Research article

STEM students prefer assessment practices known to reduce the impact of test anxiety

R. A. Costello[1,2], S. P. Hammarlund[3], E. M. Christiansen[4], M. K. Kiani[1], M. S. Glessmer[4,5], S. Cotner[3,4], and C. J. Ballen[1]

[1] Auburn University, USA
[2] University at Buffalo, USA
[3] University of Minnesota, USA
[4] University of Bergen, Norway
[5] Lund University, Sweden

*Corresponding author. E-mail: robincos@buffalo.edu

Abstract Undergraduate introductory STEM courses often rely on few, high-stakes exams to assess student learning. This assessment strategy engenders high test anxiety and negatively impacts academic performance. We know little about how students want to be assessed—particularly those who experience high test anxiety. Applying a humanist learning framework, we asked students at a university in Norway to envision their ideal assessment practice. Our analyses affirm that test anxious students performed worse in their STEM courses, and students with marginalized identities in STEM were more test anxious. Additionally, we found that students overwhelmingly want more assessments. We also found that first-generation students, a student group rarely studied in Norway, want different types of assessments to replace high-stakes exams. In sum, student preferences aligned with assessment practices known to reduce the impact of test anxiety. Our results support calls for creating STEM environments where student voices are valued.

Keywords:
assessments, first-generation students, humanist learning framework, introductory STEM courses, STEM equity, test anxiety

1 Introduction

In higher education, attrition rates among science, technology, engineering, and mathematics (STEM) majors are highest for members of groups with a history of exclusion from STEM, including members of marginalized racial/ethnic groups, first-generation college students, and women (Hill et al., 2009; Rosenberg et al., 2018; Trapani & Hale, 2022). Similar disparities exist in academic performance in introductory STEM college courses (Salehi et al., 2019). In seeking to understand predictors of performance and retention in STEM higher education, investigators have looked at incoming academic preparation (Warburton et al., 2001; Westrick et al., 2015; Salehi et al., 2020), environmental or contextual factors (Lundeberg et al., 2011; Salehi et al., 2021; Thompson et al., 2020), and student affect (e.g., sense of belonging (O'Brien et al., 2020; Hammarlund et al., 2022); science self-efficacy (Ballen et al., 2017); interest in STEM (Burnette et al., 2020)). Here, we focus on assessment practices and how they can contribute to differences in academic performance among students from marginalized groups.

Introductory STEM college courses often rely on few, high-stakes exams to assess student learning. High-stakes exams account for the majority of the course grade and are typically administered in multiple choice format in introductory STEM courses. Several studies, largely conducted at North American institutions, have documented the negative impacts of high-stakes exams on student academic performance in STEM courses, especially among students with marginalized identities (Macphee et al., 2013; Koester et al., 2016; Ballen, Salehi et al., 2017). Women, first-generation college students, and Black, Native American, and Hispanic students all perform worse on high-stakes exams relative to men, continuing-generation students, and white students (Macphee et al., 2013; Koester et al., 2016; Ballen, Salehi et al., 2017).

Here, we extend analyses of the relationships between high-stakes exams and student academic performance to a Norwegian context. Norwegian higher education presents an interesting case study because most STEM higher education coursework is characterized by grading based on one or two high-stakes tests. Where low-stakes assignments are used, they may not be graded and thus may not serve to reduce the impact of the exams. Furthermore, in Norway, gender equality is among the highest in the world (World Economic Forum, 2022). This status is somewhat counterbalanced by a pervasive gender gap in obtaining STEM degrees in Norway and other relatively gender-equal countries, a phenomenon referred to as "The Gender Equality Paradox" (Stoet & Geary, 2018). In response to the lack of women in STEM fields, some investigators have begun seeking causal explanations for gender-biased attrition (e.g., Ballen & Holmegaard, 2019; Eddy & Brownell, 2016). A possible explanation includes test anxiety-mediated performance differences.

1.1 Test anxiety impacts student performance

High-stakes exams are associated with elevated test anxiety—the anxiety caused by internal or external pressure to perform in evaluative situations (Hodapp & Benson, 1997; Cassady & Johnson, 2002; Zeidner, 2010). Test anxiety can be viewed as one phenomenon, albeit composed of multiple dimensions such as worry, emotionality, and lack of confidence (Hodapp & Benson, 1997). Here we focus on students' self-reported

trait test anxiety (hereafter simply 'test anxiety'), which are the perceptions of their "generalized and enduring predisposition to react" to assessments in a consistent way (Endler & Kocovski, 2001). This dimension of test anxiety differs from *state-based test anxiety*, which refers to students' feelings immediately in response to high-stakes assessment. Significantly, test anxiety may not be felt equally by all students, and its impacts may vary by student characteristics. For example, several studies indicate that women in STEM courses exhibit more test anxiety than do men (Chapell et al., 2005; Lowe, 2015). Further, test anxiety in women—but not in men—is negatively and significantly associated with performance on exams, possibly explaining some of the gender-biased performance gaps that have been documented in STEM fields (Ballen, Salehi et al., 2017; Salehi et al., 2019). A recent meta-analysis analysed performance in 169 biology and chemistry courses and further illustrated that women tend to underperform on high-stakes tests but not on lower-stakes assessments (Odom et al., 2021), a finding consistent with these differences in test anxiety. Similar patterns have been documented with first-generation college students; specifically, first-generation students have higher test anxiety (Guadier-Diaz et al., 2019) and lower performance (Harackiewicz et al., 2014; Canning et al., 2019; Guadier-Diaz et al., 2019), on average, than their peers. Furthermore, high test anxiety has been shown to mediate low academic performance among first-generation students (Guadier-Diaz et al., 2019; Salehi et al., 2020). In sum, test anxiety and associated outcomes likely explain some of the performance and retention discrepancies that plague STEM higher education.

