Metacognitive errors in the classroom: The role of variability of past performance on exam prediction accuracy

Lisa Geraci1 · Nayantara Kurpad1 · Robert Tirso2 · Kathryn N. Gray3 · Yan Wang1

Received: 18 April 2022 / Accepted: 21 October 2022 / Published online: 9 November 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Students often make incorrect predictions about their exam performance, with the lowest-performing students showing the greatest inaccuracies in their predictions. The reasons why low-performing students make inaccurate predictions are not fully understood. In two studies, we tested the hypothesis that low-performing students erroneously predict their exam performance in part because their past performance varies considerably, yielding unreliable data from which to make their predictions. In contrast, high-performing students tend to have consistently high past performance that they can rely on to make relatively accurate predictions of future test performance. Results showed that across different exams (Study 1) and different courses (Study 2), low-performing students had more variable past performance than high-performing students. Further, results from Study 2 showed that variability in past course performance (but not past exam performance) was associated with poor calibration. Results suggest that variability in past performance may be one factor that contributes to low-performing students’ erroneous performance predictions.

Keywords Variability · Metacognition · Low performers · High performers · Predictions

Research shows that students often make inaccurate evaluations of their knowledge (Foster et al., 2017; Hacker et al., 2008). When this happens, they are said to have poor metacognitive monitoring, which is defined as the ability to accurately assess one’s knowledge. The general pattern of results shows that the lowest performers on the task are generally overconfident in their performance compared to the highest performers on the task, who are sometimes underconfident in their performance (Al-Harthy et al., 2015; de Bruin et al., 2017; Dunning et al., 2003; Foster et al., 2017; Hacker et al., 2000, 2008; Händel & Fritzsche, 2016; Miller & Geraci, 2011; Saenz et al., 2017; Serra & DeMarree, 2016; Tirso & Geraci, 2020; Tirso et al., 2019). In terms of magnitude of these errors, the lowest performers are more poorly calibrated than the highest performers, as there is a larger
difference between low performers’ performance predictions and their actual performance than there is between high performers’ performance predictions and their performance.

Calibration errors can be problematic, especially for individuals who overestimate their knowledge and abilities. Overestimations of ability may lead people to engage in challenges that they are not prepared to tackle or to perform tasks, even dangerous ones, that they cannot accomplish successfully. The potential problems of overconfidence are perhaps best highlighted by studies showing that people overestimate their gun safety knowledge (Stark & Sachau, 2016). But, overconfidence can lead to other serious (though less dangerous) problems, especially for students. In educational settings, being able to accurately monitor one’s current knowledge allows for an efficient learning process, such that students can use the information from their monitoring to inform their study methods and select poorly-understood topics for further study (Dunlosky et al., 2005; Dunlosky & Ariel, 2011; Dunlosky & Rawson, 2012; Thiede, 1999). This process breaks down for students who have poor calibration. For low-performing students, they err on the side of predicting that they know more than they do. The problem with erring on the side of overconfidence (rather than underconfidence) is that students may decide that additional preparation is unnecessary, leading them to perform poorly on exams and other assignments.

Because of the potential problems with inaccurate metacognitive monitoring, there’s been a great deal of research aimed at understanding the source of these metacognitive errors. For example, one theory suggests that metacognitive errors arise in response to poor learning processes especially in the low performers, whereby lack of knowledge of the material also yields lack of awareness (e.g., Kruger & Dunning, 1999). There are many reasons for this lack of knowledge and accompanying lack of awareness, including the possibility that students do not always engage in effective study strategies, such as self-testing, which would yield accurate knowledge (Karpicke & Roediger, 2008). There is also evidence that low-performing students are not particularly confident in their predictions (Miller & Geraci, 2011) and that they rely on other factors when making predictions, including the desire to perform well, which leads them to make high performance predictions (Saenz et al., 2017, 2019; Serra & DeMarree, 2016). According to the information-motivation theory (Tirso & Geraci, 2020), low performers err in the direction of overconfidence when they lack useful information because they opt to rely on motivational information to aid their predictions of future performance. This theory is consistent with prior research showing that people’s predictions about the future are often based on their aspirations rather than their past behaviors (e.g., Helzer & Dunning, 2012).

The current studies were designed to examine another possible contributor to poor calibration: variability in past performance. We hypothesized that low-performing students erroneously predict their exam performance in part because they rely on a history of variable past performance when making predictions. In other words, the lowest-performing students may have obtained some high grades and some low grades on past exams. Reliance on this type of noisy data may present low-performing students with a difficult signal extraction problem whereby they may be unsure which of their prior grades is most indicative of their true level of ability, opting to focus on their higher grades. In contrast, high-performing students may have more consistently high past performance than low-performing students. Thus, high-performing students would have more consistent and reliable information upon which to base predictions of future test performance.

There is some research that is consistent with this hypothesis. We know that, in general, the association between any two constructs can be attenuated by increased measurement
error in either or both constructs (e.g., Crocker & Algina, 1986; Worthen et al., 1999). That is, when observed variables (e.g., attention and academic performance) have random measurement errors, their relationship to each other is weaker than the relationship between the true scores of these variables. So, students with less consistent academic records can be thought of as having more measurement error than those with consistent records, whereby the reduced consistency in prior performance would attenuate any association between their true ability and their metacognitive judgments. Also, if we conceptualize prior performance as a cue that students use to varying degrees when making metacognitive judgments (e.g., Koriati, 1997), then that cue will be less reliable (and thus have less utility) for students with noisier histories.

