Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Anxiety and the calibration of self-evaluation against external performance evidence

Benjamin Roszipal

benjamin.roszipal@ustp.at

St. Pölten University of Applied Sciences https://orcid.org/0000-0003-4276-9507

Martin Ernst

St. Pölten University of Applied Sciences https://orcid.org/0009-0002-1075-6285

Manuel Schwanda

St. Pölten University of Applied Sciences https://orcid.org/0000-0003-1944-0623

Michael Wagner

St. Pölten University of Applied Sciences https://orcid.org/0000-0003-0982-2310

Research Article

Keywords: state anxiety, metacognitive calibration, self-evaluation, retrospective judgment, performance assessment

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11723441/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Self-evaluations formed after performance may reflect both external performance evidence and the affective state present when the judgment is made. Yet it remains unclear whether anxiety experienced after consequential performance is associated with calibration beyond anxiety already present beforehand. This prospective observational study included 111 students completing a standardized, summative Objective Structured Clinical Examination. State anxiety was assessed immediately before and after the examination. Retrospective self-evaluations across three competence domains were compared with examiner ratings of the same performance, yielding 333 domain-level observations. Linear mixed-effects models examined the association between post-performance anxiety and signed calibration bias while accounting for pre-performance anxiety, examiner-rated performance, competence domain, and participant-level covariates. Higher post-performance anxiety was associated with more negative calibration (b = − 0.089, 95% cluster-bootstrap confidence interval [− 0.126, − 0.055]), corresponding to an approximately 8.9-percentage-point shift towards self-underestimation per standard-deviation increase in anxiety. The estimate changed little after adjustment for pre-performance anxiety and remained negative in temporal-sensitivity analyses and across all three competence domains. The association was stronger for retrospective self-evaluations than for examiner ratings. Evidence for an association with absolute calibration error was weak, and pre-performance anxiety was not associated with examiner-rated performance. Post-performance anxiety was therefore associated primarily with the direction of self-evaluation relative to external performance evidence, rather than with clearly poorer externally rated performance or greater calibration error irrespective of direction. Because the design was observational, the findings do not establish whether anxiety influenced self-evaluation, resulted from performance appraisal, or reflected unmeasured shared antecedents.

Background

People routinely evaluate their performance before objective results or feedback become available. These evaluations can inform how they interpret later evidence, seek help, and regulate subsequent effort (Butler & Winne, 1995; Panadero et al., 2017). Models of metacognition and self-regulated learning therefore treat monitoring judgments as a basis for later control decisions (Nelson, 1990). When judgments are well calibrated, they may help people identify whether further practice or a change in strategy is warranted. Miscalibrated judgments may instead direct attention and effort away from areas in which they are needed. Self-evaluation is consequently a form of judgment under uncertainty: people must reconcile subjective experience with incomplete or delayed external evidence (Andrade, 2019).

Two broad possibilities may explain why self-evaluations diverge from observed performance. First, differences in self-evaluation may partly track genuine differences in performance. People can discriminate between relatively better and worse outcomes, although this metacognitive sensitivity varies across individuals and tasks (Fleming & Lau, 2014). A more negative judgment may therefore sometimes reflect weaker performance. Second, self-evaluation may be constructed from multiple cues, only some of which are diagnostic of performance (Koriat, 1997; Scheiter et al., 2020). The cue-

utilisation account proposes that metacognitive judgments depend on the cues available at the time of judgment rather than on direct access to memory or performance quality (Koriat, 1997). More recent research has applied this account beyond memory judgments. Subjective difficulty, effort, fluency, confidence, and affect are associated with appraisals but do not always vary proportionally with external performance evidence (Hoch et al., 2023; Jaeger & Fiorella, 2024; Scheiter et al., 2020; Van De Pol et al., 2024). These possibilities are not mutually exclusive. A psychological state may be related to observed performance, to its evaluation, or to both.

Metacognitive calibration provides a way to examine these possibilities. Retrospective judgments evaluate the quality of a completed action, and calibration describes their correspondence with an external performance indicator. Signed calibration bias distinguishes underestimation from overestimation, whereas absolute calibration error measures the magnitude of disagreement regardless of direction (Krebs et al., 2022; Naujoks et al., 2022). Both differ from metacognitive sensitivity, which concerns whether judgments discriminate between relatively better and worse performance (Fleming & Lau, 2014). These measures are related but not interchangeable. A person may systematically underestimate performance while remaining close to the external reference or show substantial disagreement without a consistent directional tendency. Moreover, an external rating provides a structured comparison rather than an error-free representation of performance; its interpretation depends on the reliability and validity of the criterion (Andrade, 2019). An affective state could therefore be associated with a downward shift in self-evaluation without a comparable increase in absolute error.

Affect-as-information accounts offer one explanation for such a pattern. They propose that people may use their current feelings as information when evaluating an ambiguous or self-referential target (Forgas, 1995; Schwarz & Clore, 1983). Cue-utilisation theory provides a more specific inferential account: affect is one experiential cue among several, and its role may depend on its perceived diagnosticity and weight (Efklides, 2006; Koriat, 1997). Subjective appraisals may therefore be internally consistent yet remain imperfectly aligned with objective indicators when they draw on heuristic cues (Scheiter et al., 2020). Accordingly, affect may be associated with the direction of self-evaluation without necessarily being related to monitoring accuracy in every sense.

Anxiety is particularly relevant because it can occur before and after an evaluative event. Test anxiety is associated with lower achievement, but meta-analytic findings show substantial heterogeneity and do not support a simple performance-deficit account (Von Der Embse et al., 2018). Achievement-emotion theory also distinguishes prospective emotions experienced before an outcome from retrospective emotions arising after success or failure has been appraised (Pekrun, 2006). Before performance, anxiety may relate to preparation, threat appraisal, attentional interference, avoidance, or performance itself. Experimental evidence also indicates that anxiety and high stakes can reduce the use of effective self-testing strategies (Liu et al., 2024). After performance, anxiety is part of the context in which the event is reconstructed and judged. It may therefore be related to self-evaluation through processes that differ from those operating before or during task performance. Research has linked anxiety around assessments with both performance and metacognitive accuracy (Silaj et al., 2021), while post-exam

processing may involve rumination and self-critical thought (Möcklinghoff et al., 2024). These temporal positions suggest that post-performance anxiety should be distinguished from anticipatory anxiety rather than treated as the same state measured later.

