



Mathematics Task-Specific Self-Efficacy and Actual Performance among Secondary School Students: Examining the Relationship between Confidence and Accuracy

Yusuf Feyisara Zakariya^{1*}, Esther Vanessa Maina^{1,2}, Ramatu Ematum Umahaba¹, Patience Guyah¹, Safiya Shafiu Muhammed³

¹Department of Science Education, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

²Demonstration Secondary School, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

³Institute of Education, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

*Corresponding author: yfzakariya@abu.edu.ng

Abstract

This study examined the relationship between task-specific mathematics self-efficacy and actual performance among 248 senior secondary school students (111 males, 125 females, 12 missing gender data) in Kaduna state. Unlike studies correlating a global self-efficacy score with a separately administered achievement test, this design elicited students' confidence immediately before each of fifteen multiple-choice mathematics items, using validated self-efficacy ($\alpha = .848$) and performance (Guttman split-half = .715) instruments, analysed with descriptive and inferential statistics. At the student level, mean confidence correlated weakly with percentage correct (Pearson $r = .12$, 95% CI $[-.01, .25]$, $p = .059$; Spearman $\rho = .18$, $p = .007$, $n = 237$). A crossed-random-effects mixed logistic regression linking item-level confidence to correctness showed that higher confidence significantly predicted greater odds of a correct response (OR = 1.06, 95% CI $[1.03, 1.09]$, $p < .001$), confirmed by a cluster-robust GEE model (OR = 1.06, $p = .002$). Calibration analysis, however, revealed pronounced positive bias: mean confidence exceeded mean accuracy by .34 ($t(236) = 21.46$, $p < .001$), with 93.7% of students overconfident and 62.9% of high-confidence responses (≥ 7) incorrect. Findings indicate that confidence weakly but reliably discriminates correct from incorrect responses, yet students are substantially overconfident relative to actual performance. Implications for instruction, formative assessment, and self-monitoring are discussed.

Keywords: mathematics self-efficacy, task-specific confidence, actual performance, calibration, overconfidence, mixed-effects logistic regression, secondary school mathematics

1 Introduction

Mathematics learning requires students not only to possess mathematical knowledge and skills but also to make judgments about their own capability to successfully perform mathematical tasks. Long before a student writes down an answer, that student has already formed some impression of how likely the answer is to be correct. These judgments of capability, generally termed self-efficacy beliefs, are widely regarded as central to how students approach, persist with, and ultimately succeed or fail at mathematical tasks (Bandura, 1997; Zakariya et al., 2019). Yet confidence and competence are not the same thing. A student may feel highly confident and answer incorrectly, while another may feel distinctly unsure despite producing a correct answer. Understanding the extent to which students' confidence corresponds to what they actually accomplish is therefore a question of both theoretical and practical importance.

The present study addresses this question directly by examining the relationship between task-specific mathematics self-efficacy and actual performance among Nigerian senior secondary school students. Unlike the substantial body of prior research that measures mathematics self-efficacy using a general questionnaire administered independently of a separate mathematics achievement test (Pajares & Miller, 1994; Yang et al., 2024), the present design obtains students' confidence judgments immediately before they attempt each individual mathematics item and links each confidence rating to the corresponding item-level outcome. This item-matched design provides a considerably closer correspondence between the psychological judgment being measured and the specific performance it purports to predict, and it allows the relationship

between confidence and accuracy to be examined not only in aggregate but also at the level of the individual task. In specific terms, this study addressed the following research questions:

1. What are the levels of task-specific mathematics self-efficacy and actual performance among senior secondary school students?
2. What is the relationship between students' task-specific mathematics self-efficacy and their actual performance on mathematics tasks?
3. To what extent does task-specific mathematics self-efficacy predict students' probability of answering individual mathematics tasks correctly?
4. To what extent are students' confidence judgments calibrated with their actual performance on mathematics tasks?

This distinction matters conceptually. Self-efficacy, in Bandura's (1997) formulation, is a judgment of one's capability to organize and execute the specific courses of action required to attain a designated type of performance. It is inherently task-specific rather than a generalized trait or mood. A mathematics self-efficacy score obtained from a general questionnaire administered on one occasion and correlated with an achievement test administered on another occasion necessarily blurs this task-specificity, since the confidence being reported may not correspond to the precise items on which performance is subsequently assessed. A second contribution of the study lies in its treatment of the relationship between confidence and performance as two conceptually distinct questions rather than one. The first question is one of association: does greater confidence tend to accompany a greater probability of answering correctly? The second is one of calibration: does the level of confidence expressed correspond appropriately, in magnitude, to the level of performance actually achieved? A positive association between confidence and performance would indicate that students' confidence judgments discriminate, at least to some degree, between tasks they are more and less likely to solve successfully. It would not, by itself, establish that students are accurately calibrated, since confidence could be systematically and uniformly inflated (or deflated) relative to actual accuracy while still preserving a positive relationship with performance across items. Both questions are examined in the analyses that follow.