1.2 Test anxiety and assessment type

There are, however, pedagogical practices shown to reduce these differences in test anxiety and performance and possibly impact retention in STEM fields. Critically, educators can shift assessment strategies to favour more low-stakes assessments (e.g., weekly quizzes or assignments) over one or two high-stakes exams that constitute the bulk of a student's grade. In a series of three case studies, Cotner and Ballen (2017) showed that shifting to low-stakes assessments removed gender-biased performance gaps that existed in high-stakes testing situations. Furthermore, educators can change the type of assessments given in their course to reduce test anxiety and promote equitable student performance. Open-note exams decrease test anxiety and increase performance, especially among women (Gharib et al., 2012; Zoller & Ben-Chaim, 1990). Furthermore, the type of questions asked on exams (i.e., multiple-choice versus constructed-response questions) can impact student performance outcomes. In an experiment across two sections of an introductory biology course, Stanger-Hall (2017) found that students used more active study behaviours when preparing for constructed response exams and also experienced more equitable student outcomes. By administering open-note exams with constructed response questions, instructors can reduce their students' test anxiety and promote effective study habits (Driessen et al., 2022).

1.3 Assessment and student ownership

Other studies have suggested that students themselves may be aware of which assessment methods best support their learning, and, if given options, will choose assessments accordingly (e.g., Clack and Dommett, 2021, Chase, 2020). When students

are given ownership over how they are assessed, they feel more empowered and experience deeper learning (Chase, 2020; Clack & Dommett, 2021). Furthermore, Pretorius et al. (2017) found that, given the choice between two versus four graded assessments, the majority (66.4%) of students in an undergraduate accounting course chose more graded assessments; further, students reported that assessment choice increased their motivation and reduced stress. This study documents that student assessment preferences can align with assessment practices known to reduce test anxiety (in this case, more low-stakes assessments) and suggests that incorporating student choice into assessment practices may improve student outcomes.

Despite the positive outcomes resulting from the incorporation of student assessment preferences, questions remain about whether preferences vary across disciplines, learning contexts, and students who experience different levels of test anxiety and hold different identities. We fill this gap in knowledge by characterizing STEM student assessment preferences across several disciplines in a single institution.

1.4 A humanist learning theoretical framework

A humanist approach in education centres student voices in their learning and evaluation (Rogers & Freiberg, 1994). Student-centred learning most often considers learning as a process where students create their own understanding through active engagement, closely aligning with a constructivist approach to learning (Piaget, 1972; Vygotsky, 1978; Rogers & Freiberg, 1994; Tangney, 2014). However, student-centred learning also relies on student self-belief, self-confidence, and self-empowerment to actively create understanding, in alignment with a humanist approach to learning (Rogers & Freiberg, 1994; Tangney, 2014). A humanist education is also rooted in critical consciousness or critical pedagogy, which calls for educators to empower students in their own learning and to offer students choice in what they do and how they do it, such as in the matter of assessment (Freire, 1970). To build student empowerment, humanist learning theory recommends minimizing power differentials in classrooms and providing students with choice (Rogers & Freiberg, 1994). We apply this theory by asking students how they would prefer to be assessed in their STEM courses. We hypothesize that students with high test anxiety will prefer to be assessed by methods known to reduce test anxiety. Furthermore, we expect women and first-generation students, who typically experience high test anxiety, to likewise prefer assessment practices that reduce test anxiety. Understanding the connections between student preferences and outcomes such as test anxiety and performance could give instructors an important tool for creating student-centred (and student-empowered) spaces for learning.

1.5 Test anxiety and performance in the Norwegian context

Most of the work on test anxiety is based on studies conducted in North America, potentially limiting its applicability to global conversations around STEM performance, retention, and equity. We know of one study on gender and test anxiety in STEM higher education in Norway (Cotner et al., 2020), which demonstrated that women have higher test anxiety than men, and that test anxiety negatively and significantly predicts performance in women but not in men. However, this study was limited to a single section of an introductory biology course, and its implications are thus constrained. We also do not know how test anxiety varies across college generation status in Norway (i.e.,

whether students are the first generation in their family to attend university). Although college generation is rarely considered in Norway, recent evidence demonstrates that already in lower secondary school (*ungdomsskole*), pupils with the highest educated parents have a grade average that is about one grade higher than that of students with parents without similar educational backgrounds (Larsen, 2018). These same students whose parents are highly educated are three times more likely to get a university degree (Hovdhaugen, 2009; Hansen, 2010; Holseter, 2021). Differential test anxiety, between first- and continuing-generation college students, may be implicated in some of these findings.

In this study, we used an end-of-term survey to ask STEM students in a Norwegian university about their test anxiety and about what, in their opinion, constituted ideal assessment practices for their current science course. Our specific research questions were:

1. *Does test anxiety affect course grades?*
2. *Does test anxiety vary based on gender or generation in college?*
3. *What types of assessment do students prefer in their STEM courses?*
4. *Are student preferences related to test anxiety, gender, or generation in college?*

2 Methods

2.1 Data collection

We surveyed 236 students in four introductory-level STEM courses (Informatics, two Chemistry courses, and Geology) at the University of Bergen in Norway during Spring 2022. University of Bergen is a research-intensive public University with approximately 18,000 students. The four introductory-level STEM courses all included a final exam that constituted 60 to 100 percent of the grade (Table 1). We surveyed students during the last two weeks of class but before the final exam.

Table 1. *Description of the four STEM courses at the University of Bergen included in our study. Each course had its own grading scheme. However, all courses included final exams (bolded) that constituted the majority of the course grade, ranging from 60% to 100% of the grade. The number of students indicates the number of students who filled out our survey and consented to participate in the study, not students enrolled in the course.*

Course	**Grading Scheme**	**Number of Students**
Informatics	mandatory submissions (ungraded); **written final exam (5 h) (100% of grade)**	97
Chemistry A	**written final exam (4 h) (100% of grade)**	44
Chemistry B	compulsory assignments (pass/fail); laboratory course (20% of grade); mid-semester assessment (2 h) (20% of grade); **written final exam (4 h) (60% of grade)**	53
Geology	5 seminar assignments (ungraded); group field report (33% of grade); **written final exam (66% of grade)**	42

Our survey was designed to meet several different research and evaluation needs. Here, we used survey responses to items from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991), an open-ended question about assessment preferences, and demographic questions. We describe these survey items in detail below.