The association between anxiety and negative self-evaluation can therefore be interpreted in at least two ways. Anxiety may be associated with poorer performance, or with a more negative judgment of comparable performance. Negative achievement emotions have been linked to lower predictions and postdictions and to greater underconfidence, even when their associations with objective performance are smaller or inconsistent (Camacho-Morles et al., 2021; Prinz-Weiß et al., 2023). Test anxiety has also been related to comparison processes (Talić et al., 2024), metacognitive beliefs about worry and threat monitoring (Huntley et al., 2022, 2023), and habitual negative self-evaluation (Dundas & Nygård, 2024). Post-event rumination is likewise associated with anxiety and negative interpretations of completed evaluative situations (Edgar et al., 2024). This evidence is compatible with associations involving both performance and its subsequent evaluation. However, much of it concerns anticipated performance, classroom tests, self-reported outcomes, or experimentally constrained tasks. Less is known about whether anxiety experienced after a consequential, directly observed performance remains associated with self-evaluation when prior anxiety and external ratings of the same performance are considered. Complex performance settings are especially informative because judgments must integrate several behavioral and experiential cues rather than evidence from a single response (Mason et al., 2022).

The present prospective observational study examined these associations in a standardized, consequential performance setting by comparing participants' retrospective self-evaluations with independent examiner ratings of the same behavior. The primary aim was to assess whether post-performance state anxiety was associated with signed calibration bias after accounting for measured pre-performance state anxiety. Exploratory analyses examined whether the association with post-performance anxiety was expressed primarily in students' retrospective judgments rather than in examiner-rated performance of the same behavior; whether these associations differed across the three competence domains; and whether they also applied to absolute calibration error. The study sought to characterize the pattern and temporal specificity of these associations, without inferring their origins or causal direction.

Methods

Study design and assessment context

This prospective observational study examined the association between state anxiety and metacognitive calibration following a summative performance assessment. Data were collected during an Objective Structured Clinical Examination (OSCE) at the end of the second semester of a Bachelor of Nursing program. The study is reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines.

The OSCE represented a standardized, summative performance assessment in which students demonstrated knowledge, procedural skills, and communication-related competencies under observation. Performance information was available at both the station and competence-domain level. For the present study, analyses focused on three overall competence domains: professional knowledge/analytical skills (PKA), methodological/procedural skills (MP), and social/communication skills (SC).

The analytic approach linked students' retrospective judgments of their own performance with examiner-rated performance within the same competence domains. Examiner ratings therefore served as the external performance reference for the calibration analyses.

Participants and setting

Participants were first-year Bachelor of Nursing students eligible to complete the summative OSCE at the end of the second semester. Students were eligible for study inclusion if they were enrolled in the program, had completed the required preceding courses, were eligible to participate in the OSCE, and provided written informed consent. Students were excluded from the respective analyses if they withdrew their participation or if data required for a given analysis were unavailable. Missing data were primarily attributable to students not attending the OSCE on the examination date.

Because the primary analyses were conducted at the competence-domain level, each student could contribute up to three observations, corresponding to the PKA, MP, and SC domains. The available sample was determined by the size of the eligible student cohort, and all eligible students were invited to participate.

Performance assessment and examiner-rated performance

The OSCE comprised five stations selected from a pool of 25 fundamental nursing skills. Each station lasted 12 minutes and was conducted under standardized summative examination conditions. Students' performance at each station was evaluated using standardized, station-specific checklists. Checklist items were scored as 0 = not performed, 1 = partially performed, or 2 = fully performed. No additional global rating scale was used.

Checklist items were assigned to one of three competence domains (PKA, MP, and SC) according to the underlying assessment framework and expert consensus (Roszipal et al., 2025). Based on this item-to-domain mapping, the OSCE system generated both station-specific performance scores and overall examiner-rated scores for each competence domain. The present analyses used the overall domain-level examiner scores, which were scaled from 0 to 1.

Five examiners participated in the OSCE. Each station was assessed by one examiner, and examiners remained assigned to the same station throughout the examination. Before each examination day,

examiners received a standardized briefing from the examination coordinator. Examiners were blinded to students' self-ratings and anxiety scores. Because each station was rated by a single examiner and no duplicate ratings were obtained, interrater reliability could not be estimated. Examiner ratings were therefore treated as the structured external performance reference within the assessment context.

Procedure and timing of measurements

State anxiety and self-evaluation measures were collected shortly before and shortly after the OSCE. Pre-performance state anxiety was assessed during the opening session before students began the examination. Post-performance state anxiety and retrospective performance judgments were collected shortly after students had completed the OSCE during the closing session.

Post-performance state anxiety and retrospective performance judgments were collected within the same questionnaire. Students first completed the state anxiety measure and subsequently rated their own performance. At the time these judgments were made, students had not yet received their OSCE results, station-level scores, domain-level scores, or formal performance feedback. The calibration analyses therefore used retrospective performance judgments formed before externally provided outcome information was available.

Following completion of the post-OSCE questionnaire, students received access to their examination results through the OSCE software. This feedback included competence-domain scores as well as station-specific performance scores.

Measures

State anxiety

State anxiety was assessed immediately before and after the OSCE using the German version of the state scale of the State-Trait Anxiety Inventory (STAI-S; Laux et al., 1981; Spielberger et al., 1983). The complete 20-item state scale was administered at both measurement occasions. The STAI-S total score was used in all analyses. Higher scores indicate higher levels of current state anxiety. Internal consistency was estimated separately for the pre-performance and post-performance assessments using Cronbach's alpha.