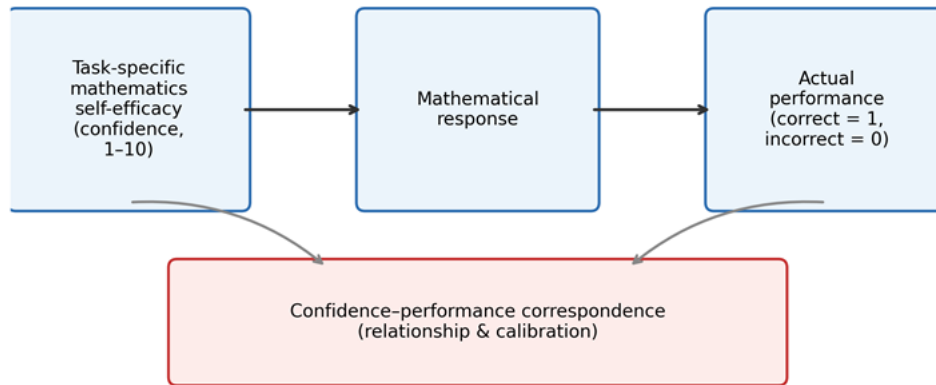
2 Theoretical Foundation

The study is anchored in Bandura's (1997) Social Cognitive Theory, and specifically in the construct of self-efficacy — people's judgments of their capability to organize and execute the actions required to attain particular performances. Self-efficacy's task-specific nature makes it particularly well suited to the present design, since students' judgments were elicited in direct relation to individual mathematics problems rather than to mathematics as a general domain. Within social cognitive theory, self-efficacy beliefs are proposed to influence students' expectations regarding successful task completion, which in turn shape effort, persistence, and cognitive engagement during problem solving; these motivational and cognitive processes are related to performance experiences, which may consequently be reflected in subsequent performance outcomes (Bandura, 1997; Pajares, 1996).

At the same time, social cognitive theory does not predict that confidence and performance will correspond perfectly. Self-efficacy is a perceived rather than an objectively verified capability, and students can hold beliefs about their own capability that are inaccurate in either direction — overestimating what they can do, or underestimating it. This possibility is captured in the broader literature on confidence–accuracy calibration, which distinguishes the discriminative value of a confidence judgment (whether it tracks variation in performance across items) from its calibration (whether its absolute level matches observed accuracy) (Fischhoff, Slovic, & Lichtenstein, 1977; Stankov, Morony, & Lee, 2014). Accordingly, confidence and actual competence are treated in the present study as related but empirically distinguishable constructs, and both the relationship between them and their calibration are examined as separate analytical questions. Figure 1 presents the conceptual model guiding this study.

Figure 1

Conceptual framework linking task-specific self-efficacy, mathematical response, actual performance, and confidence–performance correspondence



As shown in Figure 1, a mathematics task is presented to a student, who first renders a task-specific self-efficacy judgment. This is, a confidence rating, on a scale from 1 to 10, of the student's perceived capability to solve that particular task successfully. After rating the 15 items of the self-efficacy scale, the student then produces a mathematical response to each of the items, which is scored as either correct or incorrect. The central analytical interest of the study lies in the correspondence between the confidence judgment rendered before the response and the actual performance observed afterward. This constitutes a correspondence that has both a relationship component (the strength of statistical association between confidence and correctness) and a calibration component (the degree to which the level of confidence matches the level of actual accuracy).

3 Materials and Methods

3.1 Research Design

The study employed a quantitative cross-sectional survey design to examine the relationship between secondary school students' task-specific mathematics self-efficacy and their actual performance on mathematics tasks. The distinctive feature of the design is the item-level measurement of students' confidence immediately before they attempted each mathematics problem, so that confidence judgments were obtained in direct relation to the same tasks for which subsequent performance was recorded.

3.2 Participants

The participants comprised 248 senior secondary school two students with a mean age of approximately 16 years, selected through simple random sampling. The sample included 111 males (44.8%), 125 females (50.4%), and 12 students (4.8%) for whom gender information was not recorded. Because gender was missing for a nontrivial minority of participants, these 12 cases were retained as missing rather than assigned to either category, and gender was examined only as a secondary, exploratory variable rather than as a central focus of the study.

3.3 Measurement of Task-Specific Mathematics Self-Efficacy

We used a 15 item-level confidence ratings (MSE1–MSE15) to measure of task-specific mathematics self-efficacy. This instrument was developed and piloted by researchers with the reliability coefficient of .848 using Cronbach's alpha (Zakariya, 2022). Before attempting each mathematics item, students rated their confidence in their ability to successfully complete the mathematical task using a 10-point scale ranging from 1 (not confident at all) to 10 (completely confident).