We tested the hypotheses that low-performing students have more variable past performance than high-performing students and that this increased variability is associated with worse calibration. We focused on global predictions using a measure of absolute accuracy (how close a student’s overall exam score prediction is to the exam score itself), rather than on relative accuracy of individual test items, for example, which is often measured using gamma to correlate judgments with target performance (see Dunlosky & Thiede, 2013 for review). We chose to focus on global predictions of exam performance because we think these are common assessments that students naturally make, the accuracy of which can have important consequences for their study behaviors and assessments of the course or their major. In Study 1, we examined performance data from several undergraduate psychology courses to determine whether low-performing students’ exam grades were more variable than high-performing students’ exam grades. In Study 2, we examined whether variability in academic performance across students’ classroom tests and across their undergraduate careers predicted calibration on tests within a single course.

**Study 1**

To examine whether low-performing students showed greater variability in test performance than high-performing students, we examined variability in exam performance using data from several introductory and intermediate-level psychology courses that each included multiple exams.

**Method**

**Participants**

Six hundred and forty-nine undergraduate students’ deidentified class exam data from a large southwestern university were analyzed. The data were originally collected across multiple semesters from Introduction to Psychology and Cognitive Psychology courses, all taught by the same instructor. Because the policy for these classes was to allow students to drop their lowest exam score, many students opted not to take one of the four exams. After excluding these participants’ data, the final dataset consisted of 368 students with data for four exams. We chose to examine data from students with scores on all four exams (rather than just three exams) to obtain a more reliable measure of variability.
Materials and procedure

Data used in the current analyses were gathered from Introduction to Psychology and Cognitive Psychology courses offered across eight semesters (6 classes from fall semesters and 2 classes from spring semesters). In each course, students took four exams as part of the normal course requirements. The format of the exams was multiple-choice, and each exam consisted of 33 questions. The final exam in all courses was cumulative.

Students’ exam scores were compiled and de-identified. To examine whether variability in performance was greater for low-performing students compared to high-performing students, we sorted students into low- and high-performing groups based on their performance on a cumulative exam score (the total of all exam scores). This score was then used to divide the students into four quartiles, such that high performers’ average grades were above the 75th percentile and low performers’ average grades were in the 25th percentile and below. This process resulted in 93 low-performing students and 93 high-performing students.

Results

Variability was assessed using the standard deviation of each student’s exam scores (see Table 1 for students’ average exam scores by performance quartile). Based on examination of standard deviations, the results were consistent with our hypothesis that low performers exhibited more variability in their exam performance compared to high performers, $t(138.28) = 13.20, d = 1.935, p < 0.001$, 95% CI [6.8, 9.19] (Levene’s test for homogeneity of variance was violated, therefore, the corrected degrees of freedom are reported). One might worry that sorting students into high and low performers may ensure that variability will be lower in high performers due to a ceiling effect. Thus, it may be helpful to examine variability in performance for students performing in the upper middle-range, as they would be expected to be less variable than the low performers, but this pattern of reduced variability would not be attributable to a ceiling effect. Therefore, we also compared mean variability for high performers (above the 75th percentile), upper middle performers (between the 50th and 75th percentiles), and low performers (below the 25th percentile); see Table 2. Results showed that the high and upper middle performers had significantly lower mean variability than low performers (all $t$’s > 5.16, $p$’s < 0.05), which supports our hypothesis that performance and variability are negatively related.

Though our goal was to use courses with the best opportunity for observing performance variability—those with four exams—we note that the results are consistent using the larger sample, which included courses with just three exams in addition to those with four exams: Low performers had more variability in their exam performance compared to high performers (average variability of low performers = 19.06; average variability of high performers = 10.91; $t(322) = 77.991, d = 8.413, p < 0.001$, 95% CI [9.097, 7.727]).

---

1 A Satterthwaite approximation for the degrees of freedom was used as it has been shown to have adequate Type I error control and statistical power given mild violations of normality and unequal variances (Moser et al., 1989; Nguyen et al., 2016).
Study 2

Results from Study 1 supported our hypothesis that, relative to high-performing students, low-performing students exhibit greater variability in their exam grades across the semester. The goal of Study 2 was to determine whether variability in past performance is associated with prediction accuracy. We examined two types of variability: variability in past exam performance within a single course, similar to Study 1, and variability in past academic performance as measured by past course grades, across an entire university academic record. Students were recruited from six different classes over the course of a single semester. Just before taking each of their regular exams, students were asked to indicate how well they thought they would perform on the exam. These grade predictions were compared to actual performance on the tests. Variability of past performance was determined both by students’ performance across the exams in the class and by students’ past course performance across semesters based on their overall college record. We predicted that greater variability in students’ academic records would be associated with poorer metacognition.

Method

Participants

Data were collected from a new set of six different psychology classes at a large public southwestern university. These classes included Psychology of Women (Class 1), Introduction to Statistics (Class 2 and Class 3), Research Methods (Class 4), Psychology of Women of Color (Class 5), and Cognitive Neuroscience (Class 6). All six courses used a traditional in-person lecture-based format (all data collection occurred before the COVID-19 pandemic), and students in all six classes were offered extra credit in their course for participation in the study. Out of a total of 482 students, 440 consented to participate in the study.
Of these 440 students, 436 (99.1%) consented to release their exam grades to the study authors for inclusion in the analyses, and 395 (89.8%) consented to provide an electronic copy of their unofficial university transcript so that variability of past performance could be determined. Of these 395 students, only 76 (19.2%) provided a copy of their unofficial transcripts. Thus, Study 2 analyses measuring variability in performance across students’ academic careers were based on data provided by these 76 students. In contrast, all analyses that measured variability in performance across exams within a single course were based on data from 404 students—the number of students who had both exam and prediction data. Demographic data were not collected to keep the questionnaires administered to students prior to testing as short as possible.

