ASRHE HERDSA

Advancing Scholarship and Research in Higher Education

Volume 7, Issue 1, Advancing Scholarship and Research in Higher Education

Confidence and performance in quantitative literacy: A study of second-year biochemistry and molecular biology students

Xiangyu (Victoria) Wang, Faculty of Education, The University of Melbourne, Parkville, Australia. ORCID: 0009-0002-0806-4511.

Jan van Driel, Faculty of Education, The University of Melbourne, Parkville, Australia. ORCID: 0000-0002-8185-124X.

Amber Willems-Jones*, Department of Biochemistry & Pharmacology, Faculty of Medicine, Dentistry & Health Sciences, The University of Melbourne, Parkville, Australia. ORCID: 0000-0002-8867-1137. Email: amber.willems@unimelb.edu.au

*corresponding author

Recommended Citation

Wang, X., van Driel, J., & Willems-Jones, A. (2026). Confidence and performance in quantitative literacy: A study of second-year biochemistry and molecular biology students. Advancing Scholarship and Research in Higher Education, 7(1), 1–23. https://doi.org/10.59197/1e9x2d65

Publication
Received: 10 December 2025
Accepted: 11 April 2026
Published: 18 May 2026

Confidence and performance in quantitative literacy: A study of second-year biochemistry and molecular biology students

Xiangyu Wang, Jan van Driel and Amber Willems-Jones

Abstract

Quantitative literacy (QL) is central to success in science and biomedical science degrees in higher education, yet students often report low confidence when applying quantitative skills in authentic contexts. Drawing on self-efficacy theory, this study examined the relationship between self-reported confidence and actual performance on a 30-item QL assessment administered to second-year undergraduate students (n = 137) enrolled in a laboratory-based biochemistry subject at The University of Melbourne. Each item captured student performance and a 5-point confidence rating across eight QL domains (*simple arithmetic, scientific notation, stoichiometry, unit conversion, dilutions, statistical measurements, pH, graphical interpretation*). Analyses combined descriptive statistics, correlation, repeated-measures comparisons by domain, and relationships with item difficulty. Overall, confidence and performance were moderately correlated ($r = 0.49$, $p < 0.001$). Sixteen items—concentrated in statistical measurements, pH interpretation and graphical interpretation—showed significant item-level confidence–performance correlations (ρ up to 0.52). Notably, more difficult items were associated with stronger calibration (confidence–performance alignment), while easier items tended to elicit over-confidence. The study findings emphasise the importance of contextualised QL instruction, integrating confidence judgements as a metacognitive scaffold, and deliberately using moderately challenging tasks to develop accurate self-assessment and applied quantitative competence.

Keywords

Quantitative literacy; confidence; calibration; self-efficacy; biomedical science education; undergraduate assessment

Key contributions/Pathways to collaboration

- Provides item-level evidence (30 test items across eight QL domains) on how student confidence aligns with performance, extending beyond self-efficacy measures.
- Shows moderate overall correlation between student confidence and performance and stronger effects in specific domains (statistical reasoning, graph interpretation).
- Demonstrates that greater item difficulty improves calibration (confidence–performance alignment), highlighting metacognitive activation on challenging tasks.

- Offers actionable recommendations for embedding confidence ratings in formative tasks and for scaffolding applied QL (statistics, data interpretation) in undergraduate curricula.

Introduction

Understanding the confidence of students in their own quantitative literacy skills is vital for advancing effective learning in science and biomedical science education; however, students frequently struggle to accurately evaluate their own competence when engaging with quantitative tasks. Misalignment between confidence and performance in quantitative literacy, broadly and within specific domains, has important implications for learning and engagement, but exploration at a domain-specific level remains underexamined within authentic disciplinary contexts. This study examines the relationship between self-reported confidence and performance across distinct quantitative literacy domains among second-year biochemistry students to generate empirical insight into patterns of alignment and their implications for teaching and learning in higher education.

Conceptualizing quantitative literacy in higher education

Quantitative literacy (QL) is commonly understood to involve interpreting, reasoning with, and applying numerical and data-based information in academic and real-world contexts (Prince et al., 2021; Richman & Dietiker, 2025). Steen (2001) described QL as a "habit of mind", encompassing the capacity to understand the role of mathematics in the world, to make data-informed decisions, and apply mathematical skills to practical needs. In science and biomedical contexts, QL is essential as students must analyse data, interpret results, and engage in quantitative reasoning and mathematical thinking in experimental and clinical settings (Rozgonjuk et al., 2020).

Despite its importance, US data from the Program for the International Assessment of Adult Competencies (PIAAC) indicate that only a minority of adults with bachelor's degrees achieve high levels of numeracy proficiency (Goodman et al., 2016). Similarly, Wilkins (2016) and Cuddington et al. (2023) reported persistent gaps in QL skills among science undergraduates, often linked to low confidence, mathematics anxiety, and/or inadequate preparation. Gabriel et al. (2020) highlighted the role of mathematics anxiety in reducing engagement with quantitative tasks, while Millar et al., (2023) identified low self-efficacy in quantitative skill development among biomedical science students in the UK.

There is ongoing debate about whether QL should be conceptualised as a general cognitive skill or as domain specific and context dependent. Hence, a clear empirical gap persists in understanding how students' confidence aligns with their performance across varied quantitative tasks in authentic disciplinary contexts. In particular, few studies have empirically investigated calibration at the level of specific quantitative domains relevant to biomedical science. While some argue for a broad, cross-disciplinary approach (Wilkins, 2016), others advocate for contextualised models tailored to specific academic contexts (Madison & Steen, 2008). To address both the conceptual considerations and the existing empirical gap, this study employs a domain-specific framework and investigates confidence–performance relationships using an assessment instrument structured around eight quantitative literacy domains (described in the Research design and methodology section). This approach enables a more fine-grained analysis of how alignment varies across different types of quantitative reasoning and responds to calls for contextually grounded and diagnostically informative assessments in higher education (Grotlüschen et al., 2020).

Self-efficacy and academic performance

The relationship between self-perceived confidence and actual performance in QL represents an intersection of educational psychology, science education, and metacognition (Gabriel et al., 2020). Exploring key theoretical frameworks, previous findings, and existing gaps in understanding that relationship, this introduction justifies the need to explore how confidence, which has been conceptualized through self-efficacy theory, aligns with students' QL outcomes in a biomedical science education context.

