensuring that assessment not only measures but also supports genuine learning and engagement (Biggs et al., 2022; Ibarra-Sáiz et al., 2021). Within the context of summative assessment, this study showed that the same grade could represent very different aspects of learning, and it was possible to gain a deeper understanding of student learning by applying the innovative methods of this study.

The heatmap for Pass-grade students (Figure 2) exhibited diversity in knowledge patterns, with some students excelling in specific modules but struggled in others, while other students demonstrated more balanced proficiency across all modules. This diversity reflected the fragmented understanding among some Pass-grade students, challenging the notion that a Pass-grade signified proficiency across the whole course. The novelty of the heatmap approach lay in its ability to visualise inconsistencies in learning trajectories, allowing educators to identify where students might need additional support and/or allow educators to create more aligned assessment (Biggs et al., 2022) for better learning outcomes. Combination analysis and heatmaps showed that traditional assessment metrics alone do not fully

encapsulate the complexity of student understanding on a granular level. Addressing these disparities through aligned teaching and assessment strategies (Garfield et al., 2011) could migrate fragmented learning outcomes.

The dendrograms associated with the heatmap analysis revealed associations among modules, showing that strong performance in one module was often associated with weaker performance in another. These associations varied among the grade categories, with Pass-grade students exhibiting greater variability in knowledge connections. One possible interpretation was that students prioritize effort differently across modules, reinforcing certain concepts while neglecting others. However, alternative explanations, such as misalignment between teaching emphasis and assessment weighting, could also contribute to these trends. Addressing these possibilities in future studies could refine our understanding of assessment variability and fragmented knowledge on graduation. These finding supports the view that assessment models should go beyond standard metrics to capture the full scope of student learning (Huber et al., 2024).

This comprehensive approach in this study to the analysis of marks and grades not only broadened the educators view of student knowledge and achievement, but also underscored the importance of employing both traditional and composite analysis approaches for a nuanced understanding. The study's findings have implications for pedagogy and instructional design, emphasizing the need for a more in-depth exploration of the diverse composition of students' knowledge profiles. However, it is not enough for educators to have this comprehensive understanding of student achievement; this understanding must be translated into constructive feedback that helps students take ownership of their learning. From the perspective of formative assessment theory, this feedback plays a pivotal role in helping students foster reflective study habits (Cizek, 2010; Ismail et al., 2022; Wiliam, 2010).

Ensuring that Pass-grade students have a clear pathway to construct a cohesive knowledge base through thoughtful assessment design is crucial for fostering effective learning outcomes. Thoughtful assessment design could also boost academic engagement, and improve overall statistical literacy, aligning with research the re-evaluate assessment practices to promote genuine understanding (Biggs et al., 2022; Carless, 2015). The finding from this study highlighted the need to refine assessment practices to account for fragmented learning, with a particular emphasis on reinforcing fundamental competencies in later modules. This aligns with ongoing discussions in educational research advocating for adaptive assessment strategies that support student learning beyond just grade achievement (Strielkowski et al., 2025).

While these findings provided valuable insights, it is important to acknowledge some methodological constraints and external factors that may have influenced the results. The COVID-19 pandemic moved all assessment and learning activities online in 2020, which may have altered student engagement and assessment outcomes (Sato et al., 2023). Variability in students' access to resources and shifts in learning environments could have contributed to differences in knowledge distributions, although the years prior to the pandemic did also show similar fragmented knowledge in students.

Conclusion

This study discovered a complex landscape of student knowledge distribution in a tertiary introductory statistics course. Standard evaluation of assessment, while valuable, does not reveal the granular diversity and inconsistencies in student understanding. Exploratory techniques such as combination analysis and heatmaps provided a more comprehensive view, shedding light on the intricate patterns within student knowledge. These techniques could only be performed as all part-marks for all assessment items were clearly identified by the relevant markers, which then allowed mapping of achievement to topics for all students. These findings offered educators valuable insights into enhancing the learning experience and addressing the needs of passing students with diverse knowledge profiles. Although transitioning to competency-based assessment/learning (Holmes et al., 2021) or modularized based learning (Anzaldo, 2021) could potentially address these issues, such transformations require extensive planning and are not always feasible in the short term. The methods explored in this study offer an alternative, enabling educators to visualize student learning patterns more effectively within existing course structures.