Metacognitive performance judgments

Immediately after completing the OSCE, students provided retrospective judgments of their own performance. Judgments were recorded on a percentage scale ranging from 0% to 100% in 10-percentage-point increments. Students provided judgments for each individual station as well as overall judgments for the three competence domains (PKA, MP, and SC). The present analyses used the overall

domain-level judgments, which were rescaled to proportions ranging from 0 to 1. For the purposes of the present study, these immediate retrospective self-evaluations were conceptualized as metacognitive performance judgments. Specifically, they represented learners' judgments of the quality of their own immediately preceding performance and were therefore treated as monitoring judgments rather than as independent indicators of performance. Their correspondence with examiner-rated performance was examined through measures of calibration.

Metacognitive calibration measures

Signed calibration bias was defined as the difference between students' retrospective performance judgments and examiner-rated performance within the same competence domain (retrospective performance judgment − examiner-rated performance). Values below zero indicate underestimation of performance relative to the examiner rating, whereas values above zero indicate overestimation. A value of zero indicates correspondence between the student's judgment and the external performance reference. Signed calibration bias was defined a priori as the primary calibration outcome.

Absolute calibration error was calculated as the absolute value of the difference between students' retrospective performance judgments and examiner-rated performance (|self-rated performance − examiner-rated performance|). Absolute calibration error captures the magnitude of discrepancy between the learner-generated judgment and the external performance reference irrespective of whether performance was underestimated or overestimated. Higher values indicate greater absolute discrepancy. Absolute calibration error was treated as a secondary calibration outcome.

Statistical analysis

The study hypotheses and statistical analysis plan were preregistered prior to analysis. The preregistration specified the primary calibration analysis, the included covariates, and the planned secondary and contextual analyses. Descriptive statistics were calculated for pre-performance and post-performance state anxiety, retrospective performance judgments, examiner-rated performance, signed calibration bias, and absolute calibration error. Internal consistency of the STAI-S was estimated using Cronbach's alpha separately at each measurement occasion. Changes in state anxiety from before to after the OSCE were examined using paired-samples t tests and quantified using Cohen's dz. The correspondence between retrospective performance judgments and examiner-rated performance was additionally characterized using Pearson correlations, Lin's concordance correlation coefficients, and Bland-Altman analyses (Lin, 1989; Bland & Altman, 1986). Bland-Altman bias was calculated as the mean difference between students' judgments and examiner ratings, with limits of agreement defined as the mean difference ± 1.96 SD of the difference scores. These agreement analyses were used descriptively to characterize judgment-performance correspondence.

The primary inferential analysis examined whether post-performance state anxiety was associated with signed calibration bias. Because each student contributed repeated observations across the three

competence domains, linear mixed-effects models with participant-specific random intercepts were used to account for the dependency of observations within students. Signed calibration bias was regressed on post-performance state anxiety, examiner-rated performance, competence domain, age, gender, and prior healthcare training. Competence domain was modeled as a categorical fixed effect. Age, gender, and prior healthcare training were included as prespecified participant-level covariates in accordance with the preregistered analysis plan. Gender was self-reported and recorded as female or male; no separate measure of biological sex was collected. Examiner-rated performance was included to estimate the association between post-performance anxiety and calibration conditional on externally rated performance.

To further characterize whether the association with post-performance anxiety was expressed primarily in students' performance judgments or was similarly present in examiner-rated performance, a rater-dissociation model was estimated. Students' retrospective performance judgments and examiner ratings were jointly modeled as the outcome, and an interaction between rater type and post-performance state anxiety was included. The rater × anxiety interaction tested whether the association between post-performance anxiety and rating level differed between students' retrospective performance judgments and examiner ratings.

Potential variation across competence domains was examined by estimating domain-specific anxiety slopes and by comparing common-slope models with models including a post-performance anxiety × competence-domain interaction using likelihood-ratio tests. These analyses examined whether the association between post-performance anxiety and signed calibration bias was comparable across the three performance domains. Absolute calibration error was analyzed as a secondary calibration outcome using a corresponding mixed-effects model to examine whether post-performance anxiety was associated with the magnitude of judgment-performance discrepancy irrespective of its direction.

As a contextual analysis, examiner-rated performance was regressed on pre-performance state anxiety, competence domain, age, gender, and prior healthcare training. This analysis characterized the association between anxiety before the assessment and subsequently observed performance.

Continuous predictors were z-standardized before model estimation. Robustness of the key fixed-effect estimates was assessed using 5,000 successful cluster-bootstrap resamples at the participant level, thereby preserving the dependency structure of repeated competence-domain observations within students. Bootstrap 95% confidence intervals and bootstrap p values were calculated for the key model parameters. As an additional sensitivity analysis, the primary calibration model was re-estimated using a leave-one-participant-out procedure. Marginal and conditional R^2 values were reported to quantify the variance explained by the fixed effects and by the combination of fixed and random effects, respectively. Model diagnostics, additional agreement analyses, Bland-Altman summaries, domain-specific bootstrap estimates, and leave-one-participant-out results are provided in Online Resource 1.

Analyses were conducted in R version 4.5.0 (R Core Team, 2025). All tests were two-sided with $\alpha = .05$. Interpretation emphasized effect estimates, confidence intervals, and robustness across parametric and

bootstrap analyses rather than statistical significance alone.

Ethical considerations

The study was reviewed by the institutional review board of the participating university and received external approval from the responsible regional research ethics committee. The name of the committee and approval number are omitted from the blinded manuscript and provided separately with the submission materials. Participation in the study was voluntary and had no influence on students' grades or examination outcomes. All participants provided written informed consent before inclusion.

Study data were pseudonymized before analysis. The linkage file was stored separately with restricted access and was unavailable to the research team conducting the analyses. The study was conducted in accordance with the principles of the Declaration of Helsinki. No patient data or human tissue were involved.

AI-Assisted Writing

Large language model (LLM)-based tools (Claude, Anthropic; ChatGPT, OpenAI) were used to support the preparation of this manuscript with respect to logical consistency, readability, and language optimization. All content, interpretations, and conclusions are the sole responsibility of the authors, who reviewed and approved the final version of the manuscript in its entirety.