The theoretical framework guiding this research is Self-Efficacy Theory (Bandura, 1997), which underpins the study's hypothesis that higher confidence correlates with stronger QL performance. Self-efficacy, defined as individuals' beliefs in their ability to organise and execute tasks (Bandura, 1997), plays a critical role in academic performance, particularly in QL. Self-efficacy theory posits that individuals' beliefs in their capabilities to organize and execute actions required to manage prospective situations significantly influence their choices, effort, persistence, and performance (Greco et al., 2022). In educational contexts, self-efficacy influences students' motivation, persistence, and the learning strategies they adopt (Bandura, 1997). Students with high self-efficacy are more likely to approach quantitative tasks with confidence, persist through challenges, and achieve better outcomes (Hiller et al., 2022). Conversely, low self-efficacy often leads to task avoidance and underperformance, even when students possess the necessary skills (Zhang & Wang, 2020).

QL requires interrelated competencies such as statistical reasoning, data interpretation, and measurement skills, which elicit varying levels of self-efficacy (Lu et al., 2023). However, much of the existing research relies on global self-efficacy measures, which typically ask students to assess their overall confidence in “mathematics” or “quantitative reasoning” rather than in specific tasks (Cribbs et al., 2021) or in specific domains. While such measures can offer a general indication of students' perceptions of their abilities, they often fail to capture the nuanced differences in confidence that arise when students encounter a particular type of problems (Lu et al., 2023; Wang & Richarde, 1988). This gap limits understanding of how confidence aligns with performance across different QL domains, hindering the development of targeted instructional strategies.

Empirical evidence consistently supports the positive relationship between self-efficacy and academic performance. For example, data from the Longitudinal Surveys of Australian Youth illustrated that mathematics self-efficacy predicted both enrolment in mathematics-related university courses and academic success (Parker et al., 2017). In STEM disciplines, higher self-efficacy correlates with persistence in challenging coursework and long-term engagement in science careers (Rozgonjuk et al., 2020). These findings reinforce the importance of fostering students' self-efficacy in quantitative subjects, both to enhance academic achievement and to support long-term engagement in science and biomedical science fields.

Beyond academic outcomes, self-efficacy reduces mathematics anxiety, further supporting performance (Klee et al., 2022). However, in higher education, particularly in science and biomedical fields, curricula rarely address students' self-confidence with quantitative tasks (Cuddington et al., 2023). Structured opportunities to reflect on and calibrate confidence against performance are essential to fostering both academic success and professional competence in quantitative reasoning. However, few studies have examined confidence and performance at the domain level within QL assessments, leaving limited understanding of how confidence varies across different quantitative competencies in biomedical science contexts.

The confidence–performance gap

The "confidence–performance gap" refers to the misalignment between students' self-perceived confidence and their actual performance, a common issue in science education (Stankov et al., 2017). This gap manifests as either 'over-confidence', where students overestimate their abilities and underperform, or 'under-confidence', where capable students avoid challenges, limiting their participation and career aspirations in quantitative fields (Kleitman & Stankov, 2007; Prokop, 2020). Over-confidence can result in reduced preparation and poor assessment outcomes, while low self-efficacy discourages engagement with challenging tasks (Rachmatullah & Ha, 2019). Inadequate feedback and prior experiences with failure further exacerbate this gap, particularly in diverse undergraduate cohorts with varying educational backgrounds (Gagnier et al., 2023; Greco et al., 2022).

Calibration, or the alignment of confidence with performance (Hattie, 2013), is critical for effective learning. Studies show that students who accurately assess their abilities perform better academically (Foster & Renie, 2024). However, research also highlights that students often misjudge their performance, particularly on easier tasks where reduced cognitive engagement inflates confidence. Conversely, more difficult tasks potentially encourage reflection, improving confidence calibration (Dauer et al., 2021).

Addressing this confidence–performance gap requires targeted feedback and instructional strategies to help students critically reflect on their confidence and align it with their actual abilities, particularly in quantitative literacy. Therefore, item-level confidence ratings provide a valuable approach for investigating calibration and identifying domain-specific patterns in QL development. This approach forms the basis of the present study.

Research questions

This study is informed by three limitations identified in the existing literature, including the overreliance on global self-efficacy scales, the under exploration of calibration accuracy across task types, and the lack of discipline-specific QL training in biomedical science education (Schaier, 2023). By using a domain-specific, item-level approach, in which students rate their confidence alongside each test item, this study will explore a more granular pattern of confidence-performance alignment across eight distinct domains of QL. Additionally, by incorporating item difficulty into the analysis, this study may contribute new insights into how cognitive challenge shapes metacognitive accuracy in higher education contexts.

The primary research question of this study is: '*What is the relationship between self-reported confidence levels and actual performance in specific QL domains measured through a quantitative literacy test among second-year biochemistry students?'*

To examine the primary research question in depth, three interconnected sub-questions were explored:

(1) What descriptive statistics summarise confidence and performance distributions?
(2) How do performance and confidence vary by QL domain?
(3) How is item difficulty associated with confidence–performance alignment?

Research design and methodology

This study followed a post-positivist approach (Kivunja & Kuyini, 2017), recognising that confidence ratings represent subjective perceptions while allowing for empirical examination through quantitative analysis.

Participants

The study recruited students enrolled in the second-year laboratory-based undergraduate subject *Techniques in Molecular Science*, a core subject in the Biochemistry and Molecular Biology major for Bachelor of Science (BSc) and Bachelor of Biomedicine (BBMed) students at the University of Melbourne. Of the 178 enrolled students, 137 participated.

Entry into these programs is based on the Australian Tertiary Admission Rank (ATAR, 2025), a percentile ranking used nationally to indicate a student's academic achievement relative to their peers. In addition to overall entry scores, both programs anticipate students will have completed secondary school mathematics, typically Mathematical Methods or an equivalent subject that develops algebraic reasoning, functions, and introductory calculus. However, this only a pre-requisite in the BBMed. In this study, students were not explicitly asked about their high school mathematics background.