3.4 Measurement of Actual Mathematics Performance

Actual performance was determined from students' responses to the 15 multiple-choice items (ACPT1–ACPT15) developed by the researchers and covering logarithms, circle geometry quadratic equations, approximation and accuracy, trigonometry and geometry and trigonometry. Each item was coded 1 when the student's selected option corresponded to the predetermined correct answer and 0 otherwise. An overall performance score was obtained by summing the 15 items, producing a possible range of 0 to 15. Further, a percentage score was also calculated by dividing the total number of correct responses by 15 and multiplying by 100. The reliability coefficient of the performance was .715 using Guttman split-half method.

3.5 Data Collection and Analysis

The two research instruments were administered to the classroom during a part of the lesson period. The generated datasets were coded and screened for duplicate records, out-of-range values, and inconsistencies in variable coding. No duplicate cases were found. All 15 ACPT variables were confirmed to be coded dichotomously as 0 or 1 with no missing values. Further, the MSE items were coded from 1 to 10. The generated datasets were analysed using descriptive statistics (mean, standard deviation, median, range, valid n). For each student with at least one valid MSE rating, an overall self-efficacy score was calculated as the mean of the available item-level confidence ratings, and an overall performance score was calculated as the sum of the 15 dichotomous performance items. The primary student-level relationship between task-specific self-efficacy and actual performance was examined using Pearson's product-moment correlation coefficient, accompanied by its 95% confidence interval (via Fisher's z transformation); Spearman's rank-order correlation was reported as a sensitivity analysis appropriate to the ordinal, bounded nature of the two variables.

For the item-level mixed-effects analysis, because each student contributed up to 15 item-level observations, and because items are nested within students, simple correlation of all student-item pairs would violate the independence assumption underlying ordinary correlation and regression. Thus, 3,078 valid student-item observations were restructured into long format and analyzed using a mixed-effects logistic regression model with the binary correctness outcome (ACPT) as the dependent variable, item-level confidence (mean-centred) as the fixed-effect predictor, and crossed random intercepts for student and for item, of the form $\text{logit}[P(ACPT_{ij} = 1)] = \beta_0 + \beta_1 \text{MSE}_{ij} + u_i + v_j$, where i indexes student and j indexes item. Calibration was examined separately from the association analyses. Item-level confidence ratings were rescaled from the 1–10 range to a 0–1 proportion by dividing by 10. At the student level, a confidence–performance bias score was calculated as $\text{Bias}_i = \bar{C}_i - P_i$, where $\bar{C}_i$ is the student's mean rescaled confidence and P_i is the student's proportion of items answered correctly. Then, a one-sample t-test evaluated whether the mean bias across students differed significantly from zero. At the item-response level, mean confidence was compared between correct and incorrect responses using an independent-samples (Welch) t-test, and the proportion of high-confidence (rating ≥ 7) responses that were nonetheless incorrect, and the proportion of low-confidence (rating ≤ 3) responses that were nonetheless correct, were calculated as descriptive indices of overconfidence and underconfidence, respectively. All inferential tests used a two-sided significance level of $\alpha = .05$ using SPSS version 20 and R Studio.

4 Results

4.1 Addressing Research Question 1

In order to investigate the levels of task-specific mathematics self-efficacy and actual performance among senior secondary school students, we computed the descriptive statistics of the generated datasets. Table 1 presents item-level descriptive statistics for the 15 confidence ratings alongside item difficulty (proportion correct) for the corresponding performance items.

Table 1

Item-level descriptive statistics for self-efficacy and item difficulty for actual performance

Item	Mean	SD	Median	Valid n (MSE)	Proportion correct (ACPT)
Item 1	6.33	2.83	6.0	217	0.431
Item 2	6.55	2.82	6.0	215	0.375
Item 3	6.29	2.83	6.0	206	0.218
Item 4	6.78	2.78	7.0	212	0.335
Item 5	5.49	2.73	5.0	190	0.145
Item 6	6.02	2.86	5.5	202	0.238
Item 7	7.68	2.72	9.0	224	0.403
Item 8	7.06	2.81	8.0	213	0.371
Item 9	7.35	2.72	8.0	220	0.286
Item 10	6.23	2.69	6.0	207	0.270
Item 11	5.84	2.69	5.0	205	0.286
Item 12	5.33	2.64	5.0	170	0.331
Item 13	6.18	2.55	6.0	203	0.290
Item 14	6.24	2.68	6.0	195	0.270
Item 15	6.39	2.92	6.0	199	0.310

Table 1 showed that the mean item-level confidence ranged from 5.33 (Item 12) to 7.68 (Item 7) on the 1–10 scale, indicating that students were, on average, moderately to fairly confident across the item set, with some items eliciting distinctly higher confidence than others. Item difficulty, expressed as the proportion of students answering correctly, ranged from .145 (Item 5) to .431 (Item 1), indicating that the assessment as a whole was demanding; no single item was answered correctly by a majority of students. It is notable that Item 7, which elicited the highest mean confidence (7.68), was answered correctly by only 40.3% of students, while Item 1, the easiest item by observed difficulty (43.1% correct), elicited only moderate confidence (6.33). This is an early indication, explored more formally in the calibration analysis, that item-level confidence did not consistently track item-level difficulty. Further, Table 2 presents the descriptive statistics of overall mean self-efficacy and actual performance.

