Are High School Students Accurate in Predicting Their AP Exam Scores?: Examining Inaccuracy and Overconfidence of Students' Predictions

Teresa Ober, Maxwell R. Hong, Matthew F. Carter, Alex S. Brodersen, Daniella Rebouças-Ju, Cheng Liu, & Ying Cheng

University of Notre Dame

Please address correspondence to: Dr. Ying Cheng, University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN 46556

Acknowledgements: This work was supported by the National Science Foundation CAREER award (Grant #DRL-1350787) to Dr. Ying Cheng. We would like to acknowledge students in the Learning Analytics and Measurement in Behavioral Sciences Lab for their reviews of drafts of this manuscript as well as their contributions to the broader discussion of the topic.

(Version: 14 November 2021)

Are High School Students Accurate in Predicting Their AP Exam Scores?: Examining Inaccuracy and Overconfidence of Students' Predictions

We examined whether students were accurate in predicting their test performance two testing contexts (low-stakes and high-stakes). The sample comprised U.S. high school students enrolled in an advanced placement (AP) statistics course during the 2017-2018 academic year (N=209; M_{age}=16.6 years). We found that even two months before taking the AP exam, a high stakes summative assessment, students were moderately accurate in predicting their actual scores ($\kappa_{weighted}$=.62). When the same variables were entered into models predicting inaccuracy and overconfidence bias, results did not provide evidence that age, gender, parental education, number of math classes previously taken, or course engagement accounted for variation in accuracy. Overconfidence bias differed between students enrolled at different schools. Results indicated that students' predictions of performance were positively associated with performance in both low- and high-stakes testing contexts. The findings shed light on ways to leverage students' self-assessment for learning.

keywords: self-assessment, statistics education, advanced placement, course engagement

Introduction

Self-assessment is thought to be highly subjective. It requires a one to accurately monitor and appraise their performance (Andrade, 2019) and such abilities can vary substantially from person to person (Winne, 1996). Past research on the accuracy of self-assessment among students in particular is mixed (Panadero et al., 2016; Sanchez et al., 2017). From early work in this area, Shaughnessy (1979) developed a metric for calibration derived from a ratio of judgements of performance relative to an estimate derived from actual performance on a task. While the earlier work of Shaughnessy (1979) found a positive association between students who were more confident in their actual and judged performance, more recent works suggests students generally suffer from some hubris in their estimates of performance. Foster et al. (2017) examined college students' accuracy of performance prediction across 13 exams and found that students not only tended to be overconfident but that the tendency to be overconfident in their predictions generally did not decline over each successive exam. One explanation for this effect is that students' desired performance appears likely to influence the accuracy of judgements (Händel & Bukowski, 2019). So robust is this effect that even when students are trained to reflect on their performance as a means to promote a more accurate calibration, there is little effect of such training (Nederhand et al., 2020). Importantly, most of these previous studies have considered judgments of performance in the contexts where the assessments of learning would like be considered low-stakes in that they have a relatively small bearing on decisions concerning students' academic progress.

Some research on judgements of performance in higher stakes contexts has provided evidence that students' self-monitoring skills are related to improved performance on final exam and overall course grades (Smith & Was, 2019). However, evidence has also found that student self-assessment, particularly in contexts where assessment is used for summative or high-stakes purposes, is considerably less reliable (Dunning et al., 2004). To the best of our knowledge, however, few studies to date have examined students' calibration of future performance on a high-stakes exam in an authentic academic context. Thus, there is reason to speculate whether self-assessment is accurate and associated with performance outcomes, and if so, whether it remains after accounting for other factors associated with performance. In the present study we sought to examine whether high school students are accurate in predicting their performance on the advanced placement (AP) Statistics exam. We further sought to determine whether additional factors related to the students and their learning environment might predict variation in performance, both in a low- and high-stakes testing context.

Factors Associated with Self-assessment and Performance

Accuracy of self-assessment is defined as the extent to which an individual is capable of predicting their knowledge or skills in relation to some criterion. Past research indicates that among certain factors related to an individual's demographic background (e.g., age, gender, parental education), the classroom learning environment (e.g., prior educational experience, interactions with teachers, engagement in learning), as well as the assessment context (e.g., low- or high-stakes) also influence self-assessment accuracy. Individuals' background knowledge (e.g., subject matter experience) and current behaviors (e.g., course engagement) are thought to influence both the inaccuracy and overconfidence bias of self-assessment, as well as

performance, on a given task.

Background characteristics

Age. Metacognitive abilities required for accurate self-assessment appear to develop progressively during childhood and adolescence (Paulus et al., 2014; Schneider, 2008; Weil et al., 2013). Though children as young as 7 years of age are capable of making decisions based on metacognitive judgements of learning, adults may be more skilled at clearly differentiating between types of strategies for their metacognitive monitoring (Tsalas et al., 2015). These findings implicate age as a predictor of self-assessment accuracy given evidence for the development of metacognitive ability.

Gender. Differences in the accuracy of self-assessment with respect to performance has been found to vary as a function of students' gender. Though some evidence suggests that males and females on average perform similarly in science subject areas, far fewer females tend to enroll in STEM courses or pursue a post-secondary STEM career path (Cheryan et al., 2017). The expectations and choices that drive this discrepancy is thought to emerge prior to college enrollment and may be related to an individual's interest and motivations (Wang & Degol, 2013), formed within the context of societal expectations (Stoet & Geary, 2018). For example, past research has found that female students tend to perceive that teachers have lower expectations of their math learning performance compared to male classmates (Lazarides & Watts, 2015). These perceptions potentially become internalised, leading female students to have lower expectations about their academic performance in math subject areas (MacPhee et al., 2013) which is likely to lead to decreased performance (Greene et al., 2004). As such, gender may be related to differences in both self-assessment accuracy and performance on exams in math and related subject areas.

Parental education. The influence of parental educational attainment on the development of self-assessment and performance in primary and secondary school appears to be complex. The parent socialization model describes how parents' expectations and behaviors indirectly influence the academic achievement of their children by way of early emerging differences in the child's expectations of ability (Davis-Kean, 2005; Eccles, 2005). Parents with higher educational attainment are more inclined to promote a set of values that emphasise higher academic achievement among children (Faas et al., 2013). Parental educational attainment is often treated as an indicator of family socio-economic status since it is associated with higher family income and earning potential (Sirin, 2005), as well as access to social capital and educational resources that increase access to opportunities for academic achievement (Faas et al., 2013). Thus, for several reasons, parental educational attainment is thought to be predictive of academic achievement and students' expectations of their performance.

Classroom learning environment

In addition to the factors related to the individual student, the classroom learning environment may influence the students' performance, as well as their ability to accurately self-assess. Such factors include students' prior experience in the subject area and engagement in the course.