By the end of their first year of study, students in the BSc and BBMed will have completed the core subjects of chemistry and biology. As a result, the cohort represents students with prior exposure to quantitative reasoning within scientific contexts, providing an appropriate population for examining confidence and performance in quantitative literacy tasks.

Participants were recruited from the cohort of students enrolled in *Techniques in Molecular Science* during a scheduled class session, reflecting a pragmatic recruitment approach based on participant accessibility rather than a sampling strategy intended to produce a representative sample. In-class administration was chosen to provide a consistent testing environment and was appropriate given the classroom-based and time-constrained nature of the research (Creswell & Creswell, 2018).

Students were advised of the study via announcements on the subject's Learning Management Site and recruitment occurred through an invitation in the penultimate 2-hour workshop of the semester. Participation was voluntary, with assurances that non-participation would not affect academic standing. G*Power calculations (Andrade, 2020) confirmed the sample size was sufficient for correlational analysis, requiring at least 85 participants to detect a medium effect size ($r = 0.30$) with 80% power and $\alpha = 0.05$.

Test Instrument

Self-efficacy in this study is equated with self-reported confidence on each item of a 30-item QL test, reflecting the domain-specific nature (Blomquist et al., 2016). In this study, confidence refers to students' self-reported judgement of how confident they were in the correctness of the answer to each item. Students did not receive explicit training in how to rate their confidence level, ensuring that their ratings reflected immediate judgements about their test answers. Students rated their confidence on a 5-point Likert scale corresponding to each test item, enabling an understanding of the confidence–performance relationship and how confidence can vary across different types of quantitative challenges (Lingel et al., 2019).
The QL test assessed both competence and confidence across eight QL domains that had been identified as 'challenging' for students in prior research (Tran, 2024):

- *Simple arithmetic* (e.g., basic calculations, order of operations), 3 items
- *Scientific notation* (e.g., interpreting significant figures), 3 items
- *Stoichiometry measurements* (e.g., mass-mole conversions), 3 items
- *Unit conversions* (e.g., grams to milligrams, etc), 2 items
- *Calculating dilution factors* (e.g., in buffer preparations or assays), 4 items
- *Statistical measurements* (e.g., p-values, standard deviations), 8 items
- *Understanding pH measurements*, 3 items
- *Graphical interpretation* (e.g., extracting values from trendlines and graphs, etc) 4 items.

The development and selection of the 30 test items were informed by the *Next Generation Quantitative Literacy Assessment Framework* (Roohr et al., 2014), which emphasises contextualised, discipline-relevant tasks and alignment with authentic quantitative practices. Content validity was supported using test questions developed by Pakay (2023) for the *Foundations of Biomedical Science* context, ensuring that items reflected quantitative skills encountered in the biomedical curriculum. Tran's (2024) study empirically identified eight quantitative literacy domains relevant to students' learning, which provided the conceptual structure for the current instrument. Test items were selected and mapped to the eight domains, ensuring appropriate coverage. As the questions reflect quantitative tasks commonly encountered in the taught curriculum, this alignment supports both curriculum relevance and the interpretability of results.

The dual-response format of test item and confidence rating supports the study's self-efficacy framework (Bandura, 1997), hypothesising that higher confidence correlates with better performance, though this relationship may vary by domain and task difficulty. The test instrument was reviewed by biomedical science faculty for clarity, timing, and functionality, with minor revisions based on their feedback.

Test delivery

The QL test was delivered via Qualtrics during the week 10 workshop, consisting of 30 domain-specific questions and was completed by students within 30 minutes. The test was integrated into a regular class to minimise workload. Prior to the test, to minimise student anxiety and encourage authentic engagement with the task, participants were informed that their responses would be anonymous and not contribute to formal assessment for the subject. The test link was automatically closed at the end of the workshop to maintain uniform conditions.

Data analysis

Quantitative QL test data were exported from Qualtrics and analysed using IBM SPSS (version 30). Each participant's dataset included 30 binary performance scores (1 = correct, 0 = incorrect) and 30 corresponding confidence ratings on a 5-point Likert scale. Question 30, with two sub-questions, was scored separately (0.5 points each) but followed by a single confidence rating. Total performance scores (out of 30) and total confidence scores (range: 30–150) were calculated for each student.

Descriptive and non-parametric inferential statistics were used to analyse the data. Spearman's rank-order correlation coefficients (ρ) were calculated to examine item-level relationships between binary performance and confidence scores due to the ordinal nature of confidence ratings. Items with no performance variability (e.g., 100% correct responses) were excluded from the analysis. Item difficulty was calculated using the formula ($P = R/T$), where R is the number of correct responses and T is the total number of students (Mitra et al, 2009). Correlations were compared across items to identify well-calibrated (confidence aligned with performance) and mis-calibrated (over- or under-confidence) domains. It was hypothesised, based on self-efficacy theory, that the more difficult items would show stronger confidence–performance correlations due to increased self-regulation on challenging tasks.

Ethics

Institutional ethics approval was obtained (UoM MDHS HREC 2024-30171-57620-5). Participation was voluntary, and all data are reported in aggregate, consistent with the 2025 National Statement on Ethical Conduct in Human Research (National Health and Medical Research Council et al, 2025). Students were provided with a Plain Language Statement, and each gave informed consent for their data to be used for research purposes. To reduce additional potential pressure, students were explicitly informed that participation or non-participation would have no impact on their grades or academic standing in the subject.

All responses were collected anonymously. Data were stored securely, accessible only to the research team, and reported in aggregate form.

Results

A total of 137 second-year undergraduate students completed the QL test. While all participants provided complete performance data, a small subset (n = 6) had minor missing confidence ratings for specific items. Where necessary, listwise deletion was applied, resulting in a final analytic sample of 131 students for multivariate analyses.

Descriptive statistics

Table 1 illustrates the descriptive statistics of the participants in the QL test. Participants answered 84% of items correctly on average.

Table 1

Descriptive statistics for total performance and confidence scores on QL test

	Min	**Max**	**Mean**	**Std. Deviation**	**Skewness**	**Kurtosis**
Performance Score	17.5	30	25.11	2.73	-0.38	-0.377
Confidence Score	59	148	113.59	18.46	-0.403	0.081

Total confidence scores corresponded to an average confidence rating of approximately 3.8 per item, falling between "Moderately confident" and "Very confident." Both distributions approximated normality, as supported by skewness, kurtosis (Table 1), and visual inspection of P–Q plots (data not shown).