Table 2

Descriptive statistics for overall task-specific self-efficacy and actual performance

Variable	N	Mean	SD	Median	Min	Max
Overall MSE (mean confidence, 1–10)	237	6.50	1.95	6.40	1.00	10.00
ACPT total (0–15)	248	4.56	2.59	4.00	0	13
ACPT percentage (%)	248	30.40	17.27	26.67	0.00	86.67

At the student level, Table 2 shows that the overall mean self-efficacy (averaged across each student's valid items) was 6.50 (SD = 1.95) on the 1–10 scale, indicating a moderately high average level of confidence. Overall actual performance was considerably lower in relative terms: students answered a mean of 4.56 of the 15 items correctly (SD = 2.59), equivalent to a mean of 30.4% correct (SD = 17.3%). No student answered all 15 items correctly; the highest total was 13 of 15 (86.7%), and several students answered no items correctly.

4.2 Addressing Research Question two

To address research question two, we examined the Pearson correlation between each student's overall mean task-specific self-efficacy and overall percentage correct ($n = 237$, students with at least one valid MSE rating). Figure 2 displays the scatterplot of the two variables together with the fitted regression line and 95% confidence band.

Figure 2

Student-level relationship between mean confidence and mathematics performance

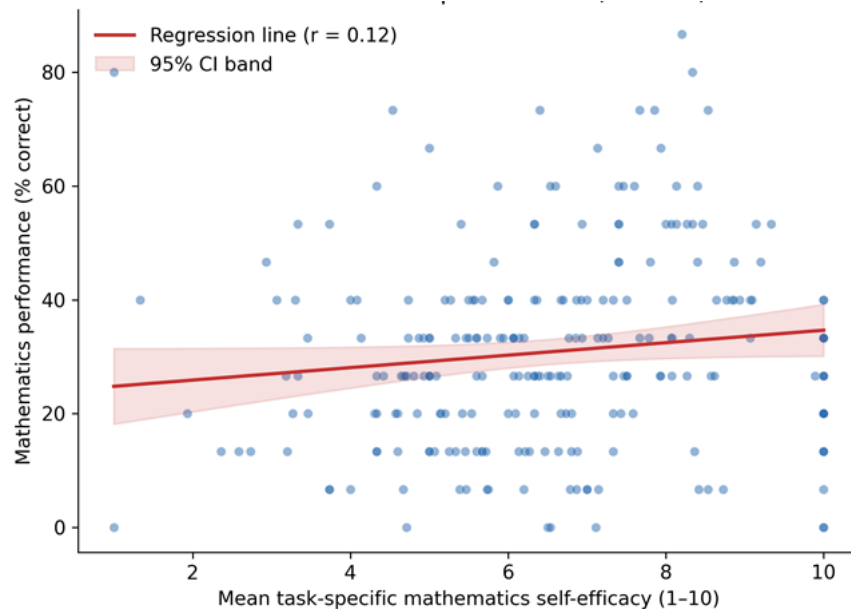


Figure 2 illustrates the considerable scatter underlying the small aggregate correlation: students at every level of mean confidence, from very low to very high, are found across almost the full range of observed performance. The correlation was positive but small in magnitude and did not reach conventional statistical significance: $r = .123$, 95% CI $[-.005, .246]$, $p = .059$. Given the ordinal nature of the confidence scale and the bounded nature of the percentage-correct outcome, Spearman's rank-order correlation was computed as a planned sensitivity analysis and was somewhat larger and statistically significant: $\rho = .175$, $p = .007$. Together, these results provide weak but not entirely absent evidence of a positive association between self-reported confidence and actual mathematics performance at the student level. An exploratory analysis by gender (Table 3) suggested that this weak association was not uniform across subgroups

Table 3

Correlation between task-specific self-efficacy and actual performance

Analysis	N	r / ρ	95% CI	p
Pearson r (overall MSE, ACPT %)	237	.123	$[-.005, .246]$	.059
Spearman ρ (sensitivity analysis)	237	.175	—	.007
Male subgroup (Pearson r)	103	-.001	—	.994
Female subgroup (Pearson r)	123	.235	—	.009

Note. Analyses based on students with at least one valid MSE rating ($n = 237$ overall; 103 male, 123 female, 11 missing gender among this subsample). CI = confidence interval.

As shown in Table 3, among male students ($n = 103$) the correlation was essentially zero ($r = -.001$, $p = .994$), whereas among female students ($n = 123$) it was larger and statistically significant ($r = .235$, $p = .009$). This gender difference was not a focus of the study's primary hypotheses and is reported as a secondary, exploratory finding that may warrant more targeted investigation in future research rather than as a confirmed effect.