To measure test anxiety, we used the four-item subscale from the short version of the MSLQ (Cotner et al., 2020; Pintrich et al., 1991). Students were asked the extent to which they agree, on a seven-point Likert scale, with the following statements:

1. I am so nervous during tests that I do not remember the facts I have learned
2. I have an uneasy, upset feeling when I take a test
3. I worry a good deal before tests
4. When I take a test, I think about how poorly I am doing

To understand student assessment preferences, our survey included an open-ended prompt: "If you could freely design a form of assessment in this course that was fair and that supported your learning, what would that assessment look like? (This can include mandatory activities, several small quizzes or tests, a single written report, a single exam, a portfolio, and so on.) Which of these activities should be a part of the final assessment? Please explain/develop your suggestion." This question was translated into Norwegian by two independent native speakers, confirmed by two additional native speakers, and then subjected to "think aloud" validation with students in the bioBERG research group at the University of Bergen.

Lastly, our survey asked students to self-identify their gender and whether neither, one, or both of their parents completed higher education, including university or university college (*høyskole*). Universities in Norway offer advanced (graduate-level) degrees, whereas university colleges are regional institutions that primarily award bachelor degrees (typically in fields such as technology, business, education, and nursing).

Our study was approved by The Norwegian Centre for Research Data (reference number 963610). Students were provided with information about the study. Specifically, students were informed that the data would be treated confidentially and anonymized in all publications. Student participants had the opportunity to withdraw from the study at any time, and only those who consented to participate in the study were included in the analyses. No rewards were given for participation.

In addition to our survey, we also collected end-of-term course grades from students who agreed to allow the researchers to access course grades. Passing grades ranged from A to E; no pluses or minuses were given.

2.2 Measuring test anxiety

To quantify test anxiety, we first verified that the survey items represented the established construct of test anxiety in our sample using confirmatory factor analysis (CFA) with the R package lavaan (Rosseel, 2012; see Knekta et al., 2019; Ballen & Salehi, 2021 for more details). We specified a one-factor CFA with four MSLQ survey items using a sample of 201 students who fully completed the test anxiety MSLQ subscale. 201 is a sufficient sample size for a one-factor CFA with four items (Wolf et al., 2013; Knekta et al., 2019). The test anxiety MSLQ subscale has been previously validated across many different student populations (Pintrich et al., 1991; McClendon, 1996; Büyüköztürk et al., 2004; Feiz & Hooman, 2013; Jakešová & Hrbáčková, 2014). Good model fit values are a comparative fit index (CFI) > 0.95 in combination with a root-mean-square error of approximation (RMSEA) < 0.06 or a standardized root-mean-square residual (SRMR) <

0.05 (Hu & Bentler, 1999; Ballen & Salehi, 2021). Our one-factor CFA model fit was good according to most of these standards (χ^2 = 7.69, df = 2, P = 0.021; CFI = 0.986; RMSEA = 0.119; SRMR = 0.026). The RMSEA of our model indicates that our model is a poor fit. However, RMSEA has been shown to often falsely indicate a poor fitting model for models with small degrees of freedom and small sample sizes, as is the case in our model (Kenny et al., 2015). Furthermore, the survey items were internally consistent in our sample (Cronbach's alpha = 0.87; Nunnally & Bernstein, 1994). Given the consistency and validity of the survey items in our sample, we used the values for the latent variable in our CFA model as our measure of test anxiety.

For ease of interpretation, we mean-variance standardized these test anxiety values across all students in our sample. Specifically, we calculated Z-scores using the formula Z-score = (X-μ)/s, where X is the test anxiety value measured from the CFA model, μ is the mean test anxiety value, and s is the standard deviation around the mean. Z-scores measure how many standard deviations a student's test anxiety is from the average test anxiety across all four courses.

2.3 Linear mixed models

We conducted a series of linear mixed models (specified below) to analyse the relationships between test anxiety, gender, college generation, course grades, and assessment preferences.

In our analyses of student gender, we compared men and women, excluding two students who did not identify as a man or a woman. We recognize that gender is a gradient, do not wish to minimize the voices of the two students who do not identify with a gender binary, and included those students in our analyses of qualitative data. In our analyses of student first-generation status, we compared students with one or more parents with higher education (continuing-generation students) to students with parents without higher education (first-generation students).

All of our mixed models were built with the R package lme4 (Bates et al., 2015), and all models included course as a random effect. The R package emmeans was used to calculate marginal means (Lenth, 2018). Figures were built in the R package ggplot2 (Wickham, 2016). All analyses were conducted in R version 4.1.1 (R Core Team, 2021).

2.4 Does test anxiety affect course grades?

To quantify how test anxiety affects course grade, we ran a linear mixed model with test anxiety as a fixed effect and course as a random effect. To measure how gender and generation in college affected the relationship between test anxiety and course grade, we added interactions between gender and test anxiety and between college generation and test anxiety in two additional linear mixed models. For these analyses, letter grades were converted to numeric grades with A as the highest (6) and E as the lowest (1). Of the 236 students who consented to participate in the study, 94 students consented to provide their course grades and responded to the test anxiety survey items.

2.5 Does test anxiety vary based on gender or generation in college?

To test how gender and generation in college impact test anxiety, we ran a linear mixed model with gender, college generation, and the interaction between gender and college

generation as fixed effects and course as a random effect. As only a subset of the 236 students provided complete demographic information and responded to the test anxiety items, this analysis included 175 students.