This research contributes to the broader conversation on assessment and student learning in higher education, emphasizing the importance of moving beyond traditional grade distribution analyses to gain a

deeper understanding of achievement distribution. One outcome from this study was the restructuring of the statistics course to create a threshold framework with aligned assessment. This approach helps reduce fragmented learning among Pass-grade students, enabling them to gain a basic understanding of higher-order concepts, even if they do not fully grasp the advanced materials. The analytical techniques used in this study provide a transferable approach that could be adapted to various educational settings, provided all marks for each module/topic areas can be granularly defined. By addressing the issues identified in this research, educators and institutions could take proactive steps to enhance the quality of education and improve student comprehension in introductory statistics courses and similar disciplines.

Ethics

Approval to conduct this study was gained from the University of Southern Queensland Research Ethics Committee (ETH2022-0065). The authors confirm that they have followed the ethical publishing practices of Cogent Education.

Disclosure statement

No potential conflict of interest was reported by the author(s).

About the authors

Taryn Axelsen is a Lecturer in Statistics at the Department of Mathematics, Physics, and Computing, University of Southern Queensland. She attained her degree in statistics, is soon to submit her PhD in statistical education and has nearly 20 years of higher education teaching experience. Her research interests include statistical education, overcoming statistics anxiety, advancements in technology for teaching university statistics, and statistical enrichment for school-aged students.

Dr. Rachel King is an Associate Professor in Statistics at the University of Southern Queensland, Australia and she obtained her PhD from Griffith University, Australia. Her research interests include approaches to the learning and teaching of statistics, and the application of statistical methods to applied health, environmental and disaster management data.

Dr. Elizabeth Curtis is a senior lecturer in the School of Education at the University of Southern Queensland in Toowoomba, Australia. She received her PhD at Queensland University of Technology in 2012. Her research interests include mentoring, values education, children's literature and teaching and learning. Elizabeth is involved in research at the national and international level with colleagues in Australia and overseas.

ORCID

Taryn Axelsen  <http://orcid.org/0000-0002-3564-4354>

Rachel King  <http://orcid.org/0000-0002-3302-0919>

Elizabeth Curtis  <http://orcid.org/0000-0001-7561-4910>

Data availability statement

The datasets generated and analysed during this current study are not publicly available due to ethical restrictions, which are in place to protect the privacy and confidentiality of the participants, but de-identified data are available from the corresponding on reasonable request.

References

- Ahshan, R. (2021). A framework of implementing strategies for active student engagement in remote/online teaching and learning during the COVID-19 pandemic. *Education Sciences*, 11(9), 483. <https://doi.org/10.3390/educsci11090483>
- Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education*, 49(4), 499–510. <https://doi.org/10.1080/02602938.2023.2271193>
- AlAfnan, M. A. (2024). Taxonomy of educational objectives: teaching, learning, and assessing in the information and artificial intelligence era. *Journal of Curriculum and Teaching*, 13(4), 173. <https://doi.org/10.5430/jct.v13n4p173>