Results

Analysis sample and state anxiety

The analytic dataset comprised 118 students with pre-OSCE state anxiety data and 116 students with post-OSCE state anxiety data. Paired pre- and post-performance state anxiety scores were available for 116 students. Domain-level retrospective performance judgments, examiner-rated performance, paired state anxiety scores, and the covariates required for the primary calibration model were available for 111 students, yielding 333 competence-domain observations. In the primary calibration sample, students had a mean age of 23.77 years (SD = 4.92; range = 19–61); 67 students (60.4%) were female and 44 (39.6%) were male, and 63 (56.8%) reported prior healthcare training (Table 1).

Table 1

Analytic Sample and Participant Characteristics

Characteristic	n / value
Eligible students	123
Students with pre-OSCE STAI-S data	118
Students with post-OSCE STAI-S data	116
Students with paired pre/post STAI-S data	116
Students included in primary calibration model	111
Domain-level observations in primary calibration model	333
Age, years, M (SD)	23.77 (4.92)
Age range, years	19–61
Female, n (%)	67 (60.4)
Male, n (%)	44 (39.6)
Prior healthcare training, n (%)	63 (56.8)

Note. Percentages for demographic characteristics are based on the primary calibration sample (n = 111). STAI-S = State-Trait Anxiety Inventory, State scale; OSCE = Objective Structured Clinical Examination.

Internal consistency of the STAI-S was high at both measurement occasions (pre-performance: Cronbach's alpha = .91; post-performance: alpha = .94). State anxiety decreased from before the OSCE (M = 58.61, SD = 9.95) to after the OSCE (M = 51.16, SD = 14.03), mean change = -7.41, 95% CI [-9.77, -5.04], t(115) = -6.21, p < .001. The within-person effect was moderate, Cohen's dz = -0.58, with a 5,000-resample bootstrap 95% CI [-0.79, -0.39]. Individual and mean pre- to post-OSCE state anxiety scores are shown in Fig. 1.

Figure 1

State Anxiety Before and After the OSCE

Note

Thin lines represent individual participants. Points and error bars represent means and 95% confidence intervals.

Descriptive metacognitive judgments and calibration

Descriptive domain-level results are presented in Table 2. Mean signed calibration bias was − 0.028, 95% CI [-0.063, 0.007], for PKA; -0.045, 95% CI [-0.083, -0.008], for MP; and 0.066, 95% CI [0.029, 0.104], for SC.

Thus, on average, students slightly underestimated performance in PKA and MP and overestimated performance in SC. Agreement between retrospective performance judgments and examiner ratings was modest: Pearson correlations ranged from r = .184 to .350 and Lin's concordance correlation coefficients from .139 to .287. Scatterplots and Bland–Altman analyses are provided in Online Resource 1.

Table 2

Descriptive Statistics for Retrospective Performance Judgments, Examiner-Rated Performance, and Calibration by Competence Domain

Domain	Self M (SD)	External M (SD)	Signed bias M [95% CI]	Absolute error M	Pearson r	Lin CCC
PKA	0.695 (0.190)	0.723 (0.103)	-0.028 [-0.063, 0.007]	0.148	.308	.254
MP	0.686 (0.208)	0.731 (0.116)	-0.045 [-0.083, -0.008]	0.154	.350	.287
SC	0.822 (0.189)	0.755 (0.100)	0.066 [0.029, 0.104]	0.174	.184	.139

Note. Ratings are scaled from 0 to 1. Signed calibration bias = retrospective performance judgment - examiner-rated performance. Negative values indicate underestimation relative to examiner ratings; positive values indicate overestimation. PKA = professional knowledge/analytical skills; MP = methodological/procedural skills; SC = social/communication skills; CCC = concordance correlation coefficient.

Figure 2

Mean Self-Other Calibration by Competence Domain

Note

Calibration is defined as self-rated minus externally rated global performance. Negative values indicate self-underestimation. Error bars represent 95% confidence intervals.

Baseline-adjusted association between post-performance anxiety and signed calibration bias

The primary inferential model examined post-performance state anxiety as a predictor of signed calibration bias while accounting for measured pre-performance state anxiety, examiner-rated performance, competence domain, age, gender, and prior healthcare training. Higher post-performance anxiety was associated with a more negative signed calibration bias after baseline adjustment, b = -0.089, SE = 0.016, t(109.88) = -5.54, p < .001, 95% cluster-bootstrap CI [-0.126, -0.055], bootstrap p < .001. Because signed calibration bias was defined as retrospective performance judgment minus examiner-rated performance, a 1-SD increase in post-performance anxiety corresponded to an

approximately 8.9-percentage-point shift toward greater underestimation relative to examiner-rated performance, conditional on pre-performance anxiety and the remaining covariates.

Measured pre-performance anxiety was not associated with signed calibration bias in the same model, b = 0.005, SE = 0.016, t(108.00) = 0.31, p = .756, 95% cluster-bootstrap CI [-0.023, 0.036], bootstrap p = .730. The primary model explained R2m = .337 of the variance through fixed effects and R2c = .735 when the participant-level random intercept was included.

To isolate the effect of baseline adjustment from changes in sample composition, the post-performance-only model was estimated in the exact same paired pre/post sample. The post-performance anxiety estimate in this same-sample model was b = -0.087, SE = 0.015, t(109.40) = -5.94, p < .001, 95% cluster-bootstrap CI [-0.123, -0.053], bootstrap p < .001. Adding pre-performance anxiety changed the focal coefficient only minimally (-0.087 to -0.089) and did not improve model fit, likelihood-ratio chi-square(1) = 0.10, p = .755. These results indicate that the observed post-performance anxiety association was not explained by measured pre-performance state anxiety.