2.6 What types of assessments do students prefer in their STEM courses?

106 of the 236 students who consented to participate in the study also answered the open-ended question about assessment preference. After an initial review of a subset of data, we found that students described their assessment preferences in relation to the current assessment of their course. We used deductive coding to match the student responses to four categories that we assigned that captured these comparative student responses. These categories were *more assessments, fewer assessments, different types of assessments* (hereafter *'different assessments'*), and *no changes to assessments*. In our definition, *'assessments'* included both graded and ungraded assignments. Student responses could be coded into multiple categories.

Two coders (RC and SC) worked independently to assign student comments to categories and then met to discuss these decisions and came to consensus. As the categories are simple and broad, the need for discussion was minimal, and agreement was nearly perfect prior to discussion.

Our findings indicated that 21 students preferred different types of assessments. We analysed these student responses to understand the types of different assessments they preferred. We used deductive coding to categorize student responses into 4 categories that we assigned after an initial review of a subset of the data. These categories were *portfolios, assignments/activities, quizzes,* and *written reports*. Student responses could be coded into multiple categories.

Two coders (RC and CJB) worked together to assign student comments to categories. As the categories are simple and broad, the need for discussion was minimal.

2.7 Are student assessment preferences related to test anxiety, gender, or generation in college?

We ran four separate mixed effects logistic regressions to measure how test anxiety impacted the probability that a student preferred more assessments, fewer assessments, different types of assessments, or no changes to assessments. 100 students were included in these models, as 100 of the 106 students who described their assessment preferences also responded to the test anxiety survey items. To test how gender and college generation affect student assessment preference, we ran two additional mixed effects logistic regressions for each student assessment preference (more assessments, fewer assessments, different types of assessments, and no changes to assessments) with either gender or college generation as a fixed effect and course as a random effect. 104 students were included in the gender models (N = 32 men and 72 women), excluding the two non-binary students, and all 106 students who described their assessment preferences were included in the generation models (N = 82 continuing-generation and 24 first-generation students). Our sample is skewed towards women and continuing-generation students, which is reflective of the college at large and higher education in Norway.

3 Results

As mentioned above, we used Z-scores of the latent variable values in our CFA model as our measure of test anxiety in our mixed effect models. However, simple averages of the four survey items show patterns of test anxiety. In our sample of students taking STEM courses at the University of Bergen in Norway, test anxiety ranged from 1 to 7 but averaged 4.24 ± 1.55 (mean ± standard deviation). Test anxiety was over a point higher on the Likert scale in women (4.55 ± 1.50) than in men (3.43 ± 1.52). Additionally, test anxiety was higher in first-generation students (5.05 ± 1.19) than in continuing-generation students (4.11 ± 1.63).

3.1 Does test anxiety affect course grades?

More test-anxious students received lower course grades (Figure 1A; β = -0.55 ± 0.27 95% CI; P = 0.000073; N = 94 students). This relationship between test anxiety and course grade depended on student college generation status (Figure 1B; P = 0.0088; N = 18 first-generation, 74 continuing-generation) but did not depend on student gender (P = 0.74; N = 63 women, 30 men). Academic performance was more sensitive to test anxiety among first-generation students compared to continuing-generation students (Figure 1B; first-generation: β = -1.64 ± 0.92; continuing-generation: β = -0.38 ± 0.31; marginal trends ±95% confidence intervals reported).

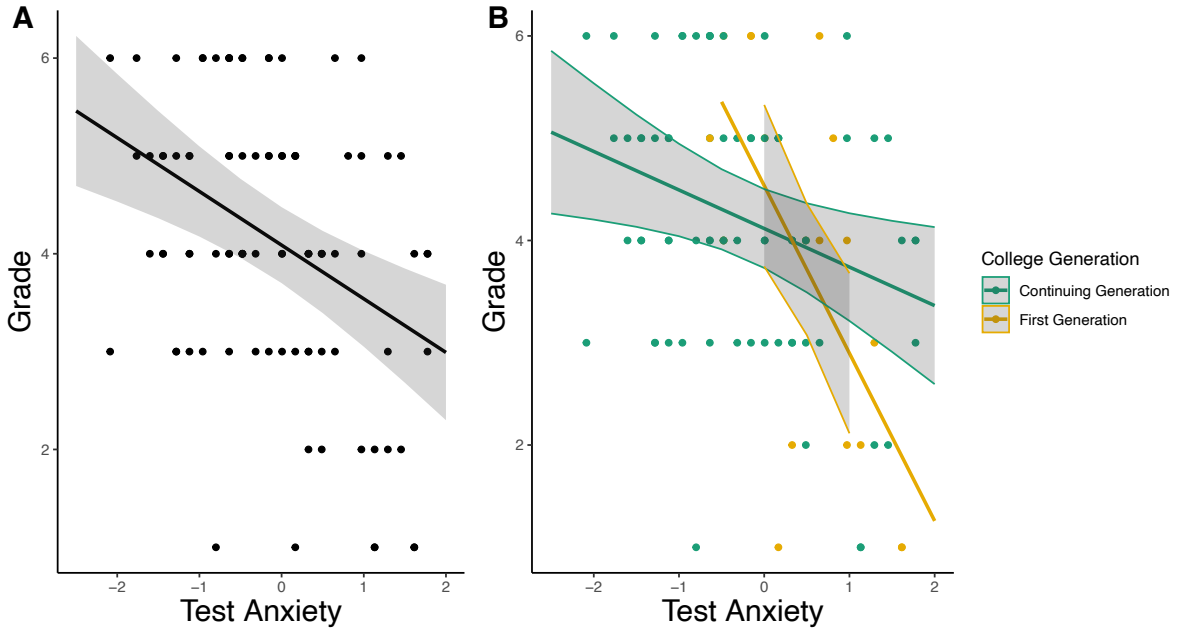

Figure 1. *Relationship between test anxiety and grade. (A) More test anxious students received lower course grades (β = -0.55 ± 0.27; N = 94 students; P = 0.000073). (B) The relationship between test anxiety and grade depends on college generation (first-generation: β = -1.64 ± 0.92, N = 18; continuing-generation: β = -0.38 ± 0.31, N = 74; P = 0.0088). Marginal trends with 95% confidence error bars were extracted from linear mixed models testing the effect of test anxiety on course grade, and points depict individual grades by averages of individual responses to the 4-item Likert questions that have been mean-variance standardized. Letter grades were converted to numbers: A is highest (6) and E is lowest (1).*