- Almond, R. G., & Sinharay, S. (2012). What can repeated cross-sectional studies tell us about student growth? *ETS Research Report Series*, 2012(2), i–20. <https://doi.org/10.1002/j.2333-8504.2012.tb02299.x>
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84(3), 261–271. <https://doi.org/10.1037/0022-0663.84.3.261>
- Anzaldo, G. D. (2021). Modular distance learning in the new normal education amidst Covid-19. *International Journal of Scientific Advances*, 2(3), 263–266. <https://doi.org/10.51542/ijscia.v2i3.6>
- Biehler, R., Durand-Guerrier, V., & Trigueros, M. (2024). New trends in didactic research in university mathematics education. *ZDM – Mathematics Education*, 56(7), 1345–1360. <https://doi.org/10.1007/s11858-024-01643-2>
- Biggs, J. B., Tang, C. S., & Kennedy, G. (2022). *Teaching for quality learning at university*. (5th ed.). Open University Press.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: Cognitive and affective domains*. David McKay Co Inc.
- Bromage, A., Pierce, S., Reader, T., & Compton, L. (2022). Teaching statistics to non-specialists: Challenges and strategies for success. *Journal of Further and Higher Education*, 46(1), 46–61. <https://doi.org/10.1080/0309877X.2021.1879744>
- Cain, J., Medina, M., Romanelli, F., & Persky, A. (2022). Deficiencies of traditional grading systems and recommendations for the future. *American Journal of Pharmaceutical Education*, 86(7), 8850. <https://doi.org/10.5688/ajpe8850>
- Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice*. (1st ed.). Routledge. <https://doi.org/10.4324/9781315740621>
- Caspari-Sadeghi, S. (2023). Learning assessment in the age of big data: Learning analytics in higher education. *Cogent Education*, 10(1), 1–11. <https://doi.org/10.1080/2331186X.2022.2162697>
- Chan, C. K. Y. (2023). A review of the changes in higher education assessment and grading policy during Covid-19. *Assessment & Evaluation in Higher Education*, 48(6), 874–887. <https://doi.org/10.1080/02602938.2022.2140780>
- Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). Routledge. <https://doi.org/10.4324/9781315166933-1>
- Collins, J. B., Harsy, A., Hart, J., Anne Haymaker, K., Hoofnagle, A. M., Kuyper Janssen, M., Stewart Kelly, J., Tyler Mohr, A., & Oshaughnessy, J. (2019). Mastery-based testing in undergraduate mathematics courses. *PRIMUS*, 29(5), 441–460. <https://doi.org/10.1080/10511970.2018.1488317>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 1–22. <https://doi.org/10.1186/s41239-023-00392-8>
- Dierker, L., Flaming, K., Cooper, J., Singer-Freeman, K., Germano, K., & Rose, J. (2018). Evaluating impact: A comparison of learning experiences and outcomes of students completing a traditional versus multidisciplinary, project-based introductory statistics course. *International Journal of Education, Training and Learning*, 2(1), 16–28. <https://doi.org/10.33094/6.2017.2018.21.16.28>
- Dunham, B., Yapa, G. Y., & Yu, E. (2015). Calibrating the difficulty of an assessment tool: The blooming of a statistics examination. *Journal of Statistics Education*, 23(3), 1–33. <https://doi.org/10.1080/10691898.2015.11889745>
- Dunn, P. K., Donnison, S., Cole, R., & Bulmer, M. (2017). Using a virtual population to authentically teach epidemiology and biostatistics. *International Journal of Mathematical Education in Science and Technology*, 48(2), 185–201. <https://doi.org/10.1080/0020739X.2016.1228015>
- Ferland, M., Molinaro, C. F., Kosovich, J. J., & Flake, J. K. (2024). Using motivation assessment as a teaching tool for large undergraduate courses: Reflections from the teaching team. *Teaching of Psychology (Columbia, Mo.)*, 51(2), 220–226. <https://doi.org/10.1177/00986283211066485>
- Fischer, J., Bearman, M., Boud, D., & Tai, J. (2024). How does assessment drive learning? A focus on students' development of evaluative judgement. *Assessment & Evaluation in Higher Education*, 49(2), 233–245. <https://doi.org/10.1080/02602938.2023.2206986>
- Galligan, L., Coupland, M., Dunn, P. K., Martinez, P. H., & Oates, G. (2020). Research into teaching and learning of tertiary mathematics and statistics. In J. Way, C. Attard, J. Anderson, J. Bobis, H. McMaster, & K. Cartwright (Eds.), *Research in mathematics education in Australasia 2016–2019*. Springer. https://doi.org/10.1007/978-981-15-4269-5_11
- Gamage, K. A. A., Silva, E. K. d., & Gunawardhana, N. (2020). Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences*, 10(11), 301. <https://doi.org/10.3390/educsci10110301>
- 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>
- Gordon, S., & Nicholas, J. (2010). Teaching with examples and statistical literacy: Views from teachers in statistics service courses. *International Journal of Innovation in Science and Mathematics Education*, 18(1), 14–25. <https://open-journals.library.sydney.edu.au/CAL/article/view/3527>
- Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, 16(1), 22–25. <https://doi.org/10.52041/serj.v16i1.209>
- Gu, Z. (2022). Complex heatmap visualization. *iMeta*, 1(3), e43. <https://doi.org/10.1002/imt2.43>
- Hamel, P., & Lee, W. K. (2024). Supporting the evaluation of authentic assessment in environmental sciences: A case study. *Cogent Education*, 11(1), 1–14. <https://doi.org/10.1080/2331186X.2024.2399433>
- Hansen, G., & Ringdal, R. (2018). Formative assessment as a future step in maintaining the mastery-approach and performance-avoidance goal stability. *Studies in Educational Evaluation*, 56, 59–70. <https://doi.org/10.1016/j.stue-duc.2017.11.005>