Rater dissociation after adjustment for pre-performance anxiety

The baseline-adjusted rater-dissociation model tested whether post-performance anxiety was differently associated with students' retrospective judgments and examiner ratings. The rater type x post-performance anxiety interaction was negative, b = -0.072, SE = 0.010, t(559.08) = -7.30, p < .001, 95% cluster-bootstrap CI [-0.106, -0.038], bootstrap p < .001. For examiner ratings, the post-performance anxiety slope was small and negative, b = -0.023, 95% cluster-bootstrap CI [-0.044, -0.004], bootstrap p = .013. The interaction therefore indicates that the association with post-performance anxiety was an additional 7.2 percentage points per SD more negative for students' retrospective judgments than for examiner ratings. The corresponding rater type x pre-performance anxiety interaction was not supported, b = 0.002, 95% cluster-bootstrap CI [-0.032, 0.034], bootstrap p = .904. The model had R2m = .246 and R2c = .535.

Variation across competence domains

The baseline-adjusted association between post-performance anxiety and signed calibration bias was negative in each competence domain. Cluster-bootstrap domain-specific slopes were b = -0.097 for PKA, 95% CI [-0.136, -0.062], bootstrap p < .001; b = -0.100 for MP, 95% CI [-0.142, -0.061], bootstrap p < .001; and b = -0.072 for SC, 95% CI [-0.116, -0.032], bootstrap p < .001. Although the association was descriptively smaller in SC, allowing the post-performance anxiety slope to vary by competence domain did not significantly improve model fit relative to the common-slope model, likelihood-ratio chi-square(2) = 4.36, p = .113. Predicted calibration across post-performance anxiety by competence domain is shown in Fig. 3.

Figure 3

Predicted Self-Other Calibration Across Post-OSCE State Anxiety by Competence Domain

Note

Predictions are derived from the baseline-adjusted domain-heterogeneity model, with pre-OSCE state anxiety and external performance held at their reference values. Shaded bands represent 95% confidence intervals.

Temporal sensitivity analyses

Two complementary temporal-sensitivity analyses yielded the same directional pattern as the baseline-adjusted primary model. When change in state anxiety (post-performance minus pre-performance STAI-S) was used as the focal predictor, a larger increase, or smaller decrease, in anxiety was associated with a more negative signed calibration bias, b = -0.063, SE = 0.015, t(109.46) = -4.26, p < .001, 95% cluster-bootstrap CI [-0.092, -0.036], bootstrap p < .001. The change-score model had R2m = .279 and R2c = .738.

Likewise, residualized post-performance anxiety - representing post-performance state anxiety not linearly predictable from measured pre-performance state anxiety - was associated with a more negative signed calibration bias, b = -0.076, SE = 0.015, t(109.94) = -5.22, p < .001, 95% cluster-bootstrap CI [-0.109, -0.046], bootstrap p < .001. The residualized post-anxiety model had R2m = .310 and R2c = .736. Figure 4 summarizes the focal estimates across the temporally informed model specifications.

Figure 4

Key Fixed Effects From the Temporally Informed Anxiety and Calibration Models

Note

Points are fixed-effect estimates and bars represent 95% confidence intervals. The baseline-adjusted, change-score, and residualized post-anxiety specifications address complementary aspects of pre-to-post anxiety variation.

Absolute calibration error

As a secondary analysis, absolute calibration error was modeled after adjustment for pre-performance state anxiety and the remaining covariates. The point estimate for post-performance anxiety was small and positive, b = 0.021, SE = 0.010, t(111.97) = 2.01, parametric p = .046. However, the participant-level cluster-bootstrap interval included zero, 95% CI [-0.002, 0.043], bootstrap p = .069. Thus, evidence that

higher post-performance anxiety was associated with greater calibration error irrespective of direction was weak and not robust across inferential specifications. The model had R2m = .097 and R2c = .533.

Pre-performance anxiety and examiner-rated performance

As a contextual analysis, pre-performance state anxiety was not associated with examiner-rated domain-level performance after adjustment for competence domain, age, gender, and prior healthcare training, b = 0.001, SE = 0.009, t(111.00) = 0.07, p = .940, 95% cluster-bootstrap CI [-0.019, 0.023], bootstrap p = .940. Allowing this association to vary across competence domains did not improve model fit, likelihood-ratio chi-square(2) = 0.56, p = .756.

Robustness to influential participants

Leave-one-participant-out analyses supported the stability of the focal estimates. In the baseline-adjusted primary model, the post-performance anxiety coefficient remained negative in every iteration and ranged from − 0.098 to -0.082, with no sign changes. Domain-specific post-performance anxiety slopes likewise remained negative in every leave-one-out iteration. The change-score effect ranged from − 0.069 to -0.058 and the residualized post-anxiety effect from − 0.083 to -0.070, again without sign changes. Full leave-one-participant-out results are provided in Online Resource 1.

Table 3

Summary of Key Inferential Results

Analysis	Focal effect	b	95% bootstrap CI	Bootstrap p	Direction / interpretation
Primary calibration	Post-STAI adjusted for pre-STAI	-0.089	[-0.126, -0.055]	< .001	Higher post-anxiety -> more underestimation
Primary calibration	Pre-STAI	0.005	[-0.023, 0.036]	.730	No clear baseline-anxiety association
Same-sample post-only	Post-STAI	-0.087	[-0.123, -0.053]	< .001	Virtually unchanged before baseline adjustment
Rater dissociation	Self vs external x post-STAI	-0.072	[-0.106, -0.038]	< .001	Stronger negative association for self-judgments
Temporal sensitivity	STAI change (post - pre)	-0.063	[-0.092, -0.036]	< .001	Relative increase in anxiety -> more underestimation
Temporal sensitivity	Residualized post-STAI	-0.076	[-0.109, -0.046]	< .001	Post-anxiety beyond pre-anxiety -> more underestimation
Absolute calibration	Post-STAI adjusted for pre-STAI	0.021	[-0.002, 0.043]	.069	Weak/non-robust evidence for larger absolute error
External performance	Pre-STAI	0.001	[-0.019, 0.023]	.940	No association with examiner-rated performance

Note. Continuous predictors were z-standardized. Outcome scores ranged from 0 to 1. Bootstrap confidence intervals and p values are based on 5,000 participant-level cluster-bootstrap resamples. For directional calibration, negative coefficients indicate a shift toward self-underestimation relative to examiner-rated performance.