3.2 Does test anxiety vary based on gender or generation in college?

Women reported more test anxiety than men (Figure 2A; women: 0.385 ± 0.264; men: -0.389 ± 0.448; P = 0.0014; N = 175 students, 122 women, 53 men; marginal means ± 95% confidence intervals reported). This means that men reported 0.77 standard deviations lower test anxiety than women in our sample. First-generation students, which we define as students whose parents have not graduated from institutions of higher education, were marginally more test anxious than continuing-generation students (Figure 2B; first-generation: 0.216 ± 0.466; continuing-generation: -0.220 ± 0.271; P = 0.072; N = 175 students, 33 first-generation, 142 continuing-generation; marginal means ± 95% confidence intervals reported). This means that continuing-generation students reported 0.44 standard deviations lower test anxiety than first-generation students. Student gender and college generation did not interact to affect test anxiety (gender x college generation: P = 0.72).

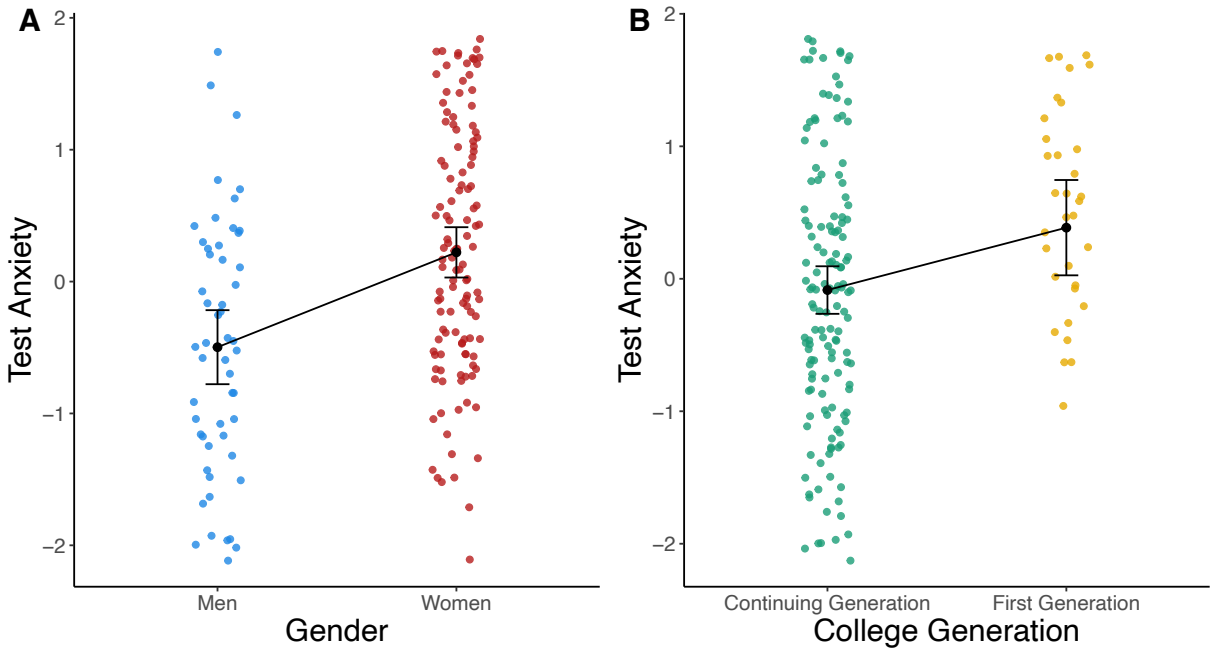

Figure 2. *Test anxiety (Z-score relative to average) across (A) gender and (B) generation in college. (A) Women reported higher test anxiety than men (women: 0.385 ± 0.264, N = 122; men: -0.389 ± 0.448, N = 53; P = 0.0014). (B) First-generation students were marginally more test anxious than continuing-generation students (first-generation: 0.216 ± 0.466, N = 33; continuing-generation: -0.220 ± 0.271, N = 142; P = 0.072). Marginal means with 95% confidence error bars were extracted from a linear mixed model testing the effects of gender, college generation, and the interaction between gender and college generation on test anxiety, and points depict averages of individual responses to the 4-item Likert questions that have been mean-variance standardized.*

3.3 What types of assessments do students prefer in their STEM courses?

The majority (63%) of students who responded to our open-ended question and described their ideal form of assessment for their STEM course (N = 106 students) wanted their course to include more assessments (Figjure 3). 9% of students wanted fewer assessments in their STEM course (Figure 3). 20% of students wanted different types of assessments in their course (Figure 3). 23% of students did not want their course assessment to change (Figure 3).

Figure 3. *Codebook from deductively coding student descriptions of their ideal form of assessment in their STEM course. The number of student responses was broken down for each code and was further categorized by student gender and college generation. The right-most panel displays the percent of student responses for each of the four categories (N = 106 students).*

3.4 Are student assessment preferences related to test anxiety, gender, or generation in college?

Test anxiety did not impact whether students wanted more assessments, fewer assessments, or no changes to assessments in their course (*more assessments*: P = 0.69; *fewer assessments*: P = 0.44; *no changes*: P = 0.40; N = 100 students). Students with higher test anxiety were marginally more likely to want different types of assessments in their STEM courses (P = 0.075; N = 100 students).