Prior math experience. Some evidence indicates that greater math attainment correlates with more accurate self-assessment of math performance (Hosein & Harle, 2018). Past experience in the subject matter may influence the association between self-assessment and actual performance

in several ways. Early experience in a subject area such as math may provide the background knowledge necessary to make more accurate predictions of performance (Watts et al., 2014). As an alternative explanation, past course experience may influence one's self-efficacy in understanding the content knowledge. This may lead a student to either overpredict or underpredict their performance (Hendy et al., 2014). Therefore, prior experience in math and related subject areas is likely to uniquely contribute to the accuracy of self-assessment, as well as performance on a statistics exam.

Engagement in the course. The extent to which students are actively engaged in learning is also thought to influence their ability to self-assess performance. Greater engagement in a course appears to be closely linked with greater motivation for learning. Motivational beliefs, which encompasses self-efficacy beliefs, appear to be associated with efficiency in problem-solving (Hoffman & Spatariu, 2008). Prior research has shown that students' engagement is linked with performance in learning mathematics, even after accounting for prior math experience (Stevens et al., 2004) and other classroom processes (Greene et al., 2004). As such, course engagement is likely to influence both self-efficacy and self-assessment accuracy, as well as performance indicators of learning.

Assessment context. The assessment context can also have a distinctive influence on the accuracy of students' self-assessment and performance. High-stakes testing contexts are characterized by an evaluation that has some bearing on the examinee's progress, and are typically associated with a summative decision based on the examinee's score. By contrast, low-stakes testing contexts are most often not associated with formative decision-making such that examinee's scores are used to inform the examinee of individual progress. Students may be more inclined to experience undue worry in evaluative contexts (i.e., test anxiety; Brown, Forman, Herbert et al., 2011) that are high-stakes as opposed to low-stakes (Putwain, 2008), and greater levels of test anxiety are associated with diminished performance (Cassady & Johnson, 2002). Students' self-efficacy has also been found to be more strongly associated with achievement in low-stakes compared to high-stakes testing contexts for mathematics knowledge (Simzar et al., 2015). Past research also indicates that test anxiety is moderately associated with math anxiety (Dew & Galassi, 1983; Hembree, 1990), another negative correlate of math achievement (Namkung et al., 2019). While students completing tests in a low-stakes context may experience less test anxiety compared with high-stakes contexts, they may also experience lower levels of engagement knowing that there are minimal consequences associated with their performance (Wise & DeMars, 2005). However, some evidence indicates that effort on a low-stakes test may be independent of its perceived importance, and thus should be unaffected by the lack of consequences resulting from testing in a low-stakes context (Barry & Finney, 2016; Barry et al., 2010).

The Current Study

Though past work has contributed to an understanding of the accuracy of self-assessment and performance in math and related subject areas (see Schneider & Artelt, 2010, for a review), few studies have considered the association between self-assessment and performance on indicators of learning (see Tanner & Jones, 1994, for a qualitative inquiry). In the present study, we were particularly interested in examining the accuracy of self-assessment within a sample of high school students enrolled in AP Statistics. We further sought to determine whether students' predictions of performance was associated with their actual performance in both contexts after

accounting for variation attributable to certain background characteristics of the student, as well as the classroom learning context. While it is not clear the extent to which the findings from this particular context could generalize to other courses, even AP or statistics courses in particular, the setting of the present study is advantageous for studying self-assessment for three main reasons.

Firstly, advanced placement courses, such as those offered through the College Board AP or International Baccalaureate (IB) programs, offer high school students the opportunity to enroll in courses taught with college-level curriculum over the span of an academic year. Many students who enroll in advanced placement coursework are within one to two years from high school graduation and typically intend to pursue a college degree (Judson & Hobson, 2015). Students who receive a certain grade or above on the AP exam, offered once at the end of each academic year in May, may be eligible for college credit equivalent to a semester's coursework in the subject area. As of 2021, AP exams scores were recognized as college credit transfer eligible by nearly 2000 colleges and universities in the U.S., as well as over 60 institutions throughout the world (CollegeBoard, 2021). Depending on the college/university transfer policy, a score of either a 4 ("well-qualified") or 5 ("extremely qualified") is considered acceptable as transfer credit. Thus, the AP exam is considered high-stakes by many who complete it as it may have an influence on the coursework students complete during college.

Secondly, the opportunity to examine high school students' learning within the context of an AP course is particularly intriguing. Past research has indicated students' metacognitive and self-monitory abilities develop gradually during childhood and adolescence, and even continue to develop into adulthood (Weil et al., 2013). Thus, we may reasonably expect that typical high school students are still acquiring the abilities to apply metacognition for accurate self-assessment of performance. Given that many students enrolled in AP coursework are preparing for college entry, their accuracy in self-assessment in the context of an AP course may be reflective of their ability to self-assess upon first-semester college enrollment.

Thirdly, statistical literacy is regarded as an essential set of knowledge and skills which enable a person to critically evaluate and make informed decisions based on data and statistical arguments (Weiland, 2017). Statistical literacy is increasingly seen as an essential component of a grade school education (Ben-Zvi & Garfield, 2008). While formal mathematics education is widely acknowledged as an essential component of K-12 education, the extent of instruction students receive to aid their understanding of statistics may still be quite inconsistent (Jones et al., 2007). Thus, for many students enrolled in such a course, there is likely to be great unfamiliarity in applying a formal approach to interpret statistical phenomenon. As such, students may be less reliant on their background knowledge of the subject matter and may rely to a greater extent on their metacognitive abilities to accurately gauge progress in learning.

Research Questions

Within the context of an AP Statistics course, we sought to address the following research questions:

1. *What influences the accuracy of students' self-assessment?* Specifically, to what extent do factors associated with the individual's demographic background (age, gender, parental education) and the classroom learning environment (school, number of previous math classes taken, engagement in the course) account for variation in the accuracy of students' self-assessment, as indicated by the absolute distance between predicted and actual scores, and bias in terms of over- and under-confident predictions?

2. *Are students predictions associated with actual performance, as measured in both a low- and high-stakes test context?* Specifically, to what extent do students' predictions account for variation in actual test performance in both a low- and high-stakes test context, after taking into account the aforementioned variables?

A better understanding in this regard may provide insights such that instructors can reliably interpret student self-assessment for formative purposes.

Methods

Participants

Participants included students enrolled in AP Statistics courses during the 2017-2018 academic year who were invited to take part in a year-long study to develop an adaptive assessment of statistics ($N = 209$, *Mean age* = 16.6 years, *SD age* = 0.9 years, 53.5% female). Recruitment of students occurred by way of participating teachers. In order to be eligible to participate, students were required to provide documentation of assent/consent prior to study enrollment.

Measures

Background questionnaire

A self-report demographics questionnaire was administered which asked respondents to provide information such as their current age, gender, and parents' education. The respondents entered their age in years as a numeric response and indicated whether they were "male" or "female." In conducting the analysis, male was coded as "0" and female was coded as "1."