Materials and procedure

All materials and procedures were compliant with the Family Educational Rights and Privacy Act (FERPA) and approved by the IRB. The researchers visited the six classes to recruit participants. Students were told that, should they choose to participate in the study, they would be asked to complete a brief questionnaire immediately prior to each of their course exams in exchange for extra credit in their course and that there would be no penalties if they ceased participation at any time. Students were also told that the data they provided by participating in the study would be kept by the research team instead of their instructor. Thus, their instructor would not know if they participated in the study. Next, consent forms were distributed, and students had the opportunity to opt in to release their exam grades and/or transcripts to the authors for inclusion in analyses; if students opted to release their transcripts, they were asked to provide an email address so that they could be contacted later to share a copy of their unofficial transcript. For all courses, this recruitment process took place within students’ classrooms before one of their regularly scheduled lectures.

After the initial recruitment session, research assistants visited each class on exam days to remind students of the study and administer the test prediction questionnaire. This questionnaire asked students to predict the numerical grade they would receive on the exam they were about to take. This questionnaire also asked students to predict what they thought the average, highest, and lowest grades on the exam would be, but these data were not used. All predictions were made on a 0–100% scale.

Classes varied in the number of exams, from three to five exams, and the dates on which their exams took place. Out of the six classes in Study 2, two classes (Class 1 and Class 5) had five exams, two classes had four exams (Class 2 and Class 3), one class had four exams with the fourth one being optional (Class 6), and the remaining class had three exams (Class 4). None of the exams were cumulative. See Table 3 for descriptive data.

Once the semester was over and final grades were posted, students who consented to be contacted for a copy of their transcript were sent an email that reminded them of the study and asked them to complete a 2-min Qualtrics survey that allowed them to securely upload a copy of their transcript. The university at which Study 2 took place allows students to download free electronic copies of their unofficial transcript. Before uploading their transcripts,

---

2 The questionnaire was given after the first exam of the research methods course (Class 4) instead of before due to a scheduling error. Analyses indicated this may have strengthened the association between predictions and performance, but this did not affect the overall pattern of results—the inclusion or exclusion of the exam one data from Class 4 did not noticeably change any results.
Table 3  Mean exam performance and grade predictions for Study 2

| Class                                | N   | Exam 1 Predicted | Actual   | Exam 2 Predicted | Actual   | Exam 3 Predicted | Actual   | Exam 4 Predicted | Actual   | Exam 5 Predicted | Actual   |
|--------------------------------------|-----|------------------|----------|------------------|----------|------------------|----------|------------------|----------|------------------|----------|
| Psychology of Women (1)              | 104 | 74.13 (9.89)     | 80.80 (7.81) | 71.47 (8.17)     | 79.65 (7.76) | 78.06 (8.10)     | 79.65 (7.88) | 81.46 (7.09)     | 82.43 (7.83) | 77.69 (8.02)     |
| Statistics (2)                       | 66  | -                | -        | 80.52 (7.51)     | 74.83 (12.59) | 75.22 (9.98)     | 72.41 (12.69) | 76.19 (7.841)    | 74.39 (10.28) | -                |
| Statistics (3)                       | 74  | -                | -        | 80.25 (7.55)     | 75.99 (12.19) | 76.46 (11.14)    | 72.41 (11.74) | 78.41 (9.62)     | 76.59 (11.89) | -                |
| Research Methods (4)                 | 65  | 84.97 (6.68)     | 79.80 (7.10) | 80.54 (16.43)    | 81.66 (11.27) | 82.15 (13.29)    | 86.14 (5.88)  | -                | -        | -                |
| Psychology of Women of Color (5)     | 32  | -                | 81.81 (8.98) | 83.61 (8.28)     | 71.08 (12.94) | 78.22 (15.23)    | 79.91 (8.00)  | 80.88 (7.52)     | 76.03 (10.31) | 79.16 (10.47)    | 81.41 (8.89) |
| Cognitive Neuroscience (6)           | 63  | 77.86 (12.54)    | 81.27 (20.69) | 74.48 (16.77)    | 74.98 (18.76) | 74.86 (14.36)    | 74.16 (16.50) | 74.52 (15.61)    | 77.02 (17.94) | -                |

Standard deviations are in parentheses
the students were informed that all identifying information from their transcripts would be redacted and assigned a code known only to the third author. Students were then given step-by-step instructions for downloading a copy of their transcript and uploading the file to the survey, after which the survey ended. Six reminder emails were sent to participants over the course of three months to boost response rates. Out of the 395 students that opted to provide a copy of their transcript, 365 students provided usable email addresses. Ultimately, 91 surveys were started, and 76 transcripts were successfully uploaded.

**Analytical plan**

**Variability in exam performance as measured within a single course** The analyses were based on data from 404 students—the number of participating students who had both exam and prediction data available—with differences in degrees of freedom being due to missing data. We began by testing whether low- and high-performing students differed in variability of prior exam grades before testing whether this variability moderated the relationship between grade predictions and actual grades. Low- and high-performing students were defined as students whose average grade fell within the first and third tertiles of the performance distribution. In Study 2 we sorted students into high and low performers based on a tertile split instead of a quartile split used in Study 1 because data were available from fewer participants since only 76 students provided transcripts. Similar to Study 1, variability of past performance was determined by calculating the standard deviation of students’ exam grades leading up to their final exam; thus, variability was based on performance from 2–3 prior exams. \( t \)-tests were first conducted to compare variability in past exam performance between low-, middle-, and high-performing students.