Confidence–performance relationships (RQ1)

The relationship between students' total performance and confidence ratings was analysed using Pearson's product-moment correlation. A significant positive correlation was observed, ($r = 0.488$, $p < 0.001$), indicating that higher overall confidence levels were associated with better performance on the QL test as a whole. This represents a medium effect size, based on Cohen's (2013) guidelines, suggesting a meaningful association between perceived and actual competence (Cohen, 2013). The coefficient of determination ($r^2 = 0.238$) indicates that 23.8% of the variance in performance scores is explained by confidence levels.

This result aligns with Bandura's (1997) self-efficacy theory, which posits a positive relationship between confidence and performance. However, confidence alone cannot serve as a moderate predictor of performance outcomes.

Differences across QL domains in performance and confidence (RQ2)

Item-level analyses examined the relationship between confidence and performance on specific QL test questions. Of the 30 items analysed (including two sub-items for Q30), 16 showed low to moderate positive correlations (ρ) at the 0.05 or 0.01 significance level. Correlations were strongest in complex QL domains, including:

- *Unit conversions* (e.g., Q10, $\rho = 0.276$, $p = 0.001$)
- *Dilution calculations* (e.g., Q13, $\rho = 0.193$, $p = 0.024$; Q15, $\rho = 0.463$, $p < 0.001$)
- *Statistical measurements* (e.g., Q16–Q23, multiple items with $\rho > 0.40$)
- *pH interpretation* (e.g., Q24, $\rho = 0.254$, $p = 0.003$; Q25, $\rho = 0.410$, $p < 0.001$)
- *Graphical interpretation* (e.g., Q27 $\rho = 0.283$, $p = 0.001$; Q28, $\rho = 0.277$, $p = 0.001$).

The highest correlation was observed for Q19 (statistical measurement of range, $\rho = 0.515$, $p < 0.001$), indicating strong alignment between confidence and performance. Similarly, Q15 (a dilution factor calculation) showed a correlation of $\rho = 0.463$, $p < 0.001$. In contrast, simpler tasks, where student performance was highest, such as *Simple Arithmetic* (Q1–Q3) and *Stoichiometry Measurements* (Q7–Q9), generally did not yield significant correlations.

Domain-level performance was analysed using repeated measures ANOVA, revealing significant differences across the eight quantitative literacy domains, ($F(7, 952) = 84.96$, $p < 0.001$). Table 2 indicates that students did not perform equally well across all domains assessed. Students performed best in *Stoichiometry Measurements* with a mean ± standard deviation of 0.99 ± 0.05 and *Simple Arithmetic* of 0.98 ± 0.08, suggesting strong mastery of routine numerical and computational skills.

The lowest performance was observed in *Interpreting graphs* and *Statistical measurements* (Table 2), which is perhaps unsurprising as these QL domains require higher-order reasoning, such as interpreting visual data and statistical values (e.g., p-values, means) which may be less familiar to students or underemphasised in their prior learning. The relatively higher standard deviations in these domains indicate greater variability, suggesting under preparedness in some students.

Post hoc comparisons on performance score (see "*Multi Comparison*" column, Table 2) revealed significant pairwise differences. For example, performance in *Simple Arithmetic* was significantly higher than in *Writing scientific notation*, *Calculating dilution factors*, *Statistical measurements*, and *Interpreting graphs*. Similarly, *Converting units* outperformed these lower-performing domains.

Table 2

Performance comparison across eight QL domains as analysed by repeated measures ANOVA (F = 84.96, p < 0.001)

QL domain	Performance Score		Multi comparison
	Mean	**S.D.**	
1. Simple arithmetic	0.98	0.08	
2. Writing scientific notation	0.92	0.15	
3. Stoichiometry measurements	0.99	0.05	
4. Converting units of measurement	0.96	0.14	1＞2＞5＞6＞8； 4＞2 ＞5＞6＞8
5. Calculating dilution factors	0.85	0.20	
6. Statistical measurements	0.72	0.19	
7. Understanding pH measurements	0.91	0.17	
8. Interpreting graphs	0.65	0.20	

Overall, these results suggest students excel in basic numeric questions but struggle with applied and conceptually complex QL questions, particularly in data interpretation and statistical reasoning.

A repeated measures ANOVA also examined differences in students' self-reported confidence across the eight QL domains, revealing significant variation (Table 3). Confidence was highest in *Simple arithmetic*, *Stoichiometry measurements*, and *Writing scientific notation*, reflecting strong confidence in foundational numerical tasks. Elevated confidence in *Stoichiometry measurements* likely stems from the recent coursework students had been involved with related to stoichiometry calculations in laboratory classes, thereby reinforcing these skills.

In contrast, confidence was lowest in *Statistical measurements*, *Interpreting graphs*, and *Calculating dilution factors* (Table 3). These QL domains require analytical and conceptual skills, such as interpreting p-values, analysing data, and calculating concentrations, which students may find more abstract or challenging due to limited exposure.

Table 3

Confidence comparison across eight QL domains as analysed by repeated measures ANOVA (F = 67.17, p < 0.001)

QL domain	Confidence Score		Multi comparison
	Mean	**S.D.**	
1. Simple arithmetic	4.68	0.66	
2. Writing scientific notation	4.42	0.58	
3. Stoichiometry measurements	4.50	0.63	1>5>6;
4. Converting units of measurement	4.30	0.81	1>2;
5. Calculating dilution factors	3.64	0.93	3>4>5;
6. Statistical measurements	3.13	0.86	7>8>6
7. Understanding pH measurements	3.70	1.03	
8. Interpreting graphs	3.49	0.82	

Post hoc comparisons on confidence score (Table 3) also confirmed significant pairwise differences. Confidence in *Simple arithmetic* exceeded that in *Calculating dilution factors*, *Statistical measurements*, and in *Writing scientific notation*. Similarly, confidence in *Stoichiometry measurements* was higher than in *Converting units*, which surpassed confidence in *Calculating dilution factors.*

These findings indicate confidence is domain-specific and generally aligned with performance data. Students felt more confident with basic numerical calculations than in interpretive or statistical reasoning, highlighting the need for targeted instructional support to build self-efficacy in these complex domains.