Respondents also indicated their parents' approximate level of education by selecting one of seven responses to a multiple-choice question ("*What is the highest level of education completed by your parent(s) or legal guardian(s)?*"). Response options were treated as ordinal factors and included the options shown in Table 1. Given the relatively small proportion of participants who indicated their parents had less than a four-year degree, we re-grouped categories of responses into ordered levels follows: (1) Did not obtain Bachelor's degree; (2) Bachelor's degree (B.A., B.S., etc.); (3) Master's degree (M.A., M.S., etc.); (4) Doctoral or professional degree (Ph.D., J.D., M.D., etc.). In the U.S., professional degrees are defined as those which meet the following criteria: (1) two or more years of college must be completed before entering the program; (2) a total of at least six academic years of college work (including both prior college work and that required during the length of the program) to complete the degree program; and (3) completion of all academic requirements is necessary to begin practice in a profession (NCES, 2021).

School. Information about the students' schools was collected during participant recruitment. Given that participants were enrolled in the same courses taught within the same schools, we were interested in accounting for this effect. Students were enrolled from six different high schools in the Midwestern United States.

Prior math experience. Respondents reported the number of mathematics classes they had previously taken in high school, not including the current statistics course in which they were enrolled. Responses greater than "8" were treated as outliers and removed from the data prior to analysis. Three cases indicated completing 21, 25, or 32 math courses in high school and were

thus removed from the sample. Given that very few participants indicated they had taken more than five math classes in high school (0 had taken six, 2 had taken seven, 1 had taken eight), these participants were binned into the same category as those who had responded they had taken five (n = 8).

Student engagement. Students' engagement in the course was measured with an instrument based on the Scale of Student Engagement in Statistics (SSE-S; Whitney et al., 2019). The term 'micro-engagement' is used to refer to engagement within the context of a specific course (Handelsman et al., 2005). The SSE-S consists of 24 items, of which eight items each reflect affective, behavioral, and cognitive dimensions of engagement. Given the present sample, the scale appeared to have acceptable reliability, α = .89 (95% CI = .87, .91). The sum score (after reverse coding negatively keyed items) was used as an indication of students' overall engagement within the context of the course.

Low-stakes test performance. Low-stakes exam performance consisted of a score derived from a practice comprehensive assessment suitable for AP or an introductory college-level statistics curriculum. The scores were computed based on a Rasch model, a type of item-response theory model that can be used to compute estimates of performance based on their responses and item parameters, such as the item's difficulty (see Embretson & Reise, 2000). The scores ranged between –2.62 and 1.88, with more positive values indicating greater ability.

High-stakes test performance. High stakes exam performance was indexed by the students' actual scores on the AP exam administered in May of the academic year in which students were enrolled in the course. The AP scores were provided to researchers by participating teachers. The AP exam scores consisted of numerical values ranging from 1 (lowest) to 5 (highest).

Predicted scores. Students were asked to predict their score on the high-stakes AP exam by responding numerically to the following question: "Even if you are not taking the AP statistics exam, what do you think your score will be on the AP statistics exam?" The question was asked approximately two months prior and within a two-week window of completing the actual exam. The scores consisted of numerical values ranging from 1 (lowest) to 5 (highest).

Procedure

Data Collection

During the 2017-2018 academic year, participants completed a series of self-report questionnaires along with a computerised practice assessment of statistics knowledge. Prior to data collection activities, Institutional Review Board approval was sought and granted by the corresponding authors' institution. The practice assessment was created to mimic the content composition and item types in the actual AP exam. Data from the background questionnaire and self-reported number of math classes previously taken were collected between late September and early November of the academic year. The self-reported measures for course engagement and students' predictions of their AP exam scores were collected in early March until early April. Finally, students' practice assessments were completed in early May and their final AP exam was completed in mid-May.

Analyses

We sought to determine which of the demographic and classroom variables accounted for the

inaccuracy of predictions, based on the distance between the actual and predicted scores, as well as the overconfidence bias in predictions. Inaccuracy was estimated as the absolute value of the distance between the actual and predicted AP score (| actual – predicted |). Overconfidence bias was indicated by predictions of AP scores that exceeded actual performance (–1), under-confidence bias was indicated by a predicted score that was less than actual performance (1), and instances of no bias where predictions were accurate was also coded (0). Two separate regression analyses were conducted with participants' inaccuracy and overconfidence bias scores as outcomes. Predictor variables consisted of those related to the individual background characteristics (i.e., age, gender, parental education), the classroom learning context (i.e., school, number of math classes previously taken, course engagement), and scores on the practice AP exam, which served as an estimate of students' proficiency in statistics.

We then examined factors that predict performance in a low-stakes test context and in a high-stakes test context, both before and after taking into account students' predictions of performance. Separate analyses were conducted using practice assessments scores as a performance outcome in a low-stakes test context, and scores from the actual AP exam as outcomes in a high-stakes test context. Block one variables consisted of those related to the individual background characteristics (i.e., age, gender, parental education) and the classroom context (i.e., school, number of math classes previously taken, course engagement). The students' prediction of their score two months prior to the exam was entered as the only additional predictor in block two. Tukey posthoc contrasts were conducted to compare the marginal means for each level of the ordinal predictors using the *emmeans* package in *R* (Lenth et al., 2019). All analyses were conducted in R version 3.6.1 (R Core Team, 2019).

Results

We were interested in examining the relation between students' predictions of their scores and their actual performance on the AP exam.

Descriptive Statistics

Table 1 shows the descriptive statistics for each of the measures based on a sample size of 209 respondents. The mean age of the sample was 16.6 years, with participants ranging between 14 to 18 years of age. The majority of participants indicated they were in their senior (64.7%) or junior (10.7%) year of high school, with approximately one quarter indicating they were in their sophomore year (24.6%). The sample was approximately evenly split between female (53.5%) and male (46.5%) participants. The majority of participants in the sample indicated they came from a household with parents who had obtained a four-year college degree or above (88.8%), with approximately a quarter indicating a parent had a doctoral or professional degree (25.2%).

Table 1. *Descriptive sample characteristics*

Variable	Levels (Factors Only)	M (SD) / Frequency (%)
Age (years)		*N*=187
M (*SD*)		*M*=16.6 (*SD*=0.9)

Biological Sex		*N*=187
N (%)	Female	100 (53.5%)
	Male	87 (46.5%)
Parent Educational		*N*=325
N (%)	Did not finish high school	6 (3.2%)
	High school diploma / G.E.D.	8 (4.3%)
	Attended college, no degree	1 (0.5%)
	Associate degree	6 (3.2%)
	Bachelor's degree	70 (37.4%)
	Master's degree	51 (27.3%)
	Doctoral/professional degree	45 (24.1%)
Math Classes Previously Taken		*N*=184
M (*SD*)		*M*=2.9 (*SD*=1.3)
Course Engagement (SSE-S sum)		*N*=209
M (*SD*)		*M*=80.9 (*SD*=12.5)
Practice AP Exam Score (z-score)		*N*=182
M (*SD*)		*M*= –.03 (*SD*=.79)

Table 2 shows the relative proportion of scores for predictions made approximately 2 months prior and within 2 weeks of the actual AP exam date, respectively, both before actual scores were known, along with the actual scores received on the AP exam. Compared with the sample size 2-months prior ($N = 209$), the sample size for predicted scores within a 2-week window ($N = 154$) was smaller by 55 participants. Over half (56.4%) of those 55 with missing values at the second time were from the same school. Despite a presumed missing not-at random mechanism, besides school, none of the other participant demographic, predicted scores, nor learning outcomes significantly differed between the sample with predicted scores 2-months prior to the exam and the subsample with predicted scores at the later time point.