Next, multilevel modeling was used to investigate how variability in performance interacted with numerical grade predictions to predict actual exam grades. Multilevel modeling was used because the data were characterized by a multilevel structure in which students were nested within different classes—a 2-level structure. The outcome variable was students’ actual exam grades, which were on the scale of 0–100. Final exam prediction, prior exam variability, and prior exam performance were level 1 predictors, and course average exam grade (grand mean centered) was a level 2 predictor. Final exam prediction was on the scale of 0–100 and centered around their respective class means so that zero represents the average prediction for a given final exam. Thus, positive values represented grade predictions that were higher than the average predictions, whereas negative values represented grade predictions that were lower than the average prediction. Prior exam variability was transformed using a tertile split (0 = lowest variability, 1 = moderate variability, 2 = highest variability). Prior exam average was centered around the class means so that zero represents the average prior exam performance within a given class. Course average exam grade was conceptualized as a measure of course difficulty and calculated as the average of all exam grades across all students for each course. Adding course difficulty to the model allowed us to determine and account for whether differing exam grades across courses

---

3 Normality assumption was checked for multilevel models and the outcome was approximately normally distributed.

4 For Classes 1 and 5, grade data were available for the first exam even though no prediction data were collected for those exams (see Table 3). Nevertheless, these grade data were included in course difficulty calculations to obtain a more accurate estimate of overall course difficulty. Students’ average grade on their previous exams was also added to the analyses to account for the relationship between variability and grades because variability decreases as grades approach ceiling (\( r(402) = -.12, \ p = .017 \)).
influenced the relationship between variability and prediction accuracy, as it was conceivable that differences in the relative difficulty of the classes used in this study may have affected students’ grade prediction accuracy.

The model-building process was an iterative one, beginning with the empty, null model (Model 1.0) to serve as a baseline and calculating the intraclass correlation coefficients (ICC). The ICC indicates the percentage of variance in final exam grades accounted for by the clustering variables themselves. We then added additional predictors with each subsequent model and compared each model’s overall fit to the fit of the preceding model using a likelihood-ratio test. Random intercept was included in all multilevel models to allow for the variation in final exam grades across classes. All multilevel models were run in R version 3.6.3 (R Core Team, 2020) using RStudio version 1.2.5042 (RStudio Team, 2020), using the packages lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), merTools (Knowles & Frederick, 2019), jtools (Long, 2020), reghelper (Hughes, 2020), and ggplot2 (Wickham, 2016). All models also used restricted maximum likelihood estimation and \( p < 0.05 \) as the threshold for statistical significance. All \( p \) values were calculated using Satterthwaite’s formula for degrees of freedom.

**Variability in performance as measured across one’s college career** We also tested whether variability in performance across one’s academic career was associated with poor prediction accuracy. We started by testing whether low-, middle-, and high-performing students differed in the variability present in their GPAs before testing whether this variability moderated the relationship between grade predictions and actual grades. We used a multilevel modeling approach to address this question, but this time we used the variability in students’ grades across their college careers (GPA variability) as the measure of variability in performance instead of variability in their exam scores. We investigated whether GPA variability interacted with the relationship between grade predictions and actual performance across all exams. The models also accounted for the effects of course difficulty and the relationship between GPA and GPA variability (GPA variability decreases as GPA approaches ceiling). The analyses that follow were based on data from the 76 students for which transcript data were available.

Because we were interested in predictions across all exams (i.e., a longitudinal design), the structure of the models used differed from the 2-level design used in Model 1. Instead, the analyses that follow all made use of a 3-level design in which exams were nested within students, and students were nested within classes. One noteworthy advantage of using multilevel modeling with a longitudinal design was that it allowed us to make use of all available observations. Therefore, the following models were based on 246 observations (exam-prediction data points) from a sample of 76 students.

The modeling process began with the empty model (Model 2.0), with additional predictors being added with each subsequent model. The grades students received on each exam served as the dependent variable. Grade predictions were group mean centered based on the average prediction of all of the participants within each class for each exam so that a prediction of zero represented the average prediction on that exam. GPA variability was calculated as the standard deviation of each student’s GPA based on their transcript and transformed using a tertile split (0 = low variability, 1 = moderate variability, 2 = high variability). Course difficulty was the average exam grade for a given course as in Model 1, but overall GPA was substituted for prior exam performance to account for the fact that variability in GPA will decrease as GPA approaches ceiling. Grade predictions were entered at level 1, overall GPA and GPA variability were entered at level 2, and course difficulty was entered at level 3.
Results