Item difficulty and calibration (RQ3)

A higher difficulty coefficient (closer to 1) indicates an easier item, whereas a lower coefficient (closer to 0) indicates a more difficult item. Question difficulty coefficients ranged from 0.38 to 1.00 (mean = 0.82) across the 30-item QL test, indicating a moderately easy test (Table 4). As expected, items requiring higher-order skills such as interpreting graphs and conducting statistical reasoning tended to have lower difficulty coefficients, indicating that students found them more challenging.

A strong positive correlation was observed between item difficulty and confidence ratings (Table 4). Students reported higher confidence on easier items and lower confidence on harder items, demonstrating sensitivity to task complexity. This suggests that students appropriately adjusted their self-perceptions when faced with more challenging tasks, while maintaining consistent confidence on simpler ones, regardless of performance.

Table 4

Correlations between item difficulty with confidence and confidence-performance alignment as measured by Spearman's rank-order correlation

Variables Compared	ρ	p	Interpretation
Item Difficulty × Confidence Rating	0.802**	< 0.001	Easier items associated with higher confidence
Item Difficulty × Confidence- Performance Alignment	–0.394*	0.038	Harder items show stronger alignment

Note. *p < 0.05, **p < 0.01.

The analysis revealed a statistically significant negative correlation between item difficulty and the alignment of students' confidence with performance. Alignment was measured by item-level confidence–performance correlation coefficient for each question—that is, the correlation between students' confidence ratings and their correctness on that specific item, as calculated in RQ2. Item difficulty was calculated such that a value of 1 represented the easiest items (most students answered correctly) and 0 the hardest (fewer correct responses). The negative correlation indicates that as items became harder, confidence–performance alignment improved. Conversely, for easier items, students reported high confidence regardless of accuracy, resulting in weaker alignment.

A possible explanation for the above observation is that more complex tasks prompted greater cognitive engagement and self-monitoring. In other words, when students encountered challenging problems, they engaged in more deliberate evaluation of their reasoning before expressing their confidence. In contrast, simpler tasks may have encouraged automatic responding, leading to high confidence, even when their answers were incorrect. This finding aligns with prior research by Hattie (2013) suggesting that cognitively demanding tasks can enhance self-awareness and reflective judgment. In this study, item difficulty appears to influence not only performance but also the accuracy of students' confidence judgments resulting in better calibration; that is, increased alignment of confidence with performance.

Discussion

This study examined the relationship between self-reported confidence and performance on a QL test among second-year biochemistry students at The University of Melbourne. Grounded in Bandura’s self-efficacy theory (1997), the study took a domain-specific approach to assess the alignment between perceived and demonstrated performance across eight QL domains.

Overall relationship

The findings reveal a statistically significant moderate positive correlation between total confidence and performance scores ($r = 0.488$, $p < 0.001$), supporting the hypothesis that higher confidence is associated with better performance. This aligns with Bandura’s (1997) self-efficacy theory which suggests that belief in one’s abilities influences motivation, effort, and performance, and is consistent with prior studies linking mathematical self-efficacy to academic outcomes (Zhang & Wang, 2020; Živković et al., 2023). However, confidence accounted for only 24% of variance in performance, suggesting other factors, such as prior knowledge, test anxiety, and task familiarity, also contribute to outcomes (Rozgonjuk et al., 2020; Warren et al., 2021).

At the item level, 16 of the 30 QLT items showed significant confidence–performance correlations, (at the $p < 0.05$ or < 0.01 levels), indicating that students were able to accurately assess their competence within unique domains of QL specifically, *Unit conversions, Dilution calculations*, *Statistical measurements*, *pH interpretation and Graphical interpretation*. These task types correspond directly to the conceptual domains assessed on the QL test; therefore, significant confidence–performance correlations within these tasks indicate domain-specific calibration, or students’ ability to judge their performance accurately for particular skill sets.

Confidence judgments can be understood as a form of metacognitive monitoring; i.e. an individuals’ evaluation of the accuracy of their own responses. While the present study did not directly measure metacognitive processes, the inclusion of item-level confidence ratings provided an indirect indicator of students’ abilities to evaluate their own understanding. If we consider Bandura’s (1997) framework, self-efficacy refers to beliefs about capability, whereas metacognitive monitoring concerns judgements about performance accuracy. Thus, the observed alignment between confidence and correct performance evident in this study, suggests that in some domains students’ efficacy beliefs were reasonably well calibrated to their demonstrated competence. In other words, confidence is not uniform across all items; instead, it varies based on the type of quantitative operation required, the underlying conceptual demands, and the difficulty level of the task (Lu et al., 2023; Wang & Richarde, 1988).

However, calibration was not consistent across all items. There was misalignment for the remaining 14 items, suggesting instances of miscalibration. This misalignment may reflect that differences in item complexity, familiarity, or ambiguity can influence students’ ability to accurately judge their performance.

Responding to the third research sub-question, the findings at the item level provide stronger empirical evidence that students’ self-perceived confidence is meaningfully linked to their actual performance, particularly when they engage in metacognitive reflection and judgment during assessment.

Domain-specific patterns

The study also revealed considerable variation in students' performance and confidence across the eight QL domains. Addressing research sub-question 2, students performed best and reported the highest confidence in *Simple Arithmetic* and *Stoichiometry Measurements*, suggesting that students are generally familiar with basic mathematical operations, which require less cognitive processes (Pontius & McIntosh, 2020). These results in the *Stoichiometry Measurements* domain also likely reflect students' recent reinforcement in their laboratory subject content. Once again, the alignment between confidence and performance supports Bandura's (1997) self-efficacy theory, as repeated success in these domains likely strengthened students' belief in their abilities (Greco et al., 2022).

Contrastingly, students performed poorly and reported lower confidence in the domains of *Statistical Measurements* and *Interpreting Graphs*, which require abstract reasoning and multi-step problem solving-skills often underemphasised in secondary mathematics curricula (Grotlüschen et al., 2020; Wilkins, 2016). These findings highlight a knowledge transfer gap, whereby students may excel in procedural tasks but struggle with the applied interpretation (Gabriel et al., 2020). Embedding contextualised statistical literacy thorough targeted instructional support early and consistently within higher education subjects is recommended to address this.