The proportions of predicted scores at each level did not appear to vary extensively between the two time points. There was considerable agreement between the two score predictions, as evidenced by a moderate-to-large kappa (κ) statistic (see McHugh, 2012), κ with quadratic weight ($\kappa_{weighted}$) = .78, $z = 9.78$, $p < .001$. The correlation between the two predictions was also high ($r_{polychoric} = .87$; see Table 3).

Table 2. *Frequency of predicted and actual AP Statistics scores*

Score	Predicted AP Score (2 months prior)	Predicted AP Score (2-week window)	Actual AP Score
	$N = 209$	$N = 154$	$N = 209$
1	2 (1.0%)	1 (0.6%)	6 (2.9%)
2	7 (3.3%)	10 (6.5%)	29 (13.9%)
3	73 (34.9%)	47 (30.3%)	50 (23.9%)
4	75 (35.6%)	54 (34.8%)	54 (25.8%)
5	52 (24.9%)	43 (27.7%)	70 (33.5%)

Relations Between Predicted Scores and Learning Outcomes

Correlations between predicted and actual AP scores were moderate-to-strong, see Table 3. A Williams test (Steiger, 1980; Williams, 1959), which is used to evaluate differences between two polychoric correlations derived from one common variable, did not find a significant difference between correlations ($t = .94$, $p = .34$) of actual and predicted AP scores when predictions were made 2 months ($r_{polychoric} = .73$) prior to testing, or actual and predicted AP scores within a 2-week window of the exam ($r_{polychoric} = .69$). Note that a polychoric correlations are used for estimating the correlation between two presumably normally distributed continuous latent variables and tend to be larger than Pearson correlation coefficients.

Table 3. *Polychoric correlations between predicted and actual AP scores*

	1.	2.
1. Predicted AP Score (2 months prior)	-	
2. Predicted AP Score (2-week window)	.87	-
3. Actual AP Score	.73	.69

Note: All correlations significant at the $\alpha = .05$ level with Bonferroni adjusted critical value = .017.

Figure 1 and Figure 2 present confusion matrices of the predicted and actual scores, with darker shading indicating a cell that represents a greater proportion of response tendencies. Weighted (squared) kappa revealed moderate agreement between predicted scores and actual scores both 2 months prior ($\kappa_{weighted} = .64$, $z = 21.27$, $p < .001$) and within 2 weeks of the actual test ($\kappa_{weighted} = .61$, $z = 13.42$, $p < .001$).

Figure 1. Confusion matrix with cell values representing outcome frequencies showing predicted score 2 months prior (predicted score) and actual exam score (actual score)

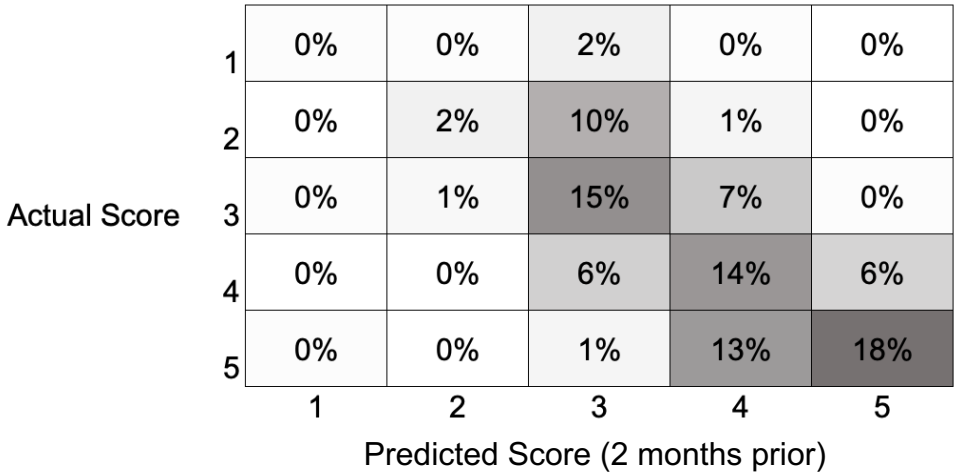
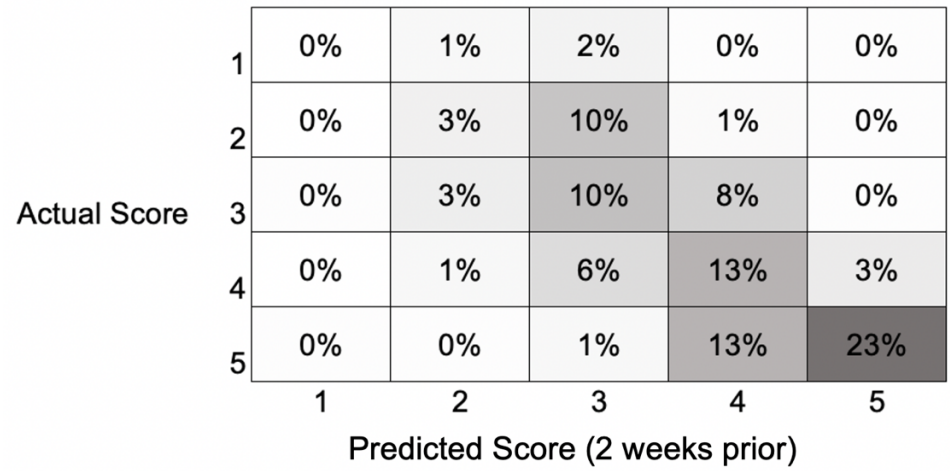


Figure 2. Confusion matrix with cell values representing outcome frequencies showing predicted score within 2 weeks of exam (predicted score) and actual exam score (actual score).



In light of the agreement between the two predictions and their similar correlations to actual performance, we opted to use the earlier predicted score to examine associations with performance indicators in the following analyses for at least two reasons. Firstly, many teachers dedicate substantial time in the spring semester review course material in preparation for the AP exam, and thus understanding the accuracy and bias earlier prediction of performance could be more useful to teachers and students. Secondly, due to attrition in the present sample, there was more data available for predicted scores measured at the earlier time point.