4.3 Addressing Research Question Three

In order to address the third research questions, a mixed-effects logistic regression with crossed random intercepts for student and item was therefore fitted, with item-level confidence (mean-centered) as the fixed-effect predictor of the binary correctness outcome. Because the 3,078 valid student-item confidence–performance pairs are nested within 237

students and 15 items, they cannot be treated as independent observations in an ordinary regression. The results of this analysis are presented in Table 4.

Table 4

Mixed-effects logistic regression predicting item-level mathematics performance from task-specific confidence

Parameter	Estimate	SE	OR	95% CI (OR)	z	p
Intercept	-0.848	0.041	0.428	—	—	—
MSE (item-level confidence, centered)	0.056	0.015	1.057	[1.028, 1.088]	3.86	<.001
Random effect SD — student	—	—	1.413	—	—	—
Random effect SD — item	—	—	1.171	—	—	—

Note. Model: $\text{logit}[P(\text{ACPT} = 1)] = \beta_0 + \beta_1 \cdot \text{MSE}(\text{centered}) + u(\text{student}) + v(\text{item})$, estimated via variational Bayes mixed generalized linear model; $n = 3,078$ student-item observations from 237 students across 15 items. OR = odds ratio.

The presented results in Table 4 show that the fixed effect of confidence was positive and statistically significant: $\beta = 0.056$ ($SE = 0.015$), $z = 3.86$, $p < .001$, corresponding to an odds ratio of 1.057 (95% CI [1.028, 1.088]). In substantive terms, each one-point increase in a student's stated confidence for a given item was associated with approximately a 5.7% increase in the odds of answering that item correctly, holding constant the general propensity of that particular student to answer correctly and the general difficulty of that particular item. The estimated random-effect standard deviations (1.413 for students, 1.171 for items) indicate that both student ability and item difficulty contributed substantial, and roughly comparable, sources of variability in the likelihood of a correct response, over and above the effect of confidence itself. As a robustness check, a cluster-robust generalized estimating equations model with item entered as a fixed factor (accounting for item difficulty directly rather than through a random effect) and standard errors clustered by student was fitted. The results are presented in Table 5.

Table 5

Robustness check: comparison of item-level models

Model	n (obs.)	Coefficient	SE	OR	p-value
Primary: crossed random-effects logistic (student, item)	3,078	0.056	0.015	1.057	<.001
Robustness: GEE, item fixed effects, student-clustered SE	3,078	0.059	0.019	1.061	.002

As presented in Table 5, the generalized estimating equations (GEE) produced a closely convergent result: $OR = 1.061$, $p = .002$. The consistency of the effect across two different modeling strategies supports the proposition that task-specific confidence is significantly, if modestly, associated with the probability of answering the corresponding item correctly.

4.4 Addressing Research Question Four

To address the fourth research question, we examined calibration independently of the association analyses earlier reported. This is necessary since a positive relationship between confidence and performance does not, by itself, establish that confidence is calibrated to the correct absolute level of accuracy. The results of this analysis are presented in Table 6.

Table 6

Confidence–performance calibration indices

Index	n	Value
Mean confidence (0–1 scale)	237	0.650 (SD = 0.195)
Mean accuracy (proportion correct)	237	0.308 (SD = 0.174)
Mean confidence–accuracy bias (confidence – accuracy)	237	0.342 (SD = 0.245)
One-sample t-test of bias vs. 0	237	$t(236) = 21.46, p < .001$
Students with positive bias (overconfident)	237	93.7%
Students with negative bias (underconfident)	237	5.9%
Mean confidence correct response (item-level)	997	6.77 (SD = 2.87)
Mean confidence incorrect response (item-level)	2,081	6.25 (SD = 2.78)
Welch t-test (correct vs. incorrect confidence)	3,078	$t = 4.80, p < .001$
High-confidence (≥ 7) responses that were incorrect	1,457	62.9%
Low-confidence (≤ 3) responses that were nonetheless correct	505	28.5%

As shown in Table 6, at the student level, mean confidence (rescaled to the 0–1 range) was .650 (SD = .195) while mean observed accuracy was only .308 (SD = .174) — a mean discrepancy (bias) of .342 (SD = .245), which was highly significantly different from zero, $t(236) = 21.46, p < .001$. The majority of students, 93.7%, showed a positive bias (confidence exceeding accuracy); only 5.9% showed the reverse pattern, and essentially no student was exactly calibrated. This constitutes strong, data-derived evidence of substantial and systematic overconfidence in this sample: students' self-reported capability to solve the mathematics tasks presented to them considerably exceeded what they actually demonstrated. Further, At the item-response level, mean confidence was higher for responses that turned out to be correct ($M = 6.77, SD = 2.87, n = 997$) than for responses that turned out to be incorrect ($M = 6.25, SD = 2.78, n = 2,081$), a statistically significant difference, Welch $t = 4.80, p < .001$. This difference is consistent with the positive item-level association reported in Section 4.2, but its magnitude is modest. We further characterize the calibration pattern descriptively. Table 7 presents the observed accuracy associated with each of the four confidence categories (Low, Moderate, High, Very high) and with each individual rating point on the 1–10 scale respectively.