Calibration, defined as the alignment between confidence and performance, also varied by domain. In *Statistical Measurements*, lower confidence correlated strongly with performance on some items (e.g., Q19 $\rho = 0.515$) indicating accurate self-assessment. Conversely, in *Simple Arithmetic*, high confidence was observed regardless of correctness, suggesting over-confidence in simpler tasks (e.g., Q2 $\rho = -0.082$), with students potentially assuming they are correct without deliberate critical reflection answers (Hattie, 2013). These findings highlight the need to address both over-confidence and under-confidence through targeted instructional strategies.

The observed pattern of over-confidence in simpler domains may also be interpreted through the lens of the Dunning–Kruger effect, which proposes that individuals with lower competence may over-estimate their performance due to limited metacognitive awareness (Kruger & Dunning, 1999). In the present study, the lack of significant confidence–performance alignment in some *Simple Arithmetic* items, combined with consistently high confidence ratings, suggests that certain students may have overestimated their accuracy without engaging in critical evaluation of their responses. However, the overall moderate positive correlation between confidence and performance indicates that the cohort did not exhibit pervasive miscalibration. Rather than reflecting a global Dunning–Kruger pattern, the findings suggest domain-specific instances of over-confidence, reinforcing the importance of fostering metacognitive monitoring skills to improve calibration in quantitative literacy contexts.

Item difficulty and metacognition

The higher difficulty levels in *Statistical Measurements*, *Understanding pH Measurements*, and *Interpreting Graphs* may be attributed to the context-rich nature of these tasks. These items required students not only to perform calculations or interpret values but also to extract relevant quantitative information from a broader narrative or scientific scenario. Such tasks demand higher-order cognitive skills, including evaluating, filtering, and prioritising information before applying appropriate mathematical procedures. As presented in the literature, contextualised tasks can increase cognitive load and challenge students to integrate mathematical reasoning with real-world knowledge (Manfreda Kolar & Hodnik, 2021; Wilkins, 2016). This aligns with the central tenet of quantitative literacy: its value lies in the ability to apply numerical reasoning in meaningful contexts (Steen, 2001, 2004).

A key finding addressing research sub-question 3 was the significant influence of item difficulty on the confidence–performance relationship. Item difficulty, coded from 1 (easiest) to 0 (hardest), showed a negative correlation with confidence–performance alignment ($\rho = -0.394$, $p = 0.038$, Table 4), suggesting that students were better calibrated on more challenging items, likely due to the activation of metacognitive processes such as self-monitoring and reflective judgement (Lingel et al., 2019). These findings highlight the role of metacognition in confidence judgements and suggest that task difficulty can prompt more deliberate and accurate self-assessment (Foster & Renie, 2024) ultimately leading to better alignment between perceived and actual performance. The findings also align with self-efficacy theory, which indicates that higher self-efficacy supports persistence and effort on challenging tasks (Bandura, 1997). Students with higher confidence were more likely to engage with difficult items, while students with lower confidence may have disengaged or skipped the question, as evidenced by missing performance data of 6 students which clustered around harder questions. This further emphasizes the role of self-efficacy in fostering perseverance in cognitively demanding tasks (Parker et al., 2017).

Overall, task difficulty not only impacts performance but also moderates confidence judgements, emphasising the need for teaching strategies that promote metacognition and self-efficacy in QL education.

Implications for educative practice and recommendations for future research

Taken together, the findings of this study indicate three key outcomes: (1) confidence and performance were moderately positively correlated, suggesting partial but incomplete calibration; (2) calibration and performance varied across the eight QL domains, with stronger alignment in certain applied domains and over-confidence evident in simpler tasks; and (3) item difficulty influenced student calibration, with more challenging items being associated with more accurate self-assessment. These findings provide an empirical basis for the pedagogical considerations and implications outlined below.

Firstly, the observations of the domain-specific differences in confidence and performance highlight the need for educators to recognise that competence in quantitative literacy is not uniform across all domains. While students may excel in basic computations, they often struggle with applied reasoning, particularly in statistical literacy, data interpretation, and graphical analysis. Research suggests that without targeted guidance and opportunities for practice, many students may struggle to apply procedural skills effectively in real-world situations (Geiger et al., 2015; Wilkins, 2016). These results indicate that applied statistical reasoning and data interpretation may require more explicit instructional support within this subject context. Rather than assuming transfer of procedural skills, targeted opportunities for practice and feedback in applied quantitative reasoning tasks may support stronger alignment between confidence and performance. This approach can foster deeper conceptual understanding, sustained engagement, and more effective application of quantitative reasoning in real-world biomedical context.

Secondly, the evidence of over-confidence in simpler tasks and stronger calibration on more difficult items suggests that task design influences metacognitive engagement. Incorporating confidence assessments into formative evaluations may offer additional benefits. Confidence measures can help identify students who require conceptual support or reassurance, while at the same time promoting metacognitive development (Foster & Renie, 2024). By encouraging students to reflect on their confidence alongside their answers, students may become more aware of their judgement processes, thereby improving their ability to self-regulate and accurately assess their abilities over time (Kleitman & Stankov, 2007). Furthermore, the integration of moderately challenging tasks routinely into subject instruction may improve students' calibration between confidence and performance. However, it is essential to ensure these tasks are adequately scaffolded to prevent disengagement of students with low self-efficacy (Mitra et al., 2009).

Finally, while this study is limited to a single second-year biochemistry subject, the domain-specific patterns observed suggest that quantitative literacy development may benefit from reinforcement across multiple stages of a program. QL should be developed iteratively across the curriculum rather than confined to isolated instruction. By integrating quantitative reasoning into meaningful, discipline-specific contexts, educators can cultivate it as a "habit of mind" rather than a discrete body of knowledge (Steen, 2001, 2004). This approach will not only enhance academic performance but will also equip students with the essential skills for evidence-based decision-making and professional competence in biomedical and healthcare careers (Pakay, 2023). However, broader curricular integration cannot be concluded directly from the present data and should be explored through coordinated, multi-subject research designs.