Prediction Inaccuracy and Bias

Inaccuracy: Distance between actual and predicted scores

To examine the association between predictors and the distance between actual and predicted scores, an ordinal logistic regression was conducted with Hessian approximation using the *MASS* package (Venables & Ripley, 2002) in *R*. None of the predictors were significant in accounting for variation in the distance between predicted and actual scores, see Table 4. Only two participants had distance scores of 3 or 4, with 4 being the maximum distance possible, and thus

were considered as potentially influential points. The results were robust, both with and without these two influential points. Excluding these two points, nearly half of the participants were accurate in predicting their scores and had a distance value of 0 (49.8%), others were off by 1 (44.5%) and a very small percentage were off by 2 points (5.7%)

Bias: Over- and under-confidence

We subsequently examined whether the same set of predictors explained students' overconfidence biases such that they either over- or under-estimated their performance. As noted, nearly half of participants were accurate, with a bias of 0 (49.8%) and the remainder were about equally split between tending towards overconfidence with a positive bias value (27.8%) or underconfidence with a negative bias value (22.5%). The results indicated a significant effect of enrolment within the same school, LR χ^2 = 29.52, p < .001, suggesting that there is some commonality in prediction bias among students enrolled in the same schools, see Table 4. The effect of course engagement was not statistically significant, LR χ^2 = 3.72, p = .054.

Table 4. *Variables accounting for prediction inaccuracy and bias (N=161)*

		Inaccuracy		*Bias*	
	df	*LR* χ^2	*p*	*LR* χ^2	*p*
1. School	5	1.84	.870	29.52***	<.001
2. Age (years)	1	<.001	.985	.15	.700
3. Gender (1=Female)	1	.04	.847	.86	.354
4. Parent Education	3	1.92	.590	1.49	.684
5. Math Classes Taken	4	4.55	.337	4.60	.331
6. Course Engagement (sum)	1	.61	.436	3.72†	.054
7. AP Practice Exam Score	1	.15	.701	.77	.380

*** p < .001, † p < .10

Associations between Predicted Score and Test Scores

Low-stakes test performance

Scores on a practice exam were used as indicators of performance in a low-stakes testing context, see Table 5. For $\eta^2_{partial}$, thresholds for small, moderate and large are 0.01, 0.06, and 0.14 (Cohen, 1988; Miles & Shevlin, 2001). Several notable effects were significant and suggested a weak-to-moderate effect size. Course engagement was significant (F(1,145) = 4.16, $\eta^2_{partial}$ =

.03, p = .038), indicating that course engagement was associated with performance in a low-stakes testing context. The students' gender also appeared to explain some variation in test performance (F(1, 145) = 5.24, $\eta^2_{partial}$ = .04, p = .023), with evidence suggesting that males (M = .17, SD = .80) tended to receive slightly higher scores on average than their female counterparts (M = –.11, SD = .82) when controlling for other factors. There was also a main effect for the categorical variable for school, indicating that differences in students' practice scores could partly be explained by the school they attended (F(1, 145) = 11.90, $\eta^2_{partial}$ = .29, p < .001). The effect of the number of math classes previously taken was not significant (F(4,145) = 2.28, $\eta^2_{partial}$ = .06, p = .063).

With the inclusion of students' predictions of their performance in block two, the pattern of significant effects changed slightly. Course engagement (F(1, 141) = 5.12, $\eta^2_{partial}$ = .04, p = .025), gender (F(1, 141) = 6.09, $\eta^2_{partial}$ = .04, p = .015), and school (F(1, 141) = 13.82, $\eta^2_{partial}$ = .33, p < .001), remained significant predictors. The number of math classes previously taken emerged as a significant predictor (F(1, 141) = 2.65, $\eta^2_{partial}$ = .07, p = .036). Contrasts revealed a significant difference in practice scores between students who had taken 2 (estimated marginal means =*emm* = .23) as compared to 3 (*emm* = –.41, t = 3.16, p = .016) or 4 previous math courses (*emm* = –.44, t = 2.90, p = .035). This indicates that students who reported only 2 previous high school math classes tended to perform better on the practice exam than those who reported taking either 3 or 4 high school math classes. Several factors may be driving this effect, including the requirements and recommended course sequence set by the schools. In support of this explanation, we found that the number of math courses previously taken was not independent of the effect of school (χ^2(df = 24) = 127.68, p < .001). Students who struggle with learning the material may also opt to enroll in more math classes for additional support. Age appeared to correlate significantly with number of math classes previously taken ($r_{polyserial}$(df = 183) = .757, p < .001), lending support for the explanation that younger and more accelerated students tended to perform better on the practice exam. Alternatively, the inclusion of predicted AP scores and the subsequent increased magnitude of the effect of math classes on the practice score is likely to indicate a suppression effect (see MacKinnon et al., 2000). Thus, the number of math classes, which may remain an indicator of subject matter background, is not necessarily a good indicator of better test performance.

Students' predictions of performance were significantly and positively associated with practice scores (F(1, 141) = 6.86, $\eta^2_{partial}$ = .04, p <.001). Contrasts revealed a significant difference in practice scores between students who predicted they would receive a 5 on the AP exam (*emm* = .34) compared with those who predicted receiving a 2 (*emm* = –.83, t = 3.34, p = .009), or 3 (*emm* = –.43, t = 4.86, p < .001) on the actual AP exam. The difference between students who predicted receiving a 5 as opposed to those who predicted receiving a 4 (*emm* = –.02, t = 2.70, p = .056) was not statistically significant. There was also a significant difference in the practice scores between students who predicted receiving a 4 on the AP exam (*emm* = –.02) as compared with a 3 (t = 3.12, p = .019). The contrasts generally showed that students who predicted doing well (i.e., a 4 or 5) tended to do better on the practice exam than their counterparts who predicted receiving a 3 or lower. At many selective colleges and universities, enrollees who receive of a score of a 4 or 5 on an AP exam may be eligible to transfer subject area college credit. Therefore, the significant contrast between students predicting a 4 or 5 and 3 or lower can suggest qualitative differences in their statistical literacy and motivation to do well on a practice exam.

Table 5. *Predictors of AP practice exam score (N = 161)*

		Block 1: *AP Practice*		**Block 2:** *AP Practice*	
	df	*F*	*p*	*F*	*p*
Block 1					
1. School	5	11.90 ***	<.001	13.83 ***	<.001
2. Age (years)	1	.76	.384	.89	.348
3. Gender (0=Female, 1=Male)	1	5.24 *	.023	6.09 *	.015
4. Parent Education	3	.09	.965	.11	.957
5. Math Classes Taken	4	2.28 †	.063	2.65 *	.036
6. Course Engagement (sum)	1	4.41 *	.038	5.12 *	.025
Block 2					
7. Predicted AP Score (2 mo. prior)	4	-	-	6.86 ***	<.001

Block 1 R^2 = .33, Block 2 R^2 = .43, ΔR^2 = .10
*** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$

High-stakes test performance

We next examined the set of predictors in relation to performance on an exam in a high-stakes testing context, see Table 6. Given the outcome being ordinal, an ordinal logistic regression model was tested with Hessian approximation using the *MASS* package in *R.* The same model building procedure was followed with entry of variables used to examine scores from the low-stakes testing context. In the first model, we included variables related to individual background characteristics and the classroom learning context. There was a significant effect of school (LR χ^2 = 53.04, $p < .001$). Gender was also significant (β = 1.00, SE = .30, LR χ^2 = 11.21, $p < .001$), indicating that males (M = 4.05, SD = 1.10) tended to perform slightly better on the actual test than their female (M = 3.58, SD = 1.21) classmates. The number of math classes previously taken (LR χ^2 = 10.48, $p < .001$) appeared to account for variation in test performance, though course engagement did not (LR χ^2 = 1.86, p = .173). Contrasts revealed significantly lower actual AP exam scores for students who had taken 4 (*emm* = .37) than 1 (*emm* = 2.71) previous math classes (z = 2.79, p = .042).