Women and men did not give significantly different answers to their ideal assessment structure (*fewer assessments*: women: 13% [7%, 22%]; men: 3% [0.4%, 19%]; P = 0.17; *more assessments*: women: 69% [33%, 91%]; men: 69% [31%, 92%]; P = 1.00; *different assessments*: women: 18% [11%, 29%]; men: 19% [9%, 36%]; P = 0.93; *no changes*: women: 18% [6%, 43%]; men: 25% [8%, 56%]; P = 0.47; back-transformed marginal means with [95% confidence intervals] reported).

First-generation and continuing-generation students likewise wanted their STEM courses to have more assessments, fewer assessments, and no changes to the assessments at similar rates (Figure 4; *more assessments*: first-generation: 76% [35%,

95%]; continuing-generation: 69% [31%, 91%]; P = 0.51; *fewer assessments*: first-generation: 10% [2%, 41%]; continuing-generation: 7% [1%, 25%]; P = 0.57; *no changes*: first-generation: 16% [4%, 48%]; continuing-generation: 21% [7%, 49%]; P = 0.62; back-transformed marginal means with [95% confidence intervals] reported). However, first-generation students more frequently asked for different types of assessments in their STEM courses than continuing-generation students (Figure 4; first-generation: 33% [18%, 54%]; continuing-generation: 16% [9%, 25%]; P = 0.065; back-transformed marginal means with [95% confidence intervals] reported).

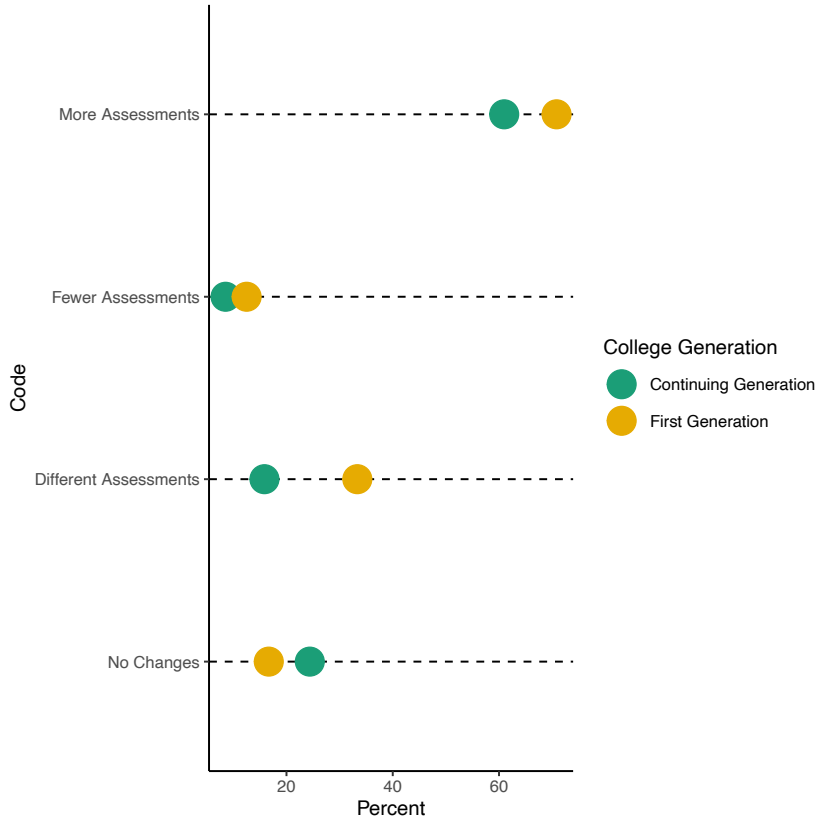

Figure 4. *Percent of student responses for each of the four categories across college generation. First-generation students (N = 24) more often asked for different types of assessments in their STEM courses than continuing-generation students (N = 82) (P = 0.065). First-generation and continuing-generation students did not differ in the rates they asked for more assessments (P = 0.51), fewer assessments (P = 0.57), and no changes to the assessments (P = 0.62) in their STEM courses.*

By aligning the open-ended survey responses about ideal assessments with student characteristics, we can shed additional light on how test anxiety can impact assessment preferences. For example, one first-generation woman with high test anxiety (6.75 out of 7) responded, “Oh please. Just please anything other than a big threatening final exam at the end. IT DOES NOT WORK.” Another first-generation woman (6.75/7) said, “…more mandatory assignments…I find this way makes it easier to work evenly throughout the semester.” In contrast, a continuing-generation man with low test anxiety (1.5/7) remarked, “Final exam. I prefer to have 100% of my grade determined in a final exam, so that there is a chance to ‘catch up’ in the course.” And another continuing-generation

woman (1.75/7) said, “The way it is now, with 14 mandatory assignments and a final exam, is good.”

Students (N = 21) suggested several different types of assessment options, including portfolios, assignments, quizzes, and written reports (Figure 5). The majority (67%) of students who prefer different types of assessments wanted to be assessed with portfolios, a common type of assessment in Norway in which students are evaluated based on multiple assignments from throughout the term (Figure 5). For example, a first-generation woman (4/7) wrote, “A report or a portfolio could be an option. I personally feel that I learn a lot more from this. Could have several exercises throughout a semester that are handed in, combined, at the end as a folder, maybe consisting of both reports and other smaller tasks.” Another first-generation woman (3.5/7) also wanted a portfolio and specified, “but then it should contain varied assessments, both report and mandatory assignments. But also tests where one has not used any aids.”