In future research, educators could explore longitudinal studies to track the development of confidence and performance throughout their university studies, evaluating the long-term effects of targeted interventions, such as the provision of e-resources. Additionally, examining demographic and educational background variables, such as prior mathematics experience and secondary school preparation, could inform more inclusive instructional strategies. Comparative studies across disciplines, such as science, engineering, and health sciences, should also be considered to identify context-specific patterns in the confidence–performance relationship, enabling tailored pedagogical approaches to support quantitative reasoning in diverse academic fields.

Limitations

This study has three main limitations. In the first instance, the use of convenience sampling may have introduced self-selection bias, as participants might have had greater interest or confidence in quantitative literacy tasks, potentially inflating confidence and performance estimates. That said, a substantial proportion of students (137/178 = 77%) chose to participate in the research study reducing effects of potential self-selection bias. The second limitation involves the reliance on self-reported confidence ratings which raise the possibility of response bias, with students' scores potentially influenced by social desirability or researcher expectations. Thirdly, the findings have limited generalisability because the study was conducted within a single subject at one Australian university (though the findings may be relevant for other university subjects that use the Pakay open-source text).

The authors acknowledge these limitations while recognizing concurrently that the research offers valuable insights into the confidence–competence relationship and has the potential to inform future studies and educative practices.

Conclusions

This study examined the relationship between self-reported confidence and performance on a QL test among second-year biochemistry students, drawing on Bandura's self-efficacy theory. Specifically, it investigated the extent to which students' confidence was calibrated with their actual performance, and how this calibration varied across QL domains and levels of task difficulty. By adopting a domain-sensitive approach, the study contributes to understanding how confidence judgements function as a metacognitive process that supports learning in quantitative contexts.

Results revealed a positive yet moderate correlation between confidence and performance, influenced by task difficulty, highlighting the multidimensional nature of QL and the need for tailored instruction. These findings have implications for teaching and assessment, including embedding confidence judgements to support calibration, designing moderately challenging tasks, and fostering reflection to promote deliberate metacognitive monitoring and self-awareness thus enhancing quantitative literacy skills. Ultimately, by improving students' abilities to accurately assess their own understanding, they will be better prepared for evidence-based decision-making in data-rich scientific and professional environments.

References

Andrade, C. (2020). Sample size and its importance in research. Indian Journal of *Psychological Medicine, 42*(1), 102–103. https://doi.org/10.4103/IJPSYM.IJPSYM_504_19

Bandura, A. (1997). *Self-efficacy: The exercise of control.* Macmillan.

Blomquist, T., Farashah, A. D., & Thomas, J. (2016). Project management self-efficacy as a predictor of project performance: Constructing and validating a domain-specific scale. *International Journal of Project Management, 34*(8), 1417–1432. https://doi.org/10.1016/j.ijproman.2016.07.010

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Cribbs, J., Huang, X., & Piatek-Jimenez, K. (2021). Relations of mathematics mindset, mathematics anxiety, mathematics identity, and mathematics self-efficacy to STEM career choice: A structural equation modeling approach. *School Science and Mathematics, 121*(5), 275–287. https://doi.org/10.1111/ssm.12470

Cuddington, K., Abbott, K. C., Adler, F. R., Aydeniz, M., Dale, R., Gross, L. J., Hastings, A., Hobson, E. A., Karatayev, V. A., & Killion, A. (2023). Challenges and opportunities to build quantitative self-confidence in biologists. *BioScience, 73*(5), 364–375. https://doi.org/10.1093/biosci/biad015

Dauer, J., Mayes, R., Rittschof, K., & Gallant, B. (2021). Assessing quantitative modelling practices, metamodelling, and capability confidence of biology undergraduate students. *International Journal of Science Education, 43*(10), 1685–1707. https://doi.org/10.1080/09500693.2021.1928325

Foster, C., & Renie, P. (2024). Changes in students' confidence calibration across a sequence of low-stakes confidence assessments. *Asian Journal for Mathematics Education, 3*(4), 406–427. https://doi.org/10.1177/27527263241298968

Gabriel, F., Buckley, S., & Barthakur, A. (2020). The impact of mathematics anxiety on self-regulated learning and mathematical literacy. *Australian Journal of Education, 64*(3), 227–242. https://doi.org/10.1177/0004944120947881

Gagnier, K. M., Holochwost, S. J., Ceren, M., & Fisher, K. R. (2023). Filling a gap: Initial evidence for reliable and valid measures of students' self-concept, self-efficacy, and interest in science with elementary students traditionally underrepresented in STEM. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1242876

Geiger, V., Goos, M., & Dole, S. (2015). The role of digital technologies in numeracy teaching and learning. *International Journal of Science and Mathematics Education, 13*(5), 1115–1137. https://doi.org/10.1007/s10763-014-9530-4

Goodman, M., Finnegan, R., Mohadjer, L., Krenzke, T., Hogan, J., Owen, E., & Provasnik, S. (2016). *Literacy, numeracy, and problemsolving technology-rich environments among U.S. adults: Results from the Program for the International Assessment of Adult Competencies 2012*. http://nces.ed.gov/pubsearch

Greco, A., Annovazzi, C., Palena, N., Camussi, E., Rossi, G., & Steca, P. (2022). Self-efficacy beliefs of university students: Examining factor validity and measurement invariance of the new academic self-efficacy scale. *Frontiers in Psychology, 12*, 498824. https://doi.org/10.3389/fpsyg.2021.498824

Grotlüschen, A., Desjardins, R., & Liu, H. (2020). Literacy and numeracy: Global and comparative perspectives. *International Review of Education, 66*(2), 127–137. https://doi.org/10.1007/s11159-020-09854-x

Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction, 24*, 62–66. https://doi.org/10.1016/j.learninstruc.2012.05.009

Hiller, S. E., Kitsantas, A., Cheema, J. E., & Poulou, M. (2022). Mathematics anxiety and self-efficacy as predictors of mathematics literacy. *International Journal of Mathematical Education in Science and Technology, 53*(8), 2133–2151. https://doi.org/10.1080/0020739X.2020.1868589

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education, 6*(5), 26–41. https://doi.org/10.5430/ijhe.v6n5p26