The second model included student's predicted scores two months prior to the exam, in addition to the variables entered into the previous model, see Table 6. Gender was no longer a significant predictor (*LR* χ^2 = 2.23, p = .135). School remained significant (*LR* χ^2 = 42.37, $p <$ .001), as did the number of math classes previously taken (*LR* χ^2 = 9.77, p = .045). In the second model, contrasts still revealed a significant difference in the actual AP exam scores between

students who had taken 4 previous high school math class (*emm* = –.38) compared with those who had taken 1 previous high school math classes (*emm* = 2.07, $z = 2.80$, $p = .045$).

Student's prediction of their exam score was significant (LR $\chi^2 = 76.75$, $p < .001$). Contrasts revealed a significant difference in actual AP exam scores between students who had predicted receiving a 5 (*emm* = 4.34) as compared to those who predicted receiving 4 (*emm* = 2.78, $z = 3.21$, $p = .012$), 3 (*emm* = .39, $z = 6.87$, $p < .001$), 2 (*emm* = –1.30, $z = 4.71$, $p < .001$), or 1 (*emm* = –2.10, $z = 4.18$, $p < .001$) on the actual exam. There was also a significant difference between students who predicted receiving a 4 compared with those who predicted receiving a 3 ($z = 5.78$, $p < .001$), 2 ($z = 3.64$, $p = .003$), or 1 ($z = 3.31$, $p = .008$). These findings indicate that students who predict receiving a score of a 4 or 5 also tend to perform better in both contexts, even after accounting for the other factors, likely for the same reasons outlined above.

Table 6. *Predictors of actual AP exam score (*N = *184)*

		Block 1: *AP Exam*		**Block 2:** *AP Exam*	
	df	*LR* χ^2	*p*	*LR* χ^2	*p*
Block 1					
1. School	5	53.04 ***	<.001	42.37 ***	<.001
2. Age (years)	1	<.001	.976	<.001	.956
3. Gender (0=Female, 1=Male)	1	11.21 ***	<.001	2.23	.135
4. Parent Education	3	2.18	.537	.91	.824
5. Math Classes Taken	4	10.48 *	.033	9.77 *	.045
6. Course Engagement (sum)	1	1.86	.173	1.49	.222
Block 2					
7. Predicted AP Score (2 mo. prior)	4	-	-	76.75 ***	<.001

*** $p < .001$, ** $p < .01$, * $p < .05$

Discussion

In the present study, we sought to examine whether students are accurate in their self-assessment, and whether their predictions of performance on a standardized AP Statistics exam was related to their actual scores. We examined self-assessment accuracy two months prior to taking the exam and within two weeks of the actual exam. Our results did not produce evidence that predictions differed between the two time points, suggesting that the predictions were relatively stable within the two-month window prior to the exam. This finding appears consistent with past research with

low-stakes assessments that found stability in students' predictions of performance (Foster et al., 2017). One explanation for such stability in students' predictions is that much of the course content had already been covered two months prior to the exam to provide an extended exam preparation period. Thus, at the two time points, students were reflecting on their knowledge of about the same amount of course material. Alternatively, such stability could be influenced by other factors, such as students' unwavering aspiration to achieve a particular level of mastery and thus receive a particular exam score (Händel & Bukowski, 2019).

The association between students' predictions of performance and actual performance was also examined in a low-stakes and high-stakes testing context. We sought to determine whether students' predictions of performance were associated with performance in both testing contexts after accounting for variation in certain background characteristics of the individual student, as well as the classroom learning context. We found that in both low- and high-stakes testing contexts, students' prediction of their performance on the AP exam was significantly associated with performance, even after accounting for other factors, such as their backgrounds and learning contexts.

We also sought to examine how these individual factors were associated with achievement in statistics. We did not find associations between either age of the participant or their parents' educational attainment with respect to their performance in either testing context. Unlike age, gender appeared to correlate with students test scores, at least when not accounting for variation in scores related to students' predictions of performance. The findings with respect to the present sample indicate that males tended to perform slightly better than female classmates with respect to the practice exam and the actual AP exam. Such findings, while troubling, may point towards the beginning of a gender achievement gap in STEM courses and subsequent career paths (Wang & Degol, 2016; You, 2010).

Prior math attainment has been shown to explain variation in the accuracy of students' prediction of performance in math education (Hatami, 2015; Hosein & Harle, 2018). The number of previous math courses accounted for variation in the high-stakes testing context both with and without accounting for students' predictions of performance. In both contexts, the association appeared to be non-linear, with contrasts indicating that a lower number of math classes previously taken was associated with high scores on both the practice and actual exam. Prior math experience, as indicated by the number of previous math classes taken, may also be confounded with students' age (Boud et al., 2013) as well as the requirements or course sequence students must meet for graduation at their schools. Furthermore, the effects found in the present analysis may be partly driven by other factors that may influence results in regression models, such as the general lack of variability in the number of math classes taken or a suppression effect.

The effect of course engagement appeared to be only robustly associated with performance in the low-stakes testing context. This effect persisted both with and without accounting for students' predictions of performance. Self-efficacy beliefs, which likely influence predictions of performance (Pajares, 1996), are believed to influence motivation and engagement (Greene, 2015). Such values and beliefs may be more likely to predict variation in performance in low-stakes assessment contexts, where intrinsic rewards (e.g., mastery of skills), as opposed to extrinsic rewards (e.g., grades), are considered motivators.

These findings build on earlier work examining the accuracy of predictions of performance. Hosein and Harle (2018) conducted a study examining accuracy both with respect to the difference between the actual and predicted scores, which they refer to as bias, as well as

the absolute value of the difference, which they refer to as inaccuracy. In the present study, we operationalized inaccuracy in much the same way; however, we opted to use an overconfidence bias estimate derived from the direction of the difference (positive, negative, neither). We felt that this approach was a more appropriate operationalization of accuracy and overconfidence bias, particularly given that the underlying associations between outcomes and predictors in each model is presumed linear. By contrast, if the raw difference between actual and predicted scores is presumed to be linearly related to the predictors, then the absolute value of that difference can hardly bear a linear association with the predictors and thus fitting linear models for both seems inappropriate. In practice, linear methods have been used to model both types of outcomes (e.g., Nietfeld & Schraw, 2002). Thus, we believe this study provides some precedent for new ways of analyzing self-assessment inaccuracy.