Variability in exam performance as measured within a single course A pair of Welch’s $t$-tests indicated that low-performing students ($N = 137, M = 7.97, SD = 5.30$) had numerically, but not significantly, more variable past performance than middle- ($N = 132, M = 6.99, SD = 5.32$; $t(266.58) = 1.51, p = 0.133, d_r = 0.18$) and high-performing students ($N = 135, M = 6.88, SD = 5.66$; $t(268.31) = 1.63, p = 0.103, d_r = 0.20$; see Tables 3 and 4 for descriptive statistics). Full results of multilevel models can be found in Table 5. In the final model, Model 1.2, there were significant main effects of prior exam performance ($b = 0.62, SE = 0.04, t(364.82) = 15.12, p < 0.001$), course difficulty ($b = 1.37, SE = 0.32, t(5.85) = 4.32, p = 0.005$), and grade predictions on the final exam ($b = 0.19, SE = 0.08, t(365.19) = 2.54, p = 0.011$). The hypothesized interaction between predictions and variability was absent—there was no difference in the strength of the association between predictions and grades when comparing the low variability group to the moderate ($b = 0.04, SE = 0.12, t(366.39) = 0.34, p = 0.732$) or high ($b = -0.03, SE = 0.10, t(365.86) = -0.31, p = 0.759$) variability groups. Interestingly, this interaction was notably closer to the $p < 0.05$ cut off prior to adding prior exam performance to the model (Model 1.1b; $p = 0.053$). This finding suggests that, although it might appear that prediction resolution decreased as grade variability increased, this potential relationship is confounded by students’ prior grades because variability decreases as grades approach ceiling.

Variability in performance as measured across one’s college career A pair of Welch’s $t$-tests indicated that low performers ($M = 0.80, SD = 0.36$) possessed more variable past academic records than high performers ($M = 0.36, SD = 0.28$; $t(31.57) = 4.85, p < 0.001, d_r = 1.41$) but did not differ significantly from middle performers ($M = 0.70, SD = 0.28$; $t(32.68) = 1.00, p = 0.325, d_r = 0.30$). The multilevel modeling process began with the empty model (Model 2.0, ICC $\rho_{students} = 0.37, \rho_{classes} = 0.12$) and full model results are reported in Table 6. In the final model (Model 2.3), there were significant main effects of grade prediction ($b = 0.75, SE = 0.34, t(157.39) = 2.21, p = 0.028$), overall GPA ($b = 8.07, SE = 2.55, t(71.61) = 3.16, p = 0.002$), and course difficulty ($b = 1.10, SE = 0.35, t(7.92) = 3.17, p = 0.013$). Additionally, and most importantly, there was a significant interaction between grade predictions and GPA variability, showing that the association between predictions and performance was stronger for the lowest variability group than for the highest variability group ($b = -0.80, SE = 0.39, t(182.92) = -2.08, p = 0.039$; see Fig. 1).

| Table 4 | Mean variability in test scores and academic records by performance group in Study 2 |
|---------|--------------------------------------------------------------------------------------|
|         | $N$ | Prior Exam Variability | Academic Record (GPA) Variability |
| High Performers | 135 | 6.88 (5.66) | 0.36 (0.28) |
| Middle Performers | 132 | 6.99 (5.32) | 0.70 (0.28) |
| Low Performers | 137 | 7.97 (5.30) | 0.80 (0.36) |

Standard deviations are in parentheses
Table 5 Full results from Model 1 for Study 2

Model 1 Parameter Estimates

| Parameters | Model 1.0 | Model 1.1a*** | Model 1.1b | Model 1.1c*** | Model 1.2* |
|------------|-----------|---------------|------------|---------------|------------|
| Intercept  | 78.24 (1.68)*** | 78.24 (1.68)*** | 78.56 (1.84)*** | 78.17 (1.85)*** | 78.39 (1.07)*** |
| Final Exam Prediction | 0.29 (0.05)*** | 0.48 (0.10)*** | 0.20 (0.08)* | 0.15 (0.08)~ |
| Prior Exam Variability (low vs. moderate) | -0.80 (1.35) | 0.03 (1.06) | -0.10 (1.07) |
| Prior Exam Variability (low vs. high) | -0.68 (1.35) | 1.00 (1.09) | 0.71 (1.11) |
| Final Exam Prediction x Prior Exam Variability (low vs. moderate) | -0.26 (0.13)~ | 0.02 (0.11) | 0.04 (0.12) |
| Final Exam Prediction x Prior Exam Variability (low vs. high) | -0.25 (0.13)~ | -0.08 (0.10) | -0.03 (0.10) |
| Prior Exam Average | 0.63 (0.04) | 0.64 (0.04)*** |
| Prior Exam Average x Final Exam Prediction | -0.01 (0.01)* | -0.02 (0.01)* |
| Prior Exam Average x Final Exam Prediction x Prior Exam Variability (low vs. moderate) | 0.01 (0.01) | 0.01 (0.01)~ |
| Prior Exam Average x Final Exam Prediction x Prior Exam Variability (low vs. high) | 0.01 (0.01) | 0.01 (0.01) |
| Course Average | 1.37 (0.31)*** |
| Course Average x Final Exam Prediction | -0.04 (0.03) |
| Course Average x Final Exam Prediction x Prior Exam Variability (low vs. moderate) | 0.02 (0.05) |
| Course Average x Final Exam Prediction x Prior Exam Variability (low vs. high) | 0.04 (0.04) |

Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model

~$p < .10$, $*p < .05$, **$p < .01$, ***$p < .001$
Table 6  Full results from Model 2 for Study 2

| Parameters                                      | Model 2.0          | Model 2.1***         | Model 2.2a***        | Model 2.2b**         | Model 2.3 ~          |
|------------------------------------------------|--------------------|----------------------|----------------------|----------------------|----------------------|
| Intercept                                      | 79.25 (2.19)***    | 78.86 (2.05)***      | 82.01 (2.02)***      | 79.08 (2.20)***      | 79.02 (1.87)***      |
| Grade Prediction                               | 0.28 (0.07)***     | 0.64 (0.17)***       | 0.71 (0.35)*         | 0.75 (0.34)*         |
| GPA Variability (low vs. moderate)             | -3.18 (2.04)       | -1.21 (2.08)         | -0.42 (2.13)         |
| GPA Variability (low vs. high)                 | -8.35 (2.01)***    | -0.66 (3.05)         | -0.95 (3.01)         |
| Grade Prediction x GPA Variability (low vs. moderate) | -0.10 (0.23)       | -0.19 (0.38)         | -0.30 (0.37)         |
| Grade Prediction x GPA Variability (low vs. high) | -0.49 (0.19)**     | -0.75 (0.39)~        | -0.80 (0.39)*        |
| Overall GPA                                    | 9.50 (2.63)***     | 8.07 (2.55)**        |
| Overall GPA x Grade Prediction                 | -0.41 (0.70)       | -0.33 (0.69)         |
| Overall GPA x Grade Prediction x GPA Variability (low vs. moderate) | 0.80 (0.86)        | 0.73 (0.86)          |
| Overall GPA x Grade Prediction x GPA Variability (low vs. high) | 0.05 (0.79)        | -0.07 (0.78)         |
| Course Average                                 | 1.10 (0.35)*       |
| Course Average x Grade Prediction              | -0.07 (0.04)       |
| Course Average x Grade Prediction x GPA Variability (low vs. moderate) | 0.03 (0.08)        |
| Course Average x Grade Prediction x GPA Variability (low vs. high) | 0.10 (0.07)        |

Standard errors are in parentheses. Asterisks indicate the significance level of effects and improvement in overall model fit compared to the preceding model

~p < .10, *p < .05, **p < .01, ***p < .001
Discussion

In summary, we examined the relationship between variability in prior performance and grade prediction accuracy in two different ways using multilevel modeling—across tests within a single course and across all our students’ prior coursework. Within a single course, we found that (1) higher performance on earlier exams was associated with higher performance on the final exam, (2) higher course averages (in other words, easier courses) were associated with higher final exam grades, and (3) higher grade predictions on the final exam were associated with higher grades on the final exam (indicating that students had some degree of accuracy in their predictions, on average). Most importantly however, the hypothesized interaction between predictions and variability (within the span of a single course) was absent. However, this interaction was notably closer to the $p < 0.05$ cutoff prior to adding average prior exam performance (not variability) to the model (Model 1.1b; $p_s = 0.053$). This finding might suggest that, although it might appear that prediction resolution decreased as grade variability increased, this potential relationship is confounded by students’ prior grades and this ceiling effect.

When variability in performance was measured across courses, greater variability in the academic record was associated with reduced grade prediction accuracy. The main effects we observed meant that (1) higher grade predictions were associated with higher actual grades on the exams, (2) higher GPAs were associated with higher exam grades, and (3) higher course averages (easier courses) were associated with higher exam grades. Relevant to our predictions, there was a significant interaction between grade predictions and GPA variability. This interaction showed that the association between predictions and performance was stronger for the lowest variability group than for the highest variability group (see Fig. 1). This finding is consistent with our hypothesis that variability in past performance, which is primarily a problem for low-performing students, leads to worse metacognitive accuracy, at least when variability is defined across one’s academic career. It is
important to note that in the final model, the interaction between GPA variability and grade predictions remained significant even after accounting for students’ overall GPAs despite the clear relationship between GPA and GPA variability (as GPA approaches ceiling, GPA variability must approach zero). This finding suggests that GPA variability provides incremental validity in predicting the relationship between students’ grade predictions and actual grades that is not captured by a student’s average level of ability; this finding also highlights the need to disentangle variability in performance from overall performance.

General discussion

Two studies were designed to test the hypotheses that low-performing students have greater variability in their past performance compared to high-performing students and that increased variability is associated with poor calibration. These hypotheses were tested by examining the variability in students’ past performance across exams within a single course (Studies 1 and 2) and across courses within an academic career (Study 2). Results showed that across different exams (Study 1) and different courses (Study 2), low-performing students had more variable past performance than high-performing students. Increased variability in performance across one’s academic career (but not across exams within a course) was associated with increased errors on future exam and class predictions in Study 2. Thus, results showed that low performers have more variable past performance than high performers, and that variability across one’s academic career is associated with increased prediction errors.

The results showing that variability in performance across tests is not associated with increased calibration errors might have arisen because variability was somewhat limited in Study 2. For example, some courses had only three exams, thus limiting our estimates of variability to just the two exams prior to the one for which students made predictions. Thus, it is possible that using more tests, one would be able to observe greater variability in exam performance, which would be associated with worse calibration. Supporting this hypothesis, the average variability in test performance within a single course was higher in Study 1, using four tests, than it was in Study 2 using between two and five tests. In Study 1, the average variability was 13.87 for low performers and 5.88 for high performers, whereas in Study 2, it was 7.97 for low performers and 6.88 for high performers.

Of course, it is possible that variability within a single course is simply not associated with students’ prediction accuracy on a later test. Other studies (Foster et al., 2017) have examined students’ use of past performance in making predictions and found no relationship between students’ prior exam performance and their predictions. However, Foster et al. (2017) examined the role of the prior test performance from a single test and not variability of all prior tests on subsequent calibration. But, the fact that students in their study did not rely on their performance on the most recent prior test when making predictions suggests that students may not pay a lot of attention to their prior test performance within a single class when making future test performance predictions.