- Harsy, A. (2020). Variations in mastery-based testing. *PRIMUS*, 30(8–10), 849–868. <https://doi.org/10.1080/10511970.2019.1709588>
- Henning, G. W., & Roberts, D. M. (2023). *Student affairs assessment* (2nd ed.). Routledge. <https://doi.org/10.4324/9781003447207>
- Hodgson, P., & Pang, M. Y. C. (2012). Effective formative e-assessment of student learning: A study on a statistics course. *Assessment & Evaluation in Higher Education*, 37(2), 215–225. <https://doi.org/10.1080/02602938.2010.523818>
- Holmes, A. G. D., Polman Tuin, M., & Turner, S. L. (2021). Competence and competency in higher education, simple terms yet with complex meanings: Theoretical and practical issues for university teachers and assessors implementing Competency-Based Education (CBE). *Educational Process International Journal*, 10(3), 39–52. <https://doi.org/10.22521/edupij.2021.103.3>
- Holmes, N. (2018). Engaging with assessment: Increasing student engagement through continuous assessment. *Active Learning in Higher Education*, 19(1), 23–34. <https://doi.org/10.1177/1469787417723230>
- Huber, E., Harris, L., Wright, S., White, A., Radulescu, C., Zeivots, S., Cram, A., & Brodzeli, A. (2024). Towards a framework for designing and evaluating online assessments in business education. *Assessment & Evaluation in Higher Education*, 49(1), 102–116. <https://doi.org/10.1080/02602938.2023.2183487>
- Ibarra-Sáiz, M. S., Rodríguez-Gómez, G., & Boud, D. (2021). The quality of assessment tasks as a determinant of learning. *Assessment & Evaluation in Higher Education*, 46(6), 943–955. <https://doi.org/10.1080/02602938.2020.1828268>
- Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: Impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, 12(1). <https://doi.org/10.1186/s40468-022-00191-4>
- Kaspi, S., & Venkatraman, S. (2023). Data-driven decision-making (DDDM) for higher education assessments: A case study. *Systems*, 11(6), 306. <https://doi.org/10.3390/systems11060306>
- Katoue, M. G., & Schwinghammer, T. L. (2020). Competency-based education in pharmacy: A review of its development, applications, and challenges. *Journal of Evaluation in Clinical Practice*, 26(4), 1114–1123. <https://doi.org/10.1111/jep.13362>
- Kehoe, J. (1994). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation*, 4(10), 1–3. <https://doi.org/10.7275/07zg-h235>
- Kolade, O., Owoseni, A., & Egbetokun, A. (2024). Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. *Heliyon*, 10(4), e25953. <https://doi.org/10.1016/j.heliyon.2024.e25953>
- Legacy, C., Le, L., Zieffler, A., Fry, E., & Corrales, P. V. (2024). The teaching of introductory statistics: Results of a National Survey. *Journal of Statistics and Data Science Education*, 32(3), 232–240. <https://doi.org/10.1080/26939169.2024.2333732>
- Maki, P. L. (2023). *Assessing for learning: Building a sustainable commitment across the institution*. (2nd ed.). Routledge. <https://doi.org/10.4324/9781003443056>
- Marriott, J., Davies, N., & Gibson, L. (2009). Teaching, learning and assessing statistical problem solving. *Journal of Statistics Education*, 17(1), 1–18. <https://doi.org/10.1080/10691898.2009.11889503>
- Matcha, W., Gašević, D., Uzir, N. A. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp 461–470). <https://doi.org/10.1145/3303772.3303787>
- McKinley, D. W., & Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*, 36(2), 97–110. <https://doi.org/10.3109/0142159X.2013.853119>
- Mitra, S. (2023). How are students learning in a business statistics course? Evidence from both direct and indirect assessment. *INFORMS Transactions on Education*, 23(2), 95–103. <https://doi.org/10.1287/ited.2022.0270>
- Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, 9(3), 1–26. <https://doi.org/10.1002/rev3.3292>
- Muir, S., Tirlea, L., Elphinstone, B., & Huynh, M. (2020). Promoting classroom engagement through the use of an online student response system: A mixed methods analysis. *Journal of Statistics Education*, 28(1), 25–31. <https://doi.org/10.1080/10691898.2020.1730733>
- Mukherjee, P., & Lahiri, S. K. (2015). Analysis of Multiple Choice Questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. *Journal of Dental and Medical Sciences*, 14(2), 47–52. <https://api.semanticscholar.org/CorpusID:46633143>
- Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—Not just for assessment. *Educational Psychology Review*, 35(3), 1–21. <https://doi.org/10.1007/s10648-023-09808-3>
- Nieminen, J. H. (2025). How does assessment shape student identities? An integrative review. *Studies in Higher Education*, 50(2), 287–305. <https://doi.org/10.1080/03075079.2024.2334844>
- Ouyang, F., Dinh, T. A., & Xu, W. (2023). A systematic review of AI-driven educational assessment in STEM education. *Journal for STEM Education Research*, 6(3), 408–426. <https://doi.org/10.1007/s41979-023-00112-x>
- Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, 100020. <https://doi.org/10.1016/j.caeai.2021.100020>