Table 7

Observed accuracy by confidence category (item-level)

Confidence category	n (obs.)	Observed correct	proportion
Low (1–3)	505	.285	
Moderate (4–6)	1,116	.280	
High (7–8)	481	.347	
Very high (9–10)	976	.382	

As shown in Table 7, among the 1,457 item responses rated with high confidence (≥ 7 on the 1–10 scale), 62.9% were nonetheless incorrect — a substantial rate of confident error. Conversely, among the 505 item responses rated with low confidence (≤ 3), 28.5% were nonetheless correct, indicating that low confidence, too, was an imperfect guide to actual performance, though markedly less so than high confidence was to success. Observed accuracy increased only modestly and non-monotonically across confidence categories, from .285 among low-confidence responses to .382 among very-high-confidence responses — a rise of roughly ten percentage points across the full breadth of the confidence scale, far smaller than the rise that would be expected under accurate calibration.

5 Discussion

This study set out to examine, with an item-matched, task-specific design, the correspondence between secondary school students' confidence in their mathematical ability and what they actually accomplished on the corresponding mathematics tasks. The findings paint a coherent and, in some respects, cautionary picture of mathematics self-efficacy in this sample.

First, students' confidence, on average, discriminated to a modest but statistically detectable degree between tasks they were more and less likely to answer correctly. The item-level mixed-effects logistic regression, showed a significant positive effect of confidence on the odds of a correct response ($OR = 1.06$ per one-point increase in confidence), a result corroborated by an independent cluster-robust estimation strategy. This is consistent with the general proposition, central to Bandura's (1997) social cognitive theory, that self-efficacy judgments are not arbitrary but instead track, at least partially, students' genuine capability. Recent research likewise indicates that mathematics self-efficacy can be meaningfully related to mathematical performance while remaining distinguishable from actual competence (Russell & Phelps-Gregory, 2026; Bozkuş & Öztürk, 2026). The convergence of two different statistical approaches on essentially the same effect size lends some confidence to this conclusion despite its modest magnitude.

Second, however, the strength of this relationship was weak in absolute terms, and this was true at both levels of analysis. The student-level Pearson correlation between mean confidence and percentage correct did not reach the conventional threshold for statistical significance ($r = .12$, $p = .059$), although the rank-based Spearman correlation, better suited to the ordinal and bounded nature of the variables, was significant ($\rho = .18$, $p = .007$). Odds ratios of approximately 1.06 per unit of a 10-point scale are, in absolute terms, small: moving from the lowest to the highest possible confidence rating (a nine-unit change) corresponds to an estimated increase in the odds of a correct response of only around 67%, a moderate effect on a task where, overall, fewer than a third of items were answered correctly. In other words, while confidence is related to performance, it is a considerably imperfect signal of it. That is a student's stated confidence for a given mathematics item provides only limited information about whether that student will, in fact, answer the item correctly. This distinction is important because contemporary research increasingly treats self-efficacy and confidence as related but non-equivalent constructs, with task-specific judgments potentially showing different degrees of correspondence with demonstrated performance (Russell & Phelps-Gregory, 2026; Cai et al., 2026). Evidence from secondary mathematics classrooms also suggests that confidence–accuracy calibration can vary across assessments rather than remaining perfectly aligned with achievement (Foster & Renie, 2024).

Third, and most strikingly, the calibration analysis revealed a substantial, systematic, and highly consistent pattern of overconfidence in this study. Mean self-reported confidence (65.0% on a rescaled 0–1 metric) exceeded mean observed accuracy (30.8%) by more than thirty percentage points on average, and this gap was present in the majority (93.7%) of individual students rather than being driven by a small number of extreme cases. Nearly two-thirds of high-confidence responses (ratings of 7 or higher) were nonetheless incorrect. This pattern echoes a broader literature on the dissociation between confidence and accuracy in academic self-assessment (Kruger & Dunning, 1999; Stankov et al., 2014) and suggests that, in the present study, students' anticipatory judgments of their mathematical capability were not merely imprecise but systematically inflated relative to their demonstrated performance. Recent mathematics education research provides a useful contemporary parallel: Foster and Renie (2024), studying secondary school mathematics students across repeated low-stakes assessments, explicitly examined confidence calibration as the correspondence between confidence and accuracy. Similarly, recent work on task-specific mathematics self-efficacy has emphasized that students' perceived capability can be incongruent with their demonstrated performance on particular mathematical tasks (Russell & Phelps-Gregory, 2026). These findings reinforce the importance of distinguishing the level of confidence from the accuracy of the underlying judgment.