Results from Study 2 showed that increased variability across one’s academic career is associated with poor calibration. Low-performing students possessed far more varied academic records than high-performing students. High variability in one’s academic record was clearly associated with worse calibration, showing that this led to near zero ability to predict one’s performance (see Fig. 1). Given that low-performing students possessed more varied prior performance than high-performing students, and that variability in prior
course performance was related to calibration, variability in past performance might provide one explanation for why low-performing students exhibit poor calibration.

We offer one explanation for why low-performing students may have poor calibration in general. But, we note that the typical pattern is not simply one of poor calibration, but also one of over predicting performance. Thus, one may ask why low-performing students are inaccurate in the direction of overconfidence, rather than sometimes under predicting and sometimes over predicting their performance. The design of Study 2 does not allow us to determine the reason for the typical direction of the calibration errors. However, we speculate that, with less reliable information available and a tendency to attribute poor performance to external rather than internal causes (Hacker et al., 2000), low-performing students may turn to other sources of information when making their predictions. For example, research suggests that low performers report relying on motivational factors, such as the grades they wish to earn, more than high performers do, and this reliance on motivational factors predicts poor calibration in these students (see Serra & DeMarre, 2016; Saenz et al., 2017). Because low-performing students’ desired grades typically exceed their actual performance (Saenz et al., 2017), this would result in overconfidence, especially when combined with a lack of reliable information. Thus, consistent with the information-motivation theory (Tirso & Geraci, 2020), we suggest that in the face of impoverished information about prior performance, low performers may rely on other factors that lead them to err in the direction of overconfidence.

It is also possible that, with increased variability in past performance, low performers may choose which past performance they believe or hope is most representative of their abilities. That is, they may select to remember their past successes when predicting the future, choosing to focus on past “wins” rather than past “losses”, so to speak. This variability in past performance could allow people to focus more on their potential rather than their past, given that their past performance doesn’t yield a clear pattern of performance. Research shows that people tend to think of themselves in terms of who they want to be, instead of who they have been in the past (e.g., Williams & Gilovich, 2008), and their predictions about the future tend to be based on their aspirations (Helzer & Dunning, 2012). So it may be that, if there is a lot of variability in past performance, this pattern allows people to believe in future possibilities. Regardless of the potential mechanism for overconfidence, the current results show that low performers have more variable past performance than high performers and that the increased variability in one’s past academic record is associated with increased calibration errors.

As with all studies, this study has some limitations. For example, our ability to examine the relationship between variability of past exam performance and predictions was limited by the number of exams in each course in Study 2. Future studies should examine the role of variable past performance in situations where there are more tests to gain a better estimate of variability to investigate the potential relationship between past performance variability and prediction accuracy. We only examined global performance predictions (“What is your predicted grade”), but future studies might examine whether past performance variability influences local performance judgments (“What is the likelihood that this answer is correct?”). On the one hand, there is likely additional variability inherent in local judgments, as even the top-performing students will miss some items and experience variable item performance. Therefore, one might predict a smaller variability difference between high and low performers for local judgments compared to global judgments. Alternatively, one might predict that low performers would miss some items and get others correct, whereas high performers might get those items mostly correct, such that differences in variability for these groups of students would persist. One might also assess the relationship
between performance variability and postdictions (performance predictions that are taken after the test is completed). One hypothesis is that postdictions might be relatively unaffected by past performance variability because people would have recent diagnostic test experience to inform their judgments. Also, the current studies examined the potential role of past performance variability in prediction accuracy in a sample of undergraduate students, so we don’t know how the results would generalize to other student or community samples. Further, we did not measure the other factors that may influence prediction accuracy (e.g., course type, perceived course difficulty, type of prior courses, college major, test type: cumulative or not, etc.). For example, it would be interesting to examine whether performance across courses varies less within a students’ major compared to outside the major, and whether this influences the accuracy of performance predictions. Also, future studies might examine which courses students are considering when they are making assessments of future course performance. Future studies might examine the contribution of these and other factors to a potential association between performance variability and performance predictions. Finally, the sample size \(N=76\) was relatively small in Study 2 due to the limited number of students that provided copies of their unofficial transcripts. Future research is needed to replicate these findings using a larger sample.

There are several practical implications of these results. For example, the results have implications for the type of feedback one might provide to help low performers, in particular, improve their calibration. One suggestion would be to provide students with the full range of their past performance, drawing their attention to the low and the high grades. The results from Study 2 showed that variability in past academic performance predicts calibration, which may mean that students are selectively recalling or choosing which past grades to consider when predicting their future performance. Therefore, it may be important to provide students with the full range of their past performance to impede the potential for selective recall. Even if students do consider the full range of their past performance, they may determine that they will perform well on an upcoming exam because they have performed well before. After all, most students have performed well in at least one course. Therefore, it may be important to draw student’s attention to, not simply the full range of past performance, but also the frequency or average of their past performance. It may also be helpful to warn students about the potential pitfalls of variable past performance.

In addition to having implications for the type of feedback that might be effective for helping students improve their calibration, the current results offer another factor for instructors and institutional researchers to consider when determining which students might benefit from extra support, and what type of support they might need. The current results suggest that variability of past performance and not just overall GPA may signal a student in need of certain forms of academic support. A student with a 2.5 GPA with consistent performance may be quite different from a student with the same GPA but variable performance. Both students could likely benefit from extra support in the form of tutoring, learning about effective study strategies, and time management, but the student with consistently poor performance may have different challenges than the student with variable performance, and so the interventions for the two students might be different. Of course, future research is needed to directly test these hypotheses and explore the efficacy of potential interventions, but for now the current results suggest that variability in past performance is an important factor to target for future interventions.