- Perkins, M. (2023). Academic Integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2), 1–24. <https://doi.org/10.53761/1.20.02.07>
- Posit Team. (2023). *RStudio: Integrated Development Environment for R*. In Posit Software, PBC. <http://www.posit.co/>
- Pullen, R., Thickett, S. C., & Bissember, A. C. (2018). Investigating the viability of a competency-based, qualitative laboratory assessment model in first-year undergraduate chemistry. *Chemistry Education Research and Practice*, 19(2), 629–637. <https://doi.org/10.1039/C7RP00249A>
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raffaghelli, J. E., & Grion, V. (2023). Beyond just metrics: For a renewed approach to assessment in higher education. In J. E. Raffaghelli & A. Sangrà (Eds.), *Data cultures in higher education: Emergent practices and the challenge ahead* (pp. 89–121). Springer International Publishing. https://doi.org/10.1007/978-3-031-24193-2_4
- Raković, M., Gašević, D., Hassan, S. U., Ruipérez Valiente, J. A., Aljohani, N., & Milligan, S. (2023). Learning analytics and assessment: Emerging research trends, promises and future opportunities. *British Journal of Educational Technology*, 54(1), 10–18. <https://doi.org/10.1111/bjet.13301>
- Rutherford, S., Pritchard, C., & Francis, N. (2025). Assessment IS learning: Developing a student-centred approach for assessment in Higher Education. *FEBS Open Bio*, 15(1), 21–34. <https://doi.org/10.1002/2211-5463.13921>
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30(2), 175–194. <https://doi.org/10.1080/0260293042000264262>
- Sánchez-Prieto, J. C., Gamazo, A., Cruz-Benito, J., Therón, R., & García-Peñalvo, F. J. (2020). *AI-driven assessment of students: Current uses and research trends*. In *Learning and collaboration technologies. Designing, developing and deploying learning experiences*. https://doi.org/10.1007/978-3-030-50513-4_22
- Sato, S. N., Condes Moreno, E., Rubio-Zarapuz, A., Dalamitros, A. A., Yañez-Sepulveda, R., Tornero-Aguilera, J. F., & Clemente-Suárez, V. J. (2023). Navigating the new normal: Adapting online and distance learning in the post-Pandemic Era. *Education Sciences*, 14(1), 19. <https://doi.org/10.3390/educsci14010019>
- Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, 143(6), 565–600. <https://doi.org/10.1037/bul0000098>
- Schultz, M., Young, K., Gunning, T. K., & Harvey, M. L. (2022). Defining and measuring authentic assessment: A case study in the context of tertiary science. *Assessment & Evaluation in Higher Education*, 47(1), 77–94. <https://doi.org/10.1080/02602938.2021.1887811>
- Sembey, R., Hoda, R., & Grundy, J. (2024). Emerging technologies in higher education assessment and feedback practices: A systematic literature review. *Journal of Systems and Software*, 211, 111988. <https://doi.org/10.1016/j.jss.2024.111988>
- Shimizu, Y., & Vithal, R. (2023). School mathematics curriculum reforms: Widespread practice but under-researched in mathematics education. In Y. V. Shimizu, R. (Ed.), *Mathematics curriculum reforms around the world* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-031-13548-4_1
- Shivshankar, S., & Acharya, N. (2024). AI in assessment and feedback. In N. V. E. Cela, & M. Fonkam (Ed.), *Next-generation AI methodologies in education* (pp. 119–146). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-7220-3.ch006>