Our findings provide some evidence in the accuracy of students' self-assessment within the context of an AP Statistics course. These findings raise the possibility that students' judgements of learning may be used for instructional purposes. Though self-assessment is not appropriate for summative assessment purposes (Andrade, 2019), it may serve at least two practical functions. First, practicing self-assessment may facilitate the development of meta-cognitive skills, which enable students to actively monitor and evaluate their knowledge and performance, and thus achievement in math (Baliram & Ellis, 2019). Students who are encouraged to practice self-monitoring and self-regulation on a recurring basis tend to become more accurate in judging their performance and adjusting behavior to achieve certain objectives (Cleary & Zimmerman, 2004). Second, as students develop such skills, they may become more attuned to their understanding of subject matter and be better able to identify strategies that lead to improved learning and mastery of the subject matter. Findings from research has indeed shown improvements in students' monitoring accuracy and performance by means of judgment training (Händel et al., 2020; Nederhand et al., 2019). Though self-assessment may not have any direct use within a classroom learning context, it may support students in developing metacognitive skills associated with academic achievement.

Limitations

Problems of measurement abound in studying the accuracy of self-assessment. Mabe and West (1982) noted that much of the variability in effect size estimates drawn from studies examining the relation between predicted and actual performance could be explained by differences in how performance was measured. In one meta-analysis examining the relations between student and teacher assessments, Falchikov and Boud (1989) found that multiple factors appeared to influence the strength of the relation between both forms of assessment, including the level of difficulty in the course and subject matter of the course. Not only how the accuracy of self-assessment is being measured, but also the context in which it is measured appears to influence the strength of the association between self-assessment and other measures of performance.

These matters aside, there are several factors that may have restricted the amount of variability in our outcome measures, and thus the reach of our findings. There were very few students who made predictions of their performance that were greater than a two-point difference (only two participants with difference of 3 or 4 in their inaccuracies). Hence, there may not have been sufficient variability in the outcome for our predictors to reach a level of significance. Nevertheless, we did find some robust effects of predictors of performance, though not with respect to predicting differences in inaccuracy or overconfidence bias estimates. Additionally, there is likely to be missing data in the practice and actual AP exams which could bias our

results. Students may not have completed the practice exam because it had little bearing on their grade, and were unmotivated to do so. Conversely, students may not have completed the actual AP exam because they may have estimated that they would receive a low score on it and did not feel that it would be productive to sit for the exam. Further analysis should account for effects of restricted range on the association between predicted and actual scores on the AP exam due to the potential for biased missingness.

There appeared to be a robust effect of schools, which had been entered as a covariate in the model. While it is tempting to examine this effect, conclusions drawn based on the current data are likely to have very limited generalizability. Students who enroll in a particular school may have other demographic commonalities. Schools may also vary in terms of their pedagogical emphasis in preparing students for AP exams. Further work should seek to distinguish between individual and school characteristics that may influence accuracy of predictions on exams, as well as performance.

Conclusion

While past research has raised questions about the use of self-assessment in teaching and learning processes, our findings suggest that students' predictions of performance may not be entirely irrelevant to future performance in the subject area. At both time points, student's predictions were associated with actual test performance in both low- and high-stakes testing contexts. These findings suggest self-assessment may be accurate in predicting test performance in specific contexts, and stable over discrete spans of time. In high-stakes testing contexts, math experience also explained variation in test performance, even after accounting for students' predictions. Engagement was found to be associated with performance in low-stakes testing contexts. Further research should examine factors that influence prediction inaccuracy and overconfidence bias, while taking into consideration that its association with performance may be non-linear.

References

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*(87). doi:10.3389/feduc.2019.00087

Baliram, N., & Ellis, A. K. (2019). The impact of metacognitive practice and teacher feedback on academic achievement in mathematics. *School Science and Mathematics*, *119*(2), 94–104. doi:10.1111/ssm.12317

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*, *29*(1), 46–64. doi:10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. doi:10.1080/15305058.2010.508569

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). *The lme4 package*. Vienna, Austria: R Foundation for Statistical Computing

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics*, *108*(8), 355–361. doi:10.1111/j.1949-8594.2008.tb17850.x

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time?. *Assessment & Evaluation in Higher Education*, *38*(8), 941–956. doi:10.1080/02602938.2013.769198

Bourke, R. (2014). Self-assessment in professional programmes within tertiary institutions. *Teaching in Higher Education*, *19*, 908–918. doi:10.1080/13562517.2014.934353

Brown, L. A., Forman, E. M., Herbert, J. D., Hoffman, K. L., Yuen, E. K., & Goetter, E. M. (2011). A randomized controlled trial of acceptance-based behavior therapy and cognitive therapy for test anxiety: A pilot study. *Behavior Modification*, *35*(1), 31-53. doi:10.1177/0145445510390930

Carr, M., Kurtz, B. E., Schneider, W., Turner, L. A., & Borkowski, J. G. (1989). Strategy acquisition and transfer: Environmental influences on metacognitive development. *Developmental Psychology*, *25*, 765–771. doi:10.1037/0012-1649.25.5.765

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295. doi:10.1006/ceps.2001.1094

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35. doi:10.1037/bul0000052

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools*, *41*(5), 537–550. doi:10.1002/pits.10177

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

CollegeBoard. (2021). Find international universities that recognize AP. https://international.collegeboard.org/students/ap/find-universities-recognize-ap

Davis-Kean, P. E. (2005) The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment, *Journal of Family Psychology*, *19*(2), 294–304. doi:10.1037/0893-3200.19.2.294

Dew, K. H., Galassi, J. P., & Galassi, M. D. (1984). Math anxiety: Relation with situational test anxiety, performance, physiological arousal, and math avoidance behavior. *Journal of Counseling Psychology*, *31*(4), 580-583. doi:10.1037/0022-0167.31.4.580

Eccles, J. S. (2005). Influences of parents' education on their children's educational attainments: The role of parent and child perceptions. *London Review of Education*, *3*(3), 191–204. doi:10.1080/14748460500372309

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Psychology Press.

Faas, C., Benson, M. J., & Kaestle, C. E. (2013). Parent resources during adolescence: Effects on education and careers in young adulthood. *Journal of Youth Studies*, *16*(2), 151–171. doi:10.1080/13676261.2012.704989

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. doi:10.1007/s11409-016-9158-6

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, *50*(1), 14–30. doi:10.1080/00461520.2014.989230

Händel, M., & Bukowski, A. K. (2019). The gap between desired and expected performance as predictor for judgment confidence. *Journal of Applied Research in Memory and Cognition*, *8*(3), 347–354. doi:10.1016/j.jarmac.2019.05.005

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, 101245. doi:10.1016/j.learninstruc.2019.101245

Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research*, *98*, 184–193. doi:0.3200/JOER.98.3.184-192

Harding, J. L., & Hbaci, I. (2015). Evaluating pre-service teachers' math teaching experience from different perspectives. *Universal Journal of Educational Research*, *3*(6), 382–389.