Categories	Examples	Number of Student Responses (N = 21)	Men (N = 6)	Women (N = 13)	Continuing Generation (N = 13)	First Generation (N = 8)
Portfolios	“I wish the course went over to a portfolio with small chapter tests, quizzes, and written reports rather than an exam.”	14	2	12	7	7
Assignments/ Activities	“More evaluation options, like mandatory assignments with feedback.”	10	1	8	4	6
Quizzes	“More small quizzes, so one is ‘forced’ to learn smaller parts of the curriculum regularly throughout the semester.”	9	4	4	5	4
Written Reports	“Something more practical in the form of, for instance, a relevant lab course with report writing.”	6	2	3	3	3

Figure 5. *Codes from student descriptions of the types of different assessments preferred in their STEM course. Example quotes for each code and the breakdown of the number of student responses for each code are provided. The number of student responses was further categorized by student gender and college generation.*

4 Discussion

We found that test anxiety was negatively correlated with performance, that women had higher test anxiety than men, and that first-generation college students had higher test anxiety than their continuing-generation peers. We also documented patterns in assessment preferences, whereby the majority of students desired more assessments, either graded or not. Further, first-generation students were especially interested in having different types of assessments. Following the humanist theoretical framework, which calls for education to provide students with opportunities for empowerment (Freire, 1970), and in which students are seen as active participants in their education rather than passive receivers (Tangney, 2014), our results hold valuable lessons and recommendations for instructors. Our results call for instructors to allow space for students to make choices to maximize their potential, such as shifting from using high-

stakes exams towards multiple, low-stakes assessments that vary in format. We summarize our main findings below and place them in the context of the broader literature.

4.1 Does test anxiety affect course grades?

High test anxiety is associated with lower grades. We add to extensive previous literature demonstrating test anxiety has a negative impact on grades throughout a student's education (reviews by Hembree, 1988; Richardson et al., 2012), and across disciplines, such as physics (Oludipe, 2009; Malespina & Singh, 2022), math (Caviola et al., 2021), and economics (Kader, 2016). Several perspectives suggested different mechanisms underlying test anxiety and its impacts (Zeidner, 2010). One of the most common explanations is that test anxiety causes attentional interference, which is the diversion of mental resources that would otherwise be devoted to critical tasks such as short-term memory and cognitive processing (Wine, 1971; Eysenck & Calvo, 1992). A compatible explanation posits test anxiety causes individuals to associate normal physical arousal with potential failure during an assessment (Zeidner, 2014). Some researchers push back against the notion that test anxiety has sizable impacts on student performance outcomes. For example, using data from 309 medical students who prepared for a high-stakes exam, Theobald et al. (2022) showed test anxiety did not predict exam performance over students' knowledge level, concluding with calls for alternative explanations for lower academic performance among test-anxious students. While medical students represent a unique student demographic, the results align with several studies showing it is not test anxiety that directly impacts performance, but rather test-anxious students also have ineffective study behaviours, which explains lower knowledge acquisition and performance during an exam (Cassady, 2004; Culler & Holahan, 1980; Ewell et al. 2022; Theobald et al., 2022). Another explanation offers that test anxiety (in this case, math anxiety) undermines student performance on exams directly and indirectly through study behaviours, which partially mediates the relationship between anxiety and performance (Jenifer et al., 2022). In other words, more math-anxious students engaged less while studying, which in turn related to their performance on the exam. Implications of our results, which encourage instructors to use multiple, low-stakes assessments rather than a single high-stakes exam, may help students acquire knowledge through practice and active study behaviours (Kirkland & Hollandsworth, 1980). Thus, using assessments that fundamentally align with better learning will benefit students regardless of whether test anxiety, ineffective study habits, or other unforeseen factors underlie lower grades.

4.2 Does test anxiety vary based on gender or generation in college?

Women and first-generation students experience high test anxiety. Our results align with previous work showing women report higher test anxiety than men and that first-generation college students report higher test anxiety than their continuing-generation peers (Chapell et al., 2005; Lowe, 2015; Guadier-Diaz et al., 2019; Salehi et al., 2020). The majority of this work has focused on gender disparities in test anxiety (but see Guadier-Diaz et al., 2019). For example, Hembree (1988) conducted a meta-analysis of 562 studies on test anxiety and observed that gender-based differences were small in the 'early school years', peaked in grades 5-10, and then declined through upper high

school and college. More recent work has shown how test anxiety is negatively correlated with performance in women, but not necessarily in men (Ballen, Salehi et al., 2017; Salehi et al., 2019)–including in Norway (Cotner et al., 2020). From an equity perspective, it is particularly problematic that students with identities already marginalized in science are also more susceptible to test anxiety.

While a sizable body of literature has found high test anxiety in girls and women, other results do not support the link between test anxiety and performance. Harris et al. (2019) administered several test anxiety interventions and found they had no impact on students' test anxiety, but they did positively impact performance outcomes for both men and women. Their results suggested that men are less likely to declare that they are test-anxious, and that gender-based performance gaps may be for reasons other than differences in test anxiety. Núñez-Peña et al. (2016) showed that while self-reported levels of test anxiety were higher for women than for men, it did not affect academic performance. They concluded women students may have developed coping strategies during test situations. In agreement with this conclusion, Salehi et al. (2019) observed gender penalties (i.e., women were awarded lower grades on exams) mediated by test anxiety in lower division courses but did not observe gender penalties in upper division courses. The researchers explained their results by concluding that women were either being 'weeded out' at the introductory level or 'warming to' timed examinations through coping strategies. In fact, test anxiety may lead students to engage in productive coping strategies related to task-orientation, test preparation, and seeking social support (Carver et al., 1989). However, test anxious students are also more likely to engage in maladaptive coping processes, such as avoidance and emotion-focused coping, which negatively impact study behaviours and performance (Zeidner, 1995, 1998)).

What can instructors do to address test anxiety and maladaptive forms of coping that disproportionately impact women and first-generation students? Emotional regulation interventions during evaluative assessments, such as expressive writing and arousal reappraisal, may reduce the negative effects of anxiety and have been shown to reduce failure rates among low-income students (Rozek et al., 2019). Instructors can also provide students with several opportunities to engage in low-stakes assessments to externally lower the sense of risk students experience (Cotner & Ballen, 2017), lessening the likelihood that test-anxious students engage in unproductive behaviours.