Discussion

Higher post-performance state anxiety was associated with a more negative evaluation of performance relative to examiner ratings after measured pre-performance anxiety was included in the model. A one-standard-deviation increase in post-performance anxiety corresponded to an approximately 8.9-percentage-point shift towards self-underestimation, conditional on examiner-rated performance and the remaining covariates. The estimate changed little after baseline adjustment, and analyses using anxiety change and residualised post-performance anxiety showed the same directional pattern. The association was substantially stronger for participants' retrospective judgments than for examiner ratings and remained negative across all three competence domains. By contrast, evidence for an

association with absolute calibration error was weak and not robust across inferential specifications. Pre-performance anxiety was not associated with examiner-rated performance. These results together speak to the two interpretations of the anxiety–self-evaluation association outlined in the introduction: that higher anxiety reflects poorer performance, or that it is associated with a more negative appraisal of comparable performance.

This pattern is difficult to explain as a reflection of poorer performance alone. Test anxiety is often negatively associated with achievement, but the magnitude and consistency of this relationship vary across studies and assessment conditions (Von Der Embse et al., 2018). Anxiety has also been related to both performance and metacognitive accuracy in classroom assessments (Silaj et al., 2021). In the present study, however, pre-performance anxiety was unrelated to examiner-rated performance, and the association of post-performance anxiety was considerably more negative for self-evaluations than for examiner ratings. This pattern is more consistent with anxiety being related mainly to the evaluation of completed performance than with an equivalent difference in externally rated performance. It does not exclude a performance component, because examiner ratings showed a small negative association with post-performance anxiety, and it does not establish that anxiety influenced the judgment.

Adjustment for pre-performance anxiety provides limited but relevant temporal specificity. Achievement-emotion theory distinguishes prospective emotions from retrospective emotions arising after an outcome has been appraised (Pekrun, 2006), while longitudinal research indicates that metacognitive beliefs and anxiety before an examination can be related to later anxiety and performance (Huntley et al., 2022, 2023). In the present analyses, accounting for state anxiety immediately before the assessment did not materially alter the post-performance coefficient. The association was therefore not accounted for by measured differences in pre-performance state anxiety. This adjustment does not isolate the source of post-performance anxiety, however. Trait anxiety, prior knowledge, prospective metacognitive judgments, and other unmeasured pre-performance factors may have shaped both later anxiety and self-evaluation. Processes arising after performance may also have contributed (Möcklinghoff et al., 2024). The supported conclusion is consequently narrow: the association with calibration remained after measured pre-performance state anxiety was taken into account.

The contrast between signed bias and absolute error further specifies the finding. Post-performance anxiety was associated with a downward shift in self-evaluation, whereas its association with the overall magnitude of disagreement was small and the bootstrap interval included zero. More anxious participants were therefore not clearly less accurate in every metacognitive sense. Their judgments were more negative relative to the external reference. This distinction matters because metacognitive bias, error, and sensitivity describe different properties of judgment and need not vary together (Fleming & Lau, 2014). The directional pattern is consistent with evidence that negative achievement emotions are associated with lower postdictions and greater underconfidence, even when associations with objective performance are weaker or inconsistent (Camacho-Morles et al., 2021; Prinz-Weiß et al., 2023). The present findings extend this pattern to complex, directly observed behavior, but they do not show that anxiety reduced metacognitive ability.

Affect-as-information and cue-utilisation accounts provide one possible interpretation. Feelings can inform judgments when the target is ambiguous or self-referential, although their influence depends on how relevant those feelings appear to the judgment (Forgas, 1995; Schwarz & Clore, 1983). Cue-utilisation theory similarly proposes that metacognitive judgments are constructed from available cues rather than from direct access to performance quality (Koriat, 1997). Anxiety was measured immediately before the retrospective judgments and before participants received scores or formal feedback. It was therefore available alongside memories of particular actions, perceived mistakes, effort, difficulty, confidence, and anticipated evaluation. Research shows that such experiential cues can shape appraisals without corresponding closely to external performance indicators (Efklides, 2006; Hoch et al., 2023; Scheiter et al., 2020). However, the present study did not measure cue availability or cue weighting. It cannot determine whether anxiety entered the judgment as a cue, resulted from an unfavorable appraisal, or covaried with another pre- or post-performance process.

The findings can also be situated within broader accounts of anxiety and self-evaluation. Comparison-based models describe how internal and external standards may shape self-perceptions and test anxiety (Talić et al., 2024), whereas the self-regulatory executive function model focuses on metacognitive beliefs and cognitive-attentional processes that may maintain anxiety (Huntley et al., 2022, 2023). Post-event processing provides another plausible context: rumination and self-critical thought occur after examinations and are associated with state and trait test anxiety (Möcklinghoff et al., 2024). These accounts concern different mechanisms and were not tested directly here. Their relevance lies in identifying processes through which anxiety and retrospective evaluation could become associated without requiring a corresponding difference in external performance. The present study addresses only this point of convergence, not the mechanisms themselves.

Mean calibration differed across competence domains, whereas the anxiety-related association did not. Participants slightly underestimated performance in the professional knowledge and analytical and methodological and procedural domains, while overestimating performance in the social and communication domain. Domain differences in self-evaluation may reflect differences in standards, task features, or the diagnostic cues available to the judge (Roszipal et al., 2025; Van De Pol et al., 2024). Social and communication competencies may provide fewer discrete, observable behavioral anchors relative to procedural or knowledge-based domains, leaving students with less certain evidence against which to calibrate their judgments. When such anchors are sparse, appraisals may draw more heavily on perceived interactional ease or general social norms as implicit evaluative criteria; if these heuristic standards are more permissive than the structured assessment framework, overestimation relative to examiner ratings would follow. Whether the domain asymmetry observed here reflects differences in cue diagnosticity, in the specificity of the standard implicitly applied, or in some other feature of how social performance is reconstructed and judged warrants closer examination. Higher post-performance anxiety was nevertheless associated with more negative calibration in each domain, and allowing the slopes to vary did not improve model fit. Within this sample and assessment, the association was therefore not confined to one competence domain. The sample was not large enough, however, to rule out smaller domain-specific differences.

