INSIGHTS FROM EVALUATING ADOLESCENTS' METACOGNITIVE ACCURACY DURING A MATHEMATICS STANDARDIZED TEST

Madeleine Kingan, Jon Star

Harvard Graduate School of Education

Prior work demonstrates that learners are often over-confident in their performance. However, this type of monitoring may function differently when considering its developmental trajectory in mathematics. Using item response theory, we ask: (1) What was the relationship between confidence and accuracy on a mathematics standardized test in 5th to 8th grade students? (2) Were students differentially over-confident by age, over and above proficiency, at the item level? The findings demonstrated a positive relationship between confidence and accuracy. Older students were more over-confident than younger students and were often justified in this over-confidence. These findings demonstrate both consistencies and inconsistencies with prior domain-general, mathematics-specific, and developmental work.

INTRODUCTION

Metacognitively monitoring the state of, as well as the accuracy of, one's performance plays a key role in learning, allowing students to focus their attention towards the content on which they need to improve (e.g., Schraw, 2009). Prior work demonstrates that learners are frequently over-confident about the state of their knowledge (see Fleming, 2024 for a review), but there are several components of the learners' confidence and accuracy in mathematics that are less clear. First, while domain-general findings suggest that learners are over-confident, some studies find conflicting results in mathematics (e.g., C. Foster, 2022). Second, though some research demonstrates that metacognitive monitoring improves with age through adolescence (e.g., Weil et al., 2013), the developmental trajectory of the confidence-accuracy relationship with age remains unclear, particularly in the domain of mathematics (Schneider, 2008).

This paper aims to fill these gaps in the confidence-accuracy literature by examining this relationship in a mathematics standardized test administered to Romanian 5th to 8th grade students. Specifically, we ask: (1) What was the relationship between confidence and accuracy? (2) Were students differentially over-confident by age, over and above proficiency, at the item level? These questions were examined using item response theory (IRT), which enabled us investigate students' latent mathematics reasoning proficiency and confidence as well as precisely explore the role of item difficulty in the confidence-accuracy relationship.

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 827-834. PME.

PRIOR WORK ON THE CONFIDENCE-ACCURACY RELATIONSHIP

Prior research on the confidence-accuracy relationship finds that learners typically engage poorly in this type of monitoring (e.g., Fleming, 2024). In such studies, participants often complete tasks like general knowledge tests, picture and word recalls, and detail-recognition tests, assessing their confidence in the task as a whole or in individual questions. Though confidence and accuracy are often correlated positively with one another, people consistently report over-confidence, over-estimating their accuracy in learning and the assessment of learning (Fleming, 2024). Learners tend to overestimate their confidence more in more difficult content (e.g., Lichtenstein et al., 1981). Developmentally, domain-general work suggests that young children align their confidence and accuracy particularly poorly and improve with age (e.g., Lockl and Schneider, 2002; Weil et al., 2013). Consistent with domain-general findings, adult mathematics learners typically monitor their accuracy poorly, frequently reporting over-confidence (e.g., Kortenhoeven & Serra, 2025).

However, some studies of mathematics learning contradict or complicate the domain-general findings on the confidence-accuracy relationship. Several studies demonstrate that young people are better aligned in their confidence and accuracy than adults when monitoring their mathematics performance (e.g. C. Foster, 2022).

The domain-general and mathematics-specific research described above present an interesting disparity in the developmental trajectory of the confidence-accuracy relationship. The domain-general developmental work suggests that we can expect children to poorly align their confidence and accuracy and improve in this skill with age. Mathematics-specific research involving children, though, draws different conclusions. Children's mathematics confidence and accuracy are relatively well calibrated or poorly aligned in the opposite direction from domain-general findings (under-confident rather than over-confident), and adults are typically over-confident in mathematics. This suggests that confidence-accuracy alignment in mathematics either worsens or flips direction throughout development, whereas this alignment improves throughout development in domain-general work.

Differences in Methodological Approaches

Researchers have also employed a wide variety of methodological approaches to this issue. Most prior work evaluates this relationship at the person level, comparing average confidence and accuracy scores between people rather than classrooms (a more broad level) or items (a more specific level; e.g., Nolan & Markham, 1998). While this method allows a broad understanding of the confidence-accuracy relationship, it does not provide the level of quantitative detail needed to understand the minutia of this metacognitive monitoring mechanism.

Within the prior work using these person-level methodologies, scholars have also explored differences in confidence and accuracy with differences in task difficulty (e.g., Nietfeld et al., 2005). Most of these studies operationalize difficulty qualitatively

as determined by the authors (Kortenhoeven & Serra, 2025) or as discussed by participants (Rickey et al., 2025). Therefore, these studies can only provide evidence about the relationship between researcher or participant perceived difficulty and confidence, rather than the relationship between actual difficulty and confidence. Additionally, some studies do not explicitly define task difficulty at all (e.g., Nietfeld et al., 2005), leading to unclear results and implications.