- Shulruf, B., Poole, P., Jones, P., & Wilkinson, T. (2015). The objective borderline method: A probabilistic method for standard setting. *Assessment & Evaluation in Higher Education*, 40(3), 420–438. <https://doi.org/10.1080/02602938.2014.918088>
- Slade, C., Lawrie, G., Taptamat, T., Browne, E., Sheppard, K., & Matthews, K. E. (2022). Insights into how academics reframed their assessment during a pandemic: Disciplinary variation and assessment as afterthought. *Assessment & Evaluation in Higher Education*, 47(4), 588–605. <https://doi.org/10.1080/02602938.2021.1933379>
- Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., & Vasileva, T. (2025). AI-driven adaptive learning for sustainable educational transformation. *Sustainable Development*, 33(2), 1921–1947. <https://doi.org/10.1002/sd.3221>
- Sutter, C. C., Givvin, K. B., & Hulleman, C. S. (2024). Concerns and challenges in introductory statistics and correlates with motivation and interest. *The Journal of Experimental Education*, 92(4), 662–691. <https://doi.org/10.1080/00220973.2023.2229777>
- Tallman, M. A., Carlson, M. P., Bressoud, D. M., & Pearson, M. (2016). A characterization of Calculus I final exams in U.S. colleges and universities. *International Journal of Research in Undergraduate Mathematics Education*, 2(1), 105–133. <https://doi.org/10.1007/s40753-015-0023-9>
- Vaessee, B. E., van den Beemt, A., van de Watering, G., van Meeuwen, L. W., Lemmens, L., & den Brok, P. (2017). Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: A pilot study. *Assessment & Evaluation in Higher Education*, 42(6), 872–886. <https://doi.org/10.1080/02602938.2016.1204532>
- Vashishth, T. K., Sharma, V., Sharma, K. K., Kumar, B., Panwar, R., & Chaudhary, S. (2024). AI-driven learning analytics for personalized feedback and assessment in higher education. In T. N. N. Vo (Ed.), *Using traditional design methods to enhance ai-driven decision making* (pp. 206–230). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-0639-0.ch009>

- Wang, X., & Cheng, Z. (2020). Cross-sectional studies: Strengths, weaknesses, and recommendations. *Chest*, 158(15), S65–S71. <https://doi.org/10.1016/j.chest.2020.03.012>
- Warnes, G., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & B, V. (2022). *gplots: Various R programming tools for plotting data*. <https://CRAN.R-project.org/package=gplots>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. <https://ggplot2.tidyverse.org>
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. Andrade & G. Cizek (Eds.), *Handbook of formative assessment* (pp. 18–40). Taylor & Francis.
- Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. *Studies in Higher Education*, 47(3), 656–667. <https://doi.org/10.1080/03075079.2020.1779687>
- Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*, 12(3), 9–26. <https://doi.org/10.15353/joci.v12i3.3275>
- Zeng, L. M. (2025). Who knows what will come next to interrupt our assessments': The adaptability of assessments across face-to-face and online instructional environments. *Assessment & Evaluation in Higher Education*, 50(2), 250–265. <https://doi.org/10.1080/02602938.2024.2370425>