The findings suggest a cautious implication for feedback following evaluative performance. An initial self-evaluation may be more negative among people reporting greater post-performance anxiety, but anxiety should not be treated as evidence that the judgment is inaccurate. Nor do the data show that reducing anxiety would improve calibration. Feedback may instead help individuals compare the cues underlying their judgment with explicit criteria and external observations. Self-assessment is most useful when it is embedded in processes involving criteria, feedback, and opportunities for revision (Andrade, 2019; Yan et al., 2022). Rubrics, practice tests, and timely monitoring feedback can improve judgment accuracy in educational tasks, although their effects depend on the task and implementation (Krebs et al., 2022; Naujoks et al., 2022; Wang et al., 2024). Whether comparable approaches alter anxiety-related calibration in consequential performance settings remains an empirical question.

Several limitations constrain the interpretation. Most importantly, the observational design cannot identify the source of post-performance anxiety or determine whether it influenced retrospective judgment. Anxiety and self-evaluation were measured within the same post-performance session; questionnaire order does not establish causal precedence. Anxiety may have contributed to a more negative judgment, an unfavorable appraisal may have increased anxiety, or both may have reflected shared causes. Measuring pre-performance state anxiety addresses only one potential antecedent. The study did not assess trait anxiety, prior knowledge, prospective metacognitive judgments, perceived difficulty, confidence, or the cues used in self-evaluation. It also did not measure post-event processing, which is associated with anxiety and negative interpretations following socially evaluative events (Edgar et al., 2024; Möcklinghoff et al., 2024). Baseline adjustment therefore improves temporal differentiation but does not resolve these alternative explanations.

Further limitations concern the external criterion, level of analysis, and generalizability. Examiner ratings provided structured external evidence rather than error-free ground truth. Agreement with an assessor should be interpreted in relation to the reliability and validity of that criterion (Andrade, 2019). Each station was rated by one examiner, so interrater reliability and uncertainty in the criterion could not be estimated. The domain-level design provided three repeated observations per participant and did not permit analysis of item-level monitoring resolution. Participants came from one cohort, one degree program, and one summative assessment context. Although the focal estimates were stable across bootstrap and leave-one-participant-out analyses, the sample did not support strong conclusions about small interaction effects or population heterogeneity. Complex performance judgments may also depend on how evidence is distributed across tasks, time, and actions (Mason et al., 2022). Replication is therefore needed across institutions, occupations, assessment formats, and levels of expertise.

Future research should distinguish antecedents of post-performance anxiety from its possible contribution to self-evaluation. Designs incorporating trait anxiety, prior knowledge, prospective judgments, repeated state-anxiety measures, and post-event processing could separate enduring vulnerability, pre-performance appraisal, performance-specific affective change, and retrospective evaluation. Measuring perceived difficulty, confidence, effort, and recalled errors would help identify which cues account for the observed association. Experimental variation in cue availability could then

test whether directing attention towards diagnostic behavioral evidence changes anxiety-related calibration, as research on cue diagnosticity would suggest (Scheiter et al., 2020; Van De Pol et al., 2024). Multiple raters or independent video ratings would allow uncertainty in the external criterion to be modelled. Longitudinal follow-up could assess whether initial self-underestimation is associated with feedback uptake, help seeking, subsequent practice, or later performance.

Conclusion

In a consequential real-world performance setting, post-performance anxiety was associated with more negative retrospective self-evaluation relative to examiner ratings after measured pre-performance anxiety was taken into account. The association was stronger for self-evaluations than for examiner ratings, remained negative across domains, and was robust to alternative representations of temporal change and influence. Evidence for greater absolute calibration error was weak. The findings therefore identify post-performance anxiety as a correlate of the direction of calibrated self-evaluation, rather than as evidence of general monitoring inaccuracy or clearly poorer externally rated performance. They do not explain the origin of post-performance anxiety or establish its effect on metacognitive judgment. The causal direction, underlying processes, and generalizability of the association remain to be determined. Should this pattern replicate across assessment contexts and professional groups, the timing and framing of post-performance feedback may warrant closer consideration — in particular in settings where students' initial self-evaluations inform whether and how they engage with externally provided performance information before integrating it into subsequent learning.

Declarations

Declaration of Generative AI and AI-assisted technologies in the writing process

Portions of the manuscript were refined with the assistance of ChatGPT (OpenAI, San Francisco, CA, USA) and Claude (Anthropic, PBC) for language polishing and structural clarity. All AI-assisted output was critically reviewed and edited by the authors, who take full responsibility for the final content.

Study registration

This study was preregistered in an independent public repository. To preserve anonymity during double-blind review, the registration record and repository link are provided on the title page.

References

1. Andrade HL (2019) A Critical Review of Research on Student Self-Assessment. Front Educ 4:87. https://doi.org/10.3389/feduc.2019.00087
2. Butler DL, Winne PH (1995) Feedback and Self-Regulated Learning: A Theoretical Synthesis. Rev Educ Res 65(3):245–281. https://doi.org/10.3102/00346543065003245