To address these limitations, this paper uses IRT, which captures the relationship between confidence and accuracy at the item-level. By analyzing responses at the item level rather than the person level, it is possible to explore the relationship between confidence and accuracy between and within items, people, and overall scores. Rather than relying on participant-perceived difficulty or researcher-intended difficulty, IRT also allows estimation of "objective" item difficulty using exam proficiency. Additionally, we can use item difficulty estimates to examine the difference between confidence and accuracy given underlying, or latent, exam proficiency.

METHODS

Data

The data for the present study came from the Romanian Math Kangaroo exam, an individual mathematics standardized test competition. The exam consisted of 30 multiple choice mathematics items over 90 minutes. After each mathematics question students were asked "Are you sure of your answer?" Students could mark "yes" or could leave the item blank. The data consisted of 3,421 responses from 5th to 8th grade students. The test utilized a unique scoring system, which influenced the use of missing data. The absence of a response on an item was data that contributed to students' total scores, so we could not consider unanswered items as "missing." Therefore, unanswered observations were treated as zero (i.e. incorrectly answered).

Accuracy was operationalized as answering a mathematics item correctly (leaving an item blank counted as incorrect). Confidence was operationalized as answering (marking "yes") the prompt about sureness after each mathematics item.

Analysis Plan

To answer the first research question (What is the relationship between accuracy and confidence?), a one-parameter logistic model was fit to obtain b-parameter estimates for each item, which indicate an items' difficulty:

$$\log\left(\frac{P(X_{ip} = 1)}{1 - P(X_{ip} = 1)}\right) = \theta_p - b_i \; ; \; \theta_p \sim N(0,1)$$

Here $P(X_{ip} = 1)$ is the probability that person p answered a mathematics item i correctly, θ_p is person p's latent exam proficiency, and b_i is the difficulty of item i. Past work modeling this relationship with IRT similarly used one-parameter models focused on difficulty, assuming similar information parameters across items (Attali et

al., 2020). A Pearson's correlation coefficient was computed to assess the linear relationship between the item difficulty estimates (b-parameters) and item average confidence score (the percentage of test-takers who said they were sure about that item). This correlation describes the relationship between confidence and accuracy. This logistic model also provides θ estimates for further analysis.

For the second research question (Are students differentially over-confident by age, over and above proficiency, at the item level?), over-confidence was operationalized as the probability of a test taker answering that they were sure of their item response. To control for proficiency level, over-confidence was considered by grade level. The log of the odds of answering "Sure" was regressed on the one-parameter model θ estimates and grade level for each item:

$$\text{logit}(\text{'Sure'}) = \beta_0 + \beta_1(\hat{\theta}_p) + \beta_2(Grade\ Level)$$

The coefficient of interest is β_2, the difference in the log of the odds of answering "sure" on an item for every increase in grade level. These regression coefficients were compared for each item.

The analysis next considered whether students were justified in their over-confidence, controlling for exam proficiency. To investigate this, the log of the odds of answering mathematics items correctly was regressed on the one-parameter model θ estimates and grade level for each item:

$$\text{logit}(\text{Item Correct}) = \beta_0 + \beta_1(\hat{\theta}_p) + \beta_2(Grade\ Level)$$

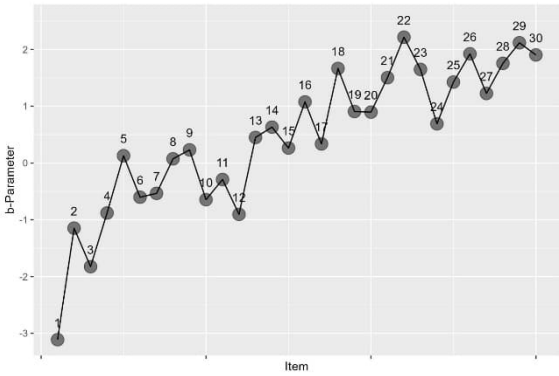
The coefficient of interest is β_2, the difference in the likelihood of answering an item correctly for every increase in grade. These coefficients were compared for each item. A Pearson's correlation coefficient was computed between the β_2 estimates from both regressions.

RESULTS

Students' Performance on the Exam as a Whole

The average accuracy for all students was .53. The average confidence for all students was .43. Cronbach's alpha for this exam was 0.937. The difficulty of each item, resulting from the first logistic model, is shown in Figure 1. The b-parameters for each item are the mathematical reasoning proficiency levels required for a student to have even odds (50% probability) of answering the item correctly. An item that required a lower proficiency was, by definition, less difficult than an item that required a higher proficiency to have even odds of answering correctly.