These two sets of findings — a statistically real but modest positive relationship, combined with pronounced and pervasive miscalibration — are not contradictory. Rather, they illustrate precisely the conceptual distinction on which the study's framework was built. A confidence judgment can retain some discriminative value (correctly signaling, on average, that a student is somewhat more likely to succeed on tasks they feel more confident about) while nevertheless being poorly calibrated in an absolute sense (systematically overstating the likelihood of success at every level of the scale). It is entirely possible for both of these things to be true simultaneously, and the present data indicate that they were: task-specific confidence discriminated weakly but reliably between correct and incorrect responses, even as it dramatically overstated the overall probability of success. This distinction is also evident in recent work examining confidence calibration in mathematics, where confidence and accuracy are treated as related dimensions rather than interchangeable indicators (Foster & Renie, 2024). Research on mathematical self-efficacy similarly cautions against assuming that perceived capability necessarily corresponds closely to demonstrated performance on individual tasks (Russell & Phelps-Gregory, 2026).

Several explanations for this pattern merit consideration. One possibility is metacognitive: students in this study may have had genuine difficulty monitoring their own problem-solving processes accurately, particularly on unfamiliar or cognitively demanding items such as those involving logarithms, quadratic equations, or circle geometry, where a

plausible-seeming solution path can feel more certain than it is. Recent research continues to identify metacognitive monitoring and calibration as important components of students' ability to evaluate their own academic performance accurately (Alfaraj et al., 2026; Bol & Hacker, 2026). A second, complementary possibility concerns the difficulty of the assessment itself: with an overall mean of only 30.4% correct, the test was demanding relative to what students may have expected based on their classroom experience, producing a gap between anticipated and actual success that need not reflect a stable trait of overconfidence so much as a mismatch between the difficulty of everyday classroom tasks and that of the assessment items used here. Recent research on task-specific mathematics self-efficacy supports the importance of considering the particular characteristics of mathematical tasks when interpreting students' efficacy judgments rather than treating self-efficacy as entirely domain-general (Russell & Phelps-Gregory, 2026; Cai et al., 2026). A third possibility is motivational or social: students completing a self-efficacy rating immediately before attempting a problem, in a context where confident responding may be socially rewarded or where uncertainty is uncomfortable to express, may report confidence levels that are inflated relative to their private assessment of their own ability. Distinguishing among these explanations was beyond the scope of the present design, which did not include independent measures of metacognitive monitoring, prior achievement, or response style, but each represents a plausible avenue for follow-up research. Recent intervention-oriented work further suggests that structured feedback may influence both students' performance and the calibration of their self-assessments, making feedback an important direction for future investigation (Razeghpour & Rolka, 2026).

The gender-differentiated pattern observed in the exploratory analysis, essentially no confidence–performance association among male students but a modest, significant one among female students, is worth noting but should be interpreted cautiously. This is because it was not among the study's primary hypotheses and because the subgroup analysis was not adjusted for multiple comparisons. If replicated, however, it would be consistent with a body of prior research suggesting that the confidence–performance relationship in mathematics can differ by gender, sometimes reflecting differences in the tendency to over- or under-report confidence relative to demonstrated ability rather than differences in mathematical competence itself (Pajares & Miller, 1994). Recent research continues to examine gender differences in mathematics self-efficacy and confidence, including evidence that self-beliefs and demonstrated performance do not necessarily exhibit identical patterns across groups (Rutkowski et al., 2026). This pattern is best regarded, at this stage, as a hypothesis for future, purpose-designed research rather than as a confirmed finding of the present study.

6 Implications for Practice and Limitations

The findings suggest that the substantial confidence–accuracy gap observed in this sample has important implications for mathematics teaching and assessment. If the pattern generalizes, teachers may benefit from explicitly developing students' metacognitive monitoring and self-assessment skills, such as predicting success before solving problems and reflecting on the accuracy of those predictions afterward. Because confidence showed only a weak relationship with actual performance, teachers should also avoid treating students' self-reported confidence as a reliable indicator of readiness for more advanced mathematics. Particular attention should be given to high-confidence errors, which may reflect strongly held misconceptions and therefore require targeted formative feedback. Meanwhile, several limitations constrain interpretation. The cross-sectional, correlational design prevents causal conclusions and does not exclude influences from factors such as prior achievement, test anxiety, or academic self-concept. Although the mixed-effects model accounted for the dependence among students and items, richer models could incorporate item difficulty, prior achievement, and domain-specific effects. Finally, because participants came from a single educational context and completed a 15-item multiple-choice assessment, the extent to which the observed overconfidence generalizes to other populations, curricula, or assessment conditions remains uncertain. These limitations highlight the need for replication using larger, more diverse samples and longitudinal or experimental designs.