3. Camacho-Morles J, Slemp GR, Pekrun R, Loderer K, Hou H, Oades LG (2021) Activity Achievement Emotions and Academic Performance: A Meta-analysis. Educational Psychol Rev 33(3):1051–1095. https://doi.org/10.1007/s10648-020-09585-3
4. Dundas I, Nygård I (2024) Mindfulness for test anxiety and negative self-evaluation in high school. Curr Psychol 43(38):30027–30037. https://doi.org/10.1007/s12144-024-06557-6
5. Edgar EV, Richards A, Castagna PJ, Bloch MH, Crowley MJ (2024) Post-event rumination and social anxiety: A systematic review and meta-analysis. J Psychiatr Res 173:87–97. https://doi.org/10.1016/j.jpsychires.2024.03.013
6. Efklides A (2006) Metacognition and affect: What can metacognitive experiences tell us about the learning process? Educational Res Rev 1(1):3–14. https://doi.org/10.1016/j.edurev.2005.11.001
7. Efklides A (2011) Interactions of Metacognition With Motivation and Affect in Self-Regulated Learning: The MASRL Model. Educational Psychol 46(1):6–25. https://doi.org/10.1080/00461520.2011.538645
8. Fleming SM, Lau HC (2014) How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00443
9. Forgas JP (1995) Mood and judgment: The affect infusion model (AIM). Psychol Bull 117(1):39–66. https://doi.org/10.1037/0033-2909.117.1.39
10. Hoch E, Sidi Y, Ackerman R, Hoogerheide V, Scheiter K (2023) Comparing Mental Effort, Difficulty, and Confidence Appraisals in Problem-Solving: A Metacognitive Perspective. Educational Psychol Rev 35(2):61. https://doi.org/10.1007/s10648-023-09779-5
11. Huntley CD, Young B, Smith CT, Jha V, Fisher PL (2022) Assessing metacognitive beliefs in test anxiety: Psychometric properties of the metacognitions questionnaire, 30 (MCQ-30) among university students. Curr Psychol 41(3):1425–1433. https://doi.org/10.1007/s12144-020-00662-y
12. Huntley CD, Young B, Tudur Smith C, Fisher PL (2023) Metacognitive beliefs predict test anxiety and examination performance. Front Educ 8:1051304. https://doi.org/10.3389/feduc.2023.1051304
13. Jaeger AJ, Fiorella L (2024) Metacognitive effects of instructional visuals: The role of cue use and judgment type. Metacognition Learn 19(1):249–291. https://doi.org/10.1007/s11409-023-09370-x
14. Koriat A (1997) Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. J Exp Psychol Gen 126(4):349–370. https://doi.org/10.1037/0096-3445.126.4.349
15. Krebs R, Rothstein B, Roelle J (2022) Rubrics enhance accuracy and reduce cognitive load in self-assessment. Metacognition Learn 17(2):627–650. https://doi.org/10.1007/s11409-022-09302-1
16. Liu S, Zhao W, Shanks DR, Hu X, Luo L, Yang C (2024) Effects of Test Anxiety on Self-Testing and Learning Performance. Educational Psychol Rev 36(2):59. https://doi.org/10.1007/s10648-024-09882-1
17. Mason LA, Thomas AK, Taylor HA (2022) On the proposed role of metacognition in environment learning: Recommendations for research. Cogn Research: Principles Implications 7(1):104. https://doi.org/10.1186/s41235-022-00454-x

18. Möcklinghoff S, Rapoport O, Neidhardt E (2024) Examining the role of post-event processing in test anxiety—Pilot testing in three student samples. Int J Educational Res Open 7:100363. https://doi.org/10.1016/j.ijedro.2024.100363
19. Naujoks N, Harder B, Händel M (2022) Testing pays off twice: Potentials of practice tests and feedback regarding exam performance and judgment accuracy. Metacognition Learn 17(2):479–498. https://doi.org/10.1007/s11409-022-09295-x
20. Nelson TO (1990) Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation*(Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5
21. Panadero E (2017) A Review of Self-regulated Learning: Six Models and Four Directions for Research. Front Psychol 8:422. https://doi.org/10.3389/fpsyg.2017.00422
22. Pekrun R (2006) The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. Educational Psychol Rev 18(4):315–341. https://doi.org/10.1007/s10648-006-9029-9
23. Prinz-Weiß A, Lukosiute L, Meyer M, Riedel J (2023) The role of achievement emotions for text comprehension and metacomprehension. Metacognition Learn 18(2):347–373. https://doi.org/10.1007/s11409-022-09331-w
24. Roszipal B, Szelesi G, Ernst M, Hoffelner A, Wagner M (2025) Competency development in early nursing training: A cross-sectional OSCE study of self-assessment versus examiner ratings. BMC Med Educ 25(1):1710. https://doi.org/10.1186/s12909-025-08279-0
25. Scheiter K, Ackerman R, Hoogerheide V (2020) Looking at Mental Effort Appraisals through a Metacognitive Lens: Are they Biased? Educational Psychol Rev 32(4):1003–1027. https://doi.org/10.1007/s10648-020-09555-9
26. Schwarz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. J Personal Soc Psychol 45(3):513–523. https://doi.org/10.1037/0022-3514.45.3.513
27. Silaj KM, Schwartz ST, Siegel ALM, Castel AD (2021) Test Anxiety and Metacognitive Performance in the Classroom. Educational Psychol Rev 33(4):1809–1834. https://doi.org/10.1007/s10648-021-09598-6
28. Talić I, Sparfeldt JR, Möller J, Renner K-H, Greiff S, Niepel C (2024) Social and dimensional comparison effects in general and domain-specific test anxiety: A nested factor modeling approach. Curr Psychol 43(11):10058–10074. https://doi.org/10.1007/s12144-023-04964-9
29. Van De Pol J, Rowan E, Janssen E, Van Gog T (2024) Effects of availability of diagnostic and non-diagnostic cues on the accuracy of teachers' judgments of students' text comprehension. Metacognition Learn 19(2):635–659. https://doi.org/10.1007/s11409-024-09383-0
30. Von Der Embse N, Jester D, Roy D, Post J (2018) Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. J Affect Disord 227:483–493. https://doi.org/10.1016/j.jad.2017.11.048

31. Wang Y, Sperling RA, Malcos JL (2024) Supporting college students' metacognitive monitoring in a biology course through practice and timely monitoring feedback. Metacognition Learn 19(3):1–40. https://doi.org/10.1007/s11409-024-09385-y
32. Yan Z, Lao H, Panadero E, Fernández-Castilla B, Yang L, Yang M (2022) Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. Educational Res Rev 37:100484. https://doi.org/10.1016/j.edurev.2022.100484