Figure 1: Item b-parameters



The average b-parameter was .438, indicating that students needed a mathematical reasoning proficiency .438 standard deviations above the mean to have a 50% chance of answering an average item correctly. There was a relatively large range of b-parameters, indicating that this exam's items had wide variety in difficulty throughout. There was a positive correlation between item number and b-parameter, meaning that, in general, items increased in difficulty throughout the exam.

Analysis of the Confidence-Accuracy Relationship

Regarding the first research question this study finds a positive relationship between confidence and accuracy on the Romanian Math Kangaroo exam. There was a strong negative correlation between item b-parameters and item confidence, $r(28) = -.993, p < .01$, indicating that test-takers reported less confidence on more difficult items. This in turn indicates a positive correlation between accuracy and confidence, since accuracy was incorporated in the difficulty calculations (i.e., by definition, students are less accurate on items that are more difficult). The highest residuals were all positive (items 2, 9, 18, and 22), indicating that the items that behaved most abnormally had higher reported confidence proportional to their difficulty than the other items.

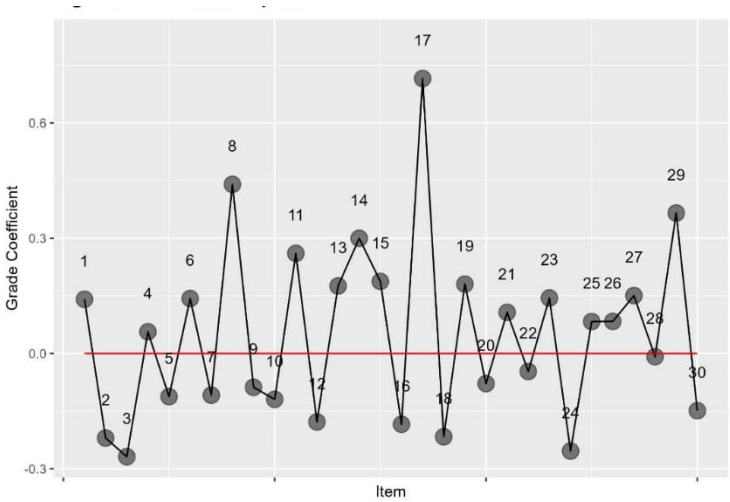
Analysis of Students' Over-Confidence

Regarding the second research question, older students were more often more over-confident than younger students (see Figure 2)and were somewhat justified in this over-confidence (see Figure 3). Each point in Figure 2 represents the coefficient of grade level in the logistic regression for a given item. A point with a coefficient of zero, which would be placed on the horizontal line on Figure 2, would indicate that students in older grades were no more or less likely to report confidence on that item than younger students, controlling for mathematical reasoning proficiency.

Figure 2 shows that students in older grades were more likely than students in younger grades to report confidence on 17 of the 30 items, over and above mathematics proficiency. Negative coefficients (below the horizontal line) indicate that older students were less likely to report confidence on that item than younger students, meaning they were less over-confident. Notably, there was no pattern between item

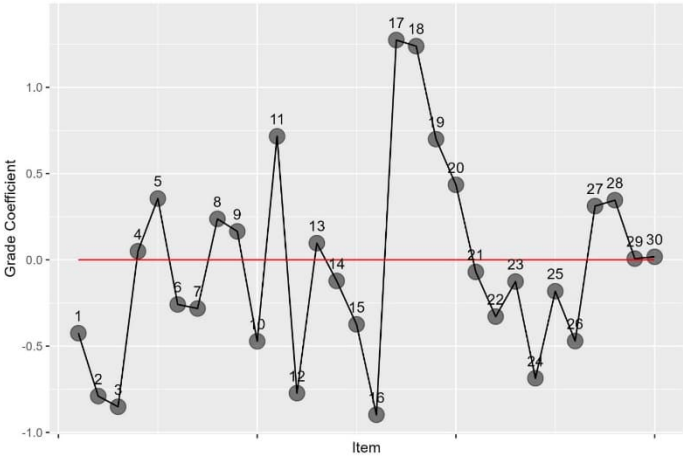
number, which correlates with difficulty, and over-confidence by grade, indicating that there was no pattern between item difficulty and over-confidence across the exam.

Figure 2: Overconfidence by Item; Difference in likelihood of older students than younger students reporting confidence



To explore whether students were justified in their over-confidence, a second logistic regression was fit for each item with accuracy as the dependent variable. The log of the odds of answering a mathematics item correctly was regressed on mathematical reasoning proficiency and grade level (see Figure 3). Each point in Figure 3 represents the coefficient of grade level in the regression for that item. A point representing an item with a coefficient of zero, which would be on the horizontal line, would indicate that students in older grades were no more or less likely than younger students to answer that item correctly, controlling for mathematical reasoning proficiency.