In sum, we found that low-performing students had more variable past performance within a single course (Study 1) and across their academic careers (Study 2) than high-performing students. We also found that variability in one’s academic career (but not in a single course) was related to calibration on an upcoming exam. These results suggest that
variability in past performance may be an important factor that contributes to students’ metacognitive knowledge, particularly among low-performing students.

Author contributions  The first author conceived of the studies, contributed to the design of the studies, facilitated data collection for Study 1, and contributed to the writing of the manuscript. The second author contributed to study design and data analysis for Study 1, and to the writing of the manuscript. The third author contributed to the study design, data collection, and data analysis for Study 2, and to the writing of the manuscript. The fourth author contributed to the data scoring and data analysis of Study 2, and to manuscript preparation. The last author contributed to the data analysis, data interpretation, and writing of the results for both studies.

Declarations

Conflict of interest  The authors report no relevant financial or non-financial interests to disclose. The authors declare that they have no conflict of interest. This study was approved by the Texas A&M University Institutional Review Board (#2019-0783D).

Informed consent  Informed consent was obtained by all participants in the studies.

References

Al-Harthy, I. S., Was, C. A., & Hassan, A. S. (2015). Poor performers are poor predictors of performance and they know it: Can they improve their prediction accuracy. Journal of Global Research in Education and Social Science, 4(2), 93–100.

Bates, D., Maechler, M., & Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

de Bruin, A. B., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. Metacognition and Learning, 12(1), 21–43. https://doi.org/10.1007/s11409-016-9159-5

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. Psychology of Learning and Motivation (Vol. 54, pp. 103–140). Academic Press.

Dunlosky, J., Hertzog, C., Kennedy, M. R., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. Cognitive Technology, 10(1), 4–11.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students’ learning and retention. Learning and Instruction, 22(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students’ judgments can improve their achievement. Learning and Instruction, 24, 58–61.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. Current Directions in Psychological Science, 12(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. Metacognition and Learning, 12(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. Metacognition and Learning, 3(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. Journal of Educational Psychology, 92(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Helzer, E. G., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspirations versus past achievement. Journal of Personality and Social Psychology, 103(1), 38–53. https://doi.org/10.1037/a0028124
Händel, M., & Fritzschke, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. Memory & Cognition, 44(2), 229–241. https://doi.org/10.3758/s13421-015-0552-0

Hughes, J. (2020). Reghelper: Helper functions for regression analysis. R package version 0.3.6. https://CRAN.R-project.org/package=reghelper

Karpieke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. Science, 319(5865), 966–968. https://doi.org/10.1126/science.1152408

Knowles, J. E., & Frederick, C. (2019). merTools: Tools for analyzing mixed effect regression models. R package version 5.0. https://CRAN.R-project.org/package=merTools

Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

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/10.1037/0027-8399.77.6.1121

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). ImerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82, 1–26. https://doi.org/10.18637/jss.v082.i13

Long, J. A. (2020). Jtools: Analysis and presentation of social scientific data. R package version 2.1.0. https://CRAN.R-project.org/package=jtools

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. Metacognition and Learning, 6(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t test versus Satterthwaite’s approximate f test. Communications in Statistics: Theory and Methods, 18(11), 3963–3975. https://doi.org/10.1080/03610928908830135

Nguyen, D. T., Kim, E., Rodriguez de Gil, P., Kellermann, A., Chen, Y.-H., Kromrey, J. D., & Bellara, A. (2016) Parametric tests for two population means under normal and non-normal distributions. Journal of Modern Applied Statistical Methods, 15(1), 9. https://doi.org/10.22237/jmasm/1462075680.

RStudio Team (2020). RStudio: Integrated development for R. RStudio, Inc.. http://www.rstudio.com/

Saenz, G. D., Geraci, L., Miller, T. M., & Tirso, R. (2017). Metacognition in the classroom: The association between students’ exam predictions and their desired grades. Consciousness and Cognition, 51, 125–139. https://doi.org/10.1016/j.concog.2017.03.002

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. Applied Cognitive Psychology, 33(5), 918–929. https://doi.org/10.1002/acp.3556

Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students’ desired grades predict their biased grade predictions. Memory & Cognition, 44(7), 1127–1137. https://doi.org/10.3758/s13421-016-0624-9

Stark, E., & Sachau, D. (2016). Lake Wobegon’s guns: Overestimating our gun-related competences. Journal of Social and Political Psychology, 4(1). https://doi.org/10.5964/jss.v4i1.464.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. Psychonomic Bulletin & Review, 6(4), 662–667. https://doi.org/10.3758/BF03212976

Tirso, R., & Geraci, L. (2020). Taking another perspective on overconfidence in cognitive ability: A comparison of self and other metacognitive judgments. Journal of Memory and Language, 114, 1–14. https://doi.org/10.1016/j.jml.2020.104132

Tirso, R., Geraci, L., & Saenz, G. D. (2019). Examining underconfidence among high-performing students: A test of the false consensus hypothesis. Journal of Applied Research in Memory and Cognition, 8(2), 154–165. https://doi.org/10.1016/j.jarmac.2019.04.003

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag.

Williams, E. F., & Gilovich, T. (2008). Do people really believe they are above average? Journal of Experimental Social Psychology, 44, 1121–1128. https://doi.org/10.1016/j.jesp.2008.01.002

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). Measurement and assessment in schools (2nd ed.). Allyn & Bacon/Longman.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.