Hatami, A. (2015). The effect of collaborative learning and self-assessment on self-regulation. *Educational Research and Reviews*, *10*(15), 2164–2167. doi:10.5897/ERR2015.2349

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*(1), 33–46. doi:10.2307/749455

Hendy, H. M., Schorschinsky, N., & Wade, B. (2014). Measurement of math beliefs and their associations with math behaviors in college students. *Psychological Assessment*, *26*(4), 1225–1234. doi:10.1037/a0037688

Hoffman, B., & Spatariu, A. (2008). The influence of self-efficacy and metacognitive prompting on math problem-solving efficiency. *Contemporary Educational Psychology*, *33*(4), 875–893. doi:10.1016/j.cedpsych.2007.07.002

Hosein, A., & Harle, J. (2018). The relationship between students' prior mathematical attainment, knowledge and confidence on their self-assessment accuracy. *Studies in Educational Evaluation*, *56*, 32–41. doi:10.1016/j.stueduc.2017.10.008

Jones, G.A., Langrall, C.W. & Mooney, E.S. (2007). Research in probability: Responding to classroom realities. In *The Second Handbook of Research on Mathematics*, F.K. Lester (Ed.), pp. 909–956. Reston, VA: National Council of Teachers of Mathematics (NCTM).

Judson, E. & Hobson, A. (2015). Growth and achievement trends of Advanced Placement (AP) exams in American high schools. *American Secondary Education*, *43*(2), 59–76.

Lazarides, R., & Watt, H. M. (2015). Girls' and boys' perceived mathematics teacher beliefs, classroom learning environments and mathematical career intentions. *Contemporary Educational Psychology*, *41*, 51–61. doi:10.1016/j.cedpsych.2014.11.005

Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. *CRAN R Project*. https://cran.r-project.org/web/packages/emmeans/index.html

Lew, M. D. N., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, *35*, 135–156. doi:10.1080/02602930802687737

Lopez-Pastor, V. M., Fernandez-Balboa, J.-M., Santos Pastor, M. L., & Aranda, A. F. (2012). Students' self-grading, professor's grading and negotiated final grading at three university programmes: Analysis of reliability and grade difference ranges and

tendencies. *Assessment & Evaluation in Higher Education*, *37*, 453–464. doi:10.1080/02602938.2010.545868

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280-296. doi:10.1037/0021-9010.67.3.280

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173–181. doi: 10.1023/a:1026595011371

MacPhee, D., Farro, S., & Canetto, S. S. (2013). Academic self-efficacy and performance of underrepresented STEM majors: Gender, ethnic, and social class patterns. *Analyses of Social Issues and Public Policy*, *13*(1), 347–369. doi:10.1111/asap.12033

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London, UK: Sage.

Namkung, J. M., Peng, P., & Lin, X. (2019). The relation between mathematics anxiety and mathematics performance among school-aged students: A meta-analysis. *Review of Educational Research*, *89*(3), 459-496. doi:10.3102/0034654319843494

National Center for Education Statistics [NCES]. (2021). Definition of "Doctor's degree-professional practice". *Integrated Postsecondary Education Data System (IPEDS) Glossary*. Retrieved 14 Sept. 2021 from https://surveys.nces.ed.gov/ipeds/public/glossary

Nederhand, M. L., Tabbers, H. K., Jongerling, J., & Rikers, R. M. (2020). Reflection on exam grades to improve calibration of secondary school students: a longitudinal study. *Metacognition and Learning*, 1–27. doi:10.1007/s11409-020-09233-9

Nederhand, M. L., Tabbers, H. K., & Rikers, R. M. (2019). Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. *Applied Cognitive Psychology*, *33*(6), 1068–1079. doi:10.1002/acp.3548

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research*, *95*(3), 131–142. doi:10.1080/00220670209596583

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578. doi:10.3102/00346543066004543

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803–830. doi:10.1007/s10648-015-9350-2

Paulus, M., Tsalas, N., Proust, J., & Sodian, B. (2014). Metacognitive monitoring of oneself and others: Developmental changes during childhood and adolescence. *Journal of Experimental Child Psychology*, *122*, 153–165. doi:10.1016/j.jecp.2013.12.011

Putwain, D. (2008). Do examinations stakes moderate the test anxiety–examination performance relationship?. *Educational Psychology*, *28*(2), 109–118. doi:10.1080/01443410701452264

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org/.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*(3), 114–121. doi:10.1006/drev.2002.0544

Schneider, W., Lingel, K., Artelt, C., & Neuenhaus, N. (2017). Metacognitive knowledge in secondary school students: assessment, structure, and developmental change. In *Competence Assessment in Education* (pp. 285–302). New York, NY: Springer, Cham.

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13*(4), 505–514. doi:10.1016/0092-6566(79)90012-6

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high-and low-stakes test achievement. Learning and individual differences, 39, 4–63. doi:10.1016/j.lindif.2015.03.002

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high-and low-stakes test achievement. *Learning and Individual Differences*, 39, 4–63. doi:10.1016/j.lindif.2015.03.002

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. doi:10.3102/00346543075003417

Sommer, M., & Arendasy, M. E. (2016). Does trait test anxiety compromise the measurement fairness of high-stakes scholastic achievement tests?. *Learning and Individual Differences*, *50*, 1–10. doi:10.1016/j.lindif.2016.06.030

Smith, F. X., & Was, C. A. (2019). Knowledge monitoring calibration: Individual differences in sensitivity and specificity as predictors of academic achievement. *Educational Sciences: Theory & Practice*, 19(4), 80–87. doi:10.12738/estp.2019.4.006

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245. doi:10.1037/0033-2909.87.2.245

Stevens, T., Olivarez, A., Lan, W. Y., & Tallent-Runnels, M. K. (2004). Role of mathematics self-efficacy and motivation in mathematics performance across ethnicity. *The Journal of Educational Research*, *97*(4), 208–222. doi:10.3200/JOER.97.4.208-222

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593. doi:10.1177/0956797617741719

Tanner, H., & Jones, S. (1994). Using peer and self-assessment to develop modelling skills with students aged 11 to 16: A socio-constructive view. *Educational Studies in Mathematics*, *27*(4), 413–431. doi:10.1007/BF01273381

Tsalas, N., Paulus, M., & Sodian, B. (2015). Developmental changes and the effect of self-generated feedback in metacognitive controlled spacing strategies in 7-year-olds, 10-year-olds, and adults. *Journal of Experimental Child Psychology*, *132*, 140–154. doi:10.1016/j.jecp.2015.01.008

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, *33*(4), 304–340. doi:10.1016/j.dr.2013.08.001

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360. doi:10.3102/0013189X14553660

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

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*(1), 33–47. doi:10.1007/s10649-017-9764-5

Whitney, B. M., Cheng, Y., Brodersen, A. S., & Hong, M. R. (2019). The scale of student engagement in statistics: Development and initial validation. *Journal of Psychoeducational Assessment*, *37*(5), 553–565. doi:10.1177/0734282918769983

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 396-399. doi:10.1111/j.2517-6161.1959.tb00346.x

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, *8*(4), 327–353. doi:10.1016/S1041-6080(96)90022-9

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

You, S. (2013). Gender and ethnic differences in precollege mathematics coursework related to science, technology, engineering, and mathematics (STEM) pathways. *School Effectiveness and School Improvement*, *24*(1), 64–86. doi:10.1080/09243453.2012.681384