Figure 3: Accuracy Difference by Item; Difference in likelihood of older students than younger students answering accurately



Older students' over-confidence and accuracy aligned on 15 out of 30 items and eight of the 17 items for which they were over-confident, indicating justification of over-confidence. This is somewhat surprising – older students were more likely to answer over-confidently on more items than they were likely to answer over-accurately in

comparison to younger students. On these items, older students had a higher probability of answering confidently controlling for mathematical reasoning levels not necessarily because they were inherently more confident on that item, but perhaps because they were simply more likely to answer the item correctly even when controlling for overall exam performance.

The Pearson's correlation coefficient between the Grade coefficient in the confidence regression and in the accuracy regression was positive, $r(28)= .485$ ($p<.01$), indicating that there was a significant positive relationship between older students' overconfidence and older students' accuracy. This suggests that older students were overall moderately justified in their over-confidence. Older students' over-confidence, though here by definition unrelated to their overall mathematics proficiency, was moderately related to their mathematics performance by item.

DISCUSSION

This study contributes evidence of a positive relationship between confidence and accuracy. Students consistently reported over-confidence, a finding common in both domain-general and mathematics-specific work. This study's results are supportive of the prior mathematics-specific developmental trend and inconsistent with the domain-general developmental trend, as the findings demonstrate that younger students were more aligned in their confidence and accuracy than older students on this exam.

We consider several possible explanations for the conflicting results between domain-general developmental results and these mathematics-specific findings. First, perhaps age and mathematics interact in the confidence-accuracy relationship. Adolescents may improve in metacognitive monitoring overall, but the task characteristics present in mathematics may cause older students to over-think the complex cues. Second, this study used data from an understudied context - a standardized mathematics exam in which students were incentivized to well-align their confidence and accuracy. Perhaps older students felt more pressure on this exam due to the setting and/or the scoring system and therefore were less able to align their confidence and accuracy. Third, the contradictory findings could be related to the use of IRT to model this relationship. By using this method, this study uncovers the "objective" difficulty of each item and operationalizes confidence as independent of mathematics proficiency.

CONCLUSION

The current study leverages item response theory to provide evidence of a positive relationship between confidence and accuracy and increased over-confidence with age using data from the Romanian Math Kangaroo exam. These findings present novel evidence of the ways in which mathematical metacognitive monitoring functions, expanding and complicating prior work. Metacognitive monitoring appears to depend on context, age, and measurement methodology. The present study begins to shed light on this nuance, and the findings presented here imply the need for further research. For example, past evidence suggests that metacognitive monitoring improves with age, but

present results find that this may not be the case given the content, context, and use of IRT. If these results are replicated, this points to the need for renewed efforts to improve this important skill in mathematics throughout development.

References

Attali, Y., Budescu, D., & Arieli-Attali, M. (2020). An item response approach to calibration of confidence judgments. *Decision*, *7*(1), 1–19. https://doi.org/10.1037/dec0000111

Fleming, S. M. (2024). Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology*, *75*(1), 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Foster, C. (2022). Implementing Confidence Assessment in Low-Stakes, Formative Mathematics Assessments. *International Journal of Science and Mathematics Education*, *20*(7), 1411–1429. https://doi.org/10.1007/s10763-021-10207-9

Kortenhoeven, A., & Serra, M. J. (2025). Experience on a prior math task affects participants' judgments of learning (JOLs) on an unrelated memory task. *Metacognition and Learning*, *20*(1), 1–32. https://doi.org/10.1007/s11409-025-09417-1

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1981). *Calibration of Probabilities: The State of the Art to 1980* (No. PTR1092816). Article PTR1092816. https://apps.dtic.mil/sti/html/tr/ADA101986/

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom. *The Journal of Experimental Education*, *74*(1), 7–28.

Nolan, J., & Markham, R. (1998). The accuracy–confidence relationship in an eyewitness task: Anxiety as a modifier. *Applied Cognitive Psychology*, *12*(1), 43–54. https://doi.org/10.1002/(SICI)1099-0720(199802)12:1%253C43::AID-ACP487%253E3.0.CO;2-1

Rickey, N., Panadero, E., & DeLuca, C. (2025). How do students self-assess? Examining the metacognitive processes of student self-assessment. *Metacognition and Learning*, *20*(1), 1–29. https://doi.org/10.1007/s11409-025-09430-4

Schneider, W. (2008). The Development of Metacognitive Knowledge in Children and Adolescents: Major Trends and Implications for Education. *Mind, Brain, and Education*, *2*(3), 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In *Applied metacognition* (pp. 224–257). Cambridge University Press. https://doi.org/10.1017/CBO9780511489976.011

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004