Klee, H. L., Miller, A. D., & Buehl, M. M. (2022). Mathematics anxiety, self-concept, and self-efficacy: A multidimensional scaling consideration of measures. *The Journal of Experimental Education, 91*(3), 494–516. https://doi.org/10.1080/00220973.2021.2024788

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*(2), 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121. https://doi.org/10.1037/0022-3514.77.6.1121

Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: Do different metacognitive monitoring measures make a difference? *ZDM, 51*(4), 587–600. https://doi.org/10.1007/s11858-019-01062-8

Lu, H., Chen, X., & Qi, C. (2023). Which is more predictive: Domain-or task-specific self-efficacy in teaching and outcomes? *British Journal of Educational Psychology, 93*(1), 283–298. https://doi.org/10.1111/bjep.12554

Madison, B. L., & Steen, L. A. (2008). Evolution of numeracy and the National Numeracy Network. *Numeracy, 1*(1). https://doi.org/10.5038/1936-4660.1.1.2

Manfreda Kolar, V., & Hodnik, T. (2021). Mathematical literacy from the perspective of solving contextual problems. *European Journal of Educational Research, 10*(1), 467–483. https://doi.org/10.12973/eu-jer.10.1.467

Millar, B. C., Tarasov, A., Ternan, N., Moore, J. E., & Murphy, C. (2023). Embedding scientific communication and digital capabilities in the undergraduate biomedical science curriculum. *British Journal of Biomedical Science, 80*, 11284. https://doi.org/10.3389/bjbs.2023.11284

Mitra, N., Nagaraja, H., Ponnudurai, G., & Judson, J. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. *IeJSME, 3*(1), 2–7. https://doi.org/10.56026/imu.3.1.2

National Health and Medical Research Council, Australian Research Council, & Universities Australia. (2025). *National statement on ethical conduct in human research (2025)*. National Health and Medical Research Council. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2025

Pakay, J. (2023). *Foundations of Biomedical Science: Quantitative literacy theory and problems.* Open Educational Resources Collective. https://oercollective.caul.edu.au/foundations-of-biomedical-science/

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S., & Abduljabbar, A. S. (2017). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. In J. Lee & L. Stankov (Eds.), *Noncognitive psychological processes and academic achievement* (pp. 39–58). Routledge.

Pontius, J., & McIntosh, A. (2020). *Critical skills for environmental professionals: Putting knowledge into practice*. Springer. https://doi.org/10.1007/978-3-030-28542-5

Prince, R. N., Frith, V., Steyn, S., & Cliff, A. (2021). Academic and quantitative literacy in higher education: Relationship with cognate school-leaving subjects. *South African Journal of Higher Education, 35*(3), 163–181. https://doi.org/10.20853/35-3-3943

Prokop, T. R. (2020). Calibration and academic performance in students of health sciences. *Health Professions Education, 6*(4), 564–573. https://doi.org/10.1016/j.hpe.2020.08.003

Rachmatullah, A., & Ha, M. (2019). Examining high-school students' overconfidence bias in biology exam: a focus on the effects of country and gender. *International Journal of Science Education, 41*(5), 652–673. https://doi.org/10.1080/09500693.2019.1578002

Richman, A., & Dietiker, L. (2025). Taking mathematical literacy seriously: a review [Book review of Learning and teaching for mathematical literacy: making mathematics useful for everyone, by H. Burkhardt, D. Pead, & K. Stacey, (Eds.)]. *Research in Mathematics Education, 28*(1), 231–235

Roohr, K. C., Graf, E. A., & Liu, O. L. (2014). Assessing quantitative literacy in higher education: An overview of existing research and assessments with recommendations for next-generation assessment. *ETS Research Report Series, 2014*(2), 1–26. https://doi.org/10.1002/ets2.12024

Rozgonjuk, D., Kraav, T., Mikkor, K., Orav-Puurand, K., & Täht, K. (2020). Mathematics anxiety among STEM and social sciences students: the roles of mathematics self-efficacy, and deep and surface approach to learning. *International journal of STEM education, 7*(1), 46. https://doi.org/10.1186/s40594-020-00246-z

Schaier, S. (2023, May 1 2023). Quantitative Confidence. *UNH Today.* https://www.unh.edu/unhtoday/2023/05/quantitative-confidence

Stankov, L., Morony, S., & Lee, Y. P. (2017). Confidence: the best non-cognitive predictor of academic achievement? In J. Lee & L. Stankov (Eds.), *Noncognitive psychological processes and academic achievement* (pp. 19–38). Routledge.

Steen, L. A. (2001). *Mathematics and democracy: the case for quantitative literacy*. NCED.

Steen, L. A. (2004). *Achieving quantitative literacy: An urgent challenge for higher education*. MAA.

Tran, T. T. G. (2024). *Exploring student challenges in transferring quantitative literacy to a biomedical science subjects* [Unpublished manuscript]. Research Project in Education (EDU91197), University of Melbourne.

Wang, A. Y., & Richarde, R. S. (1988). Global versus task-specific measures of self-efficacy. *The Psychological Record, 38*(4), 533–541. https://doi.org/10.1007/BF03395045

Warren, L., Reilly, D., Herdan, A., & Lin, Y. (2021). Self-efficacy, performance and the role of blended learning. *Journal of Applied Research in Higher Education, 13*(1), 98–111. https://doi.org/10.1108/JARHE-08-2019-0210

Wilkins, J. L. (2016). An assessment of the quantitative literacy of undergraduate students. *The Journal of Experimental Education, 84*(4), 639–665. https://doi.org/10.1080/00220973.2015.1111854

Zhang, D., & Wang, C. (2020). The relationship between mathematics interest and mathematics achievement: Mediating roles of self-efficacy and mathematics anxiety. International *Journal of Educational Research, 104*, 101648. https://doi.org/10.1016/j.ijer.2020.101648

Živković, M., Pellizzoni, S., Doz, E., Cuder, A., Mammarella, I., & Passolunghi, M. C. (2023). Math self-efficacy or anxiety? The role of emotional and motivational contribution in math performance. *Social Psychology of Education, 26*(3), 579–601. https://doi.org/10.1007/s11218-023-09760-8