7 Conclusion

This study examined the relationship between task-specific mathematics self-efficacy and actual performance among 248 Nigerian senior secondary school students, using a design in which confidence was elicited immediately before each of 15 mathematics items was attempted. The analysis found that students' confidence showed a real but modest positive association with actual performance. At the same time, and more strikingly, students' confidence was substantially and systematically miscalibrated relative to their actual accuracy. The study's central contribution is to show that these two findings — a statistically detectable relationship and a pronounced calibration gap — are not in tension but are, in fact,

complementary pieces of evidence about the nature of mathematics self-efficacy in this study. Students' confidence carried some genuine information about their likelihood of success, but that confidence was, on the whole, considerably inflated relative to what students actually achieved. Future research designed explicitly to test the metacognitive, assessment-difficulty, and motivational explanations, would help clarify both the origins of this overconfidence pattern and the extent to which it can be productively addressed through classroom practice.

8 Declarations

Ethical statements: All the participants gave individual written consent to participate in the research.

Conflict of Interest: The authors declare neither financial nor non-financial conflict of interest.

Author Contributions: YFZ, REU, and SSM: Conceptualisation, Methodology, Supervision, Writing and Reviewing Original Draft. YFZ: Data Analysis. EVM and PG: Conceptualisation, Data Collection, Validation, Writing and Reviewing Original Draft. All authors approved the manuscript.

Funding: The research reported in this manuscript was funded by the Institute of Education, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.

Data Availability Statement: The dataset used for this study is available upon request from the corresponding authors.

References

- Alfaraj, R., Ozdemir, B., Aljuffali, L., Alsabhan, J. F., et al. (2026). Metacognitive awareness and confidence as predictors of academic performance in pharmacy students: Insights from grade predictions and structural equation modelling. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2026.1720303>
- Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman.
- Bol, L., & Hacker, D. J. (2026). Understanding and improving calibration accuracy in K-12 classrooms. *Theory Into Practice*, 1–14. <https://doi.org/10.1080/00405841.2026.2714715>
- Bozkus, K., & Öztürk, H. K. (2026). Mathematics self-efficacy and performance: A mediation analysis of PISA 2022 Turkey data. *European Journal of Education*. <https://doi.org/10.1111/ejed.70635>
- Cai, J., Liu, Q., Xu, B., & Hwang, S. (2026). Posing problems confidently: Task-specific self-efficacy and self-evaluation in mathematical problem posing. *Educational Studies in Mathematics*. <https://doi.org/10.1007/s10649-026-10484-7>
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3(4), 552–564. <https://doi.org/10.1037/0096-1523.3.4.552>
- Foster, C., & Renie, P. (2024). Changes in students' confidence calibration across a sequence of low-stakes confidence assessments. *Asian Journal for Mathematics Education*. <https://doi.org/10.1177/27527263241298968>
- 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), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66(4), 543–578. <https://doi.org/10.2307/1170653>
- Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, 86(2), 193–203. <https://doi.org/10.1037/0022-0663.86.2.193>
- Razeghpour, F., & Rolka, K. (2026). Effects of guiding feedback on students' performance, calibration, and self-efficacy: Insights from a field study with engineering students. *LUMAT: International Journal on Math, Science and Technology Education*. <https://journals.helsinki.fi/lumat/article/view/2692>

- Russell, B. E., & Phelps-Gregory, C. M. (2026). Examining task-specific incongruous mathematics self-efficacy in university intermediate algebra students. *International Journal of Research in Undergraduate Mathematics Education*. <https://doi.org/10.1007/s40753-024-00261-3>
- Rutkowski, D., Rutkowski, L., & Hastedt, D. (2026). Student achievement and self-confidence in mathematics and science: Evidence from TIMSS 2023. *Large-scale Assessments in Education*. <https://doi.org/10.1186/s40536-026-00304-y>
- Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: the best non-cognitive predictor of academic achievement? *Educational Psychology*, 34(1), 9–28. <https://doi.org/10.1080/01443410.2013.814194>.
- Yang, Y., Maeda, Y., & Gentry, M. (2024). The relationship between mathematics self-efficacy and mathematics achievement: multilevel analysis with NAEP 2019. *Large-scale Assessments in Education*, 12(1). <https://doi.org/10.1186/s40536-024-00204-z>
- Zakariya, Y. F. (2022). Cronbach's alpha in mathematics education research: Its appropriateness, overuse, and alternatives in estimating scale reliability. *Frontiers in Psychology*, 13, 1074430. <https://doi.org/10.3389/fpsyg.2022.1074430>
- Zakariya, Y. F., Goodchild, S., Bjørkestøl, K., & Nilsen, H. K. (2019). Calculus self-efficacy inventory: Its development and relationship with approaches to learning. *Education Sciences*, 9(3), 170. <https://doi.org/10.3390/educsci9030170>

Publisher's Note

The views, interpretations, and conclusions expressed in this article are solely those of the author(s) and do not necessarily reflect the views or policies of the Journal of Studies in Science and Mathematics Education (JOSSME), its publisher, editorial board, editors, reviewers, or the authors' affiliated institutions. The publisher remains neutral with respect to jurisdictional claims in published maps, institutional affiliations, and other representations. Any products, materials, methods, or manufacturers mentioned or evaluated in this article are presented solely for scholarly purposes and do not constitute endorsement, recommendation, or guarantee by the publisher.