4.3 What types of assessments do students prefer in their STEM courses, and are student assessment preferences related to test anxiety, gender, or generation in college?

Students prefer assessment practices known to reduce the impact of test anxiety. We initially hypothesized that students with high test anxiety will prefer to be assessed by methods known to reduce test anxiety. We found that students, regardless of their level of test anxiety, preferred assessment practices likely to mitigate the impacts of test anxiety, namely more assessments and different types of assessments. Specifically, 63% of the students we surveyed wanted their course to include more assessments, 23% of students did not want their course assessment to change, and 20% of students wanted different types of assessments. Furthermore, first-generation students were marginally more likely to ask for different types of assessments than their continuing-generation peers. However, we avoid over-interpreting this comparison of

assessment preferences between first- and continuing-generation students due to the small sample of first-generation students in our qualitative dataset. Below we place our results in existing literature investigating the impacts of changing how we assess students.

More assessments. Several studies attest to the benefits of multiple, low-stakes assessments over one or a few high-stakes tests. Cotner and Ballen (2017) showed that lowering the amount that single assessments account for in students' grades ('lowering the stakes') resulted in relatively equitable performance outcomes across three large biology courses. These findings are in line with observational data across STEM courses and universities (Koester et al., 2016; Salehi et al., 2019; Odom et al., 2021; Malespina & Singh, 2022).

Different types of assessments. The three most common assessments that students preferred included written reports, lower stakes assessments (e.g., quizzes and assignments/activities), and portfolios. Portfolios (*mappevurdering*) are comprised of multiple assignments that are assessed as a whole. Many of these different assessment types resemble formative, rather than summative, assessments. Broadly, formative assessments aim to monitor student learning and provide ongoing feedback that can be used to improve student learning (Wiggins, 1998; Dixson & Worrell, 2016). They help students identify the strengths and weaknesses of their content knowledge and point instructors towards the areas that students need most help (Dixson & Worrell, 2016). In contrast, summative assessments evaluate student learning at the end of an instructional unit (or sometimes, a course or study program) (Dixson & Worrell, 2016). Formative assessments are often low-stakes quizzes, assignments, and activities, whereas summative assessments include final exams, college entrance exams, and term papers and typically have high point values. However, the way in which instructors implement the assessment ultimately determines its purpose (Gardner et al., 2010). For example, if a test is used to help students improve their learning, then technically it is being used formatively. Or, if a portfolio is used as an assessment of student learning, then it is being used summatively.

While we highlight the differences between formative and summative assessments, ideally, the two should complement each other and both be used throughout a course. Formative assessments can be used to help students learn the material, and summative assessments can be used at the end of a learning unit to assess how much learning has taken place. With increasing class sizes in STEM, however, there has been a recent pendulum swing towards using exclusively summative assessments, especially timed, high-stakes, multiple choice exams. As seen here, students who are taking STEM courses express preference for more formative assessments in the form of quizzes, assignments/activities, or portfolios of assignments. These findings are significant because not only are they aligned with what education researchers have revealed about the benefits of formative assessment, but, by responding to student preferences in designing assessments, educators can promote student agency and give students the freedom to determine how they are assessed.

Following a humanist learning framework, obtaining student input about how they want to be assessed can be an important first step in shifting away from the dominant use of high-stakes summative assessments in STEM and towards incorporating assessments known to reduce test anxiety (von der Embse et al., 2018). Our findings align with a humanist approach to learning, specifically that students can be trusted to

know what they need to optimize their learning and benefit from having freedom in their STEM courses (Rogers & Freiberg, 1994). While previous research shows students respond positively when given the opportunity to choose how they will be evaluated (Chase, 2020; Clack & Dommett, 2021), future research will profit from explicit investigation into the extent this varies across different learning contexts (e.g., institutions, disciplines), student identities (e.g., on the basis of gender), and student affective profiles (e.g., levels of test anxiety). For example, a natural extension of the current work is to experimentally implement student recommendations into the STEM courses we studied and gauge how students respond to these changes - both in terms of performance outcomes, but also in their levels of test anxiety, science self-efficacy, and intention to pursue a STEM career.

4.4 Placing our results in context: assessment in Norway

Norwegian higher education presents an interesting case study for anxiety-mediated performance and assessment preferences. Norwegian students have traditionally been assessed via high-stakes (typically 4-5 hour) summative exams at the end of the school term. After the Quality Reform Act of 2001 in Norway (Regjeringen, 2001), there was a push to shift universities from “exam-driven institutions” to places with a greater emphasis on student learning, emphasizing active learning, constructive alignment (Biggs, 1996), and formative assessment. The government also removed an existing requirement for two evaluators for every graded assessment, stating that “the introduction of more frequent assessments and feedback to students make it less appropriate to use external evaluators to evaluate all exams” (Regjeringen, 2001, p. 32). Despite these reforms, grading policies exist in Norwegian higher education that can make frequent assessments logistically difficult. For example, teaching assistants cannot assist with assigning grades outside of tabulating assignments graded for completion. Students in Norway also have the right to appeal any grade, and the appeals process involves regrading from two different evaluators, including one external evaluator. Furthermore, in May 2021, the Norwegian government, under pressure from the National Student Union, voted unanimously to return to the requirement for two evaluators to assess all graded student work (Regjeringen, 2021). After heated public debate, this law was essentially rescinded in summer 2022, however not before several academic programs pre-emptively returned to grading via a single, summative high-stakes exam. In sum, how students are assessed is a current pressing topic in Norwegian higher education (Harlap et al., 2022). Our hope is that studies such as this one will contribute to constructive dialogue about how best to assess student learning.

Furthermore, albeit presumably not unique to Norwegian students, most students in the courses studied here concluded secondary education (*videregaende*) during the COVID-19 pandemic and its associated restrictions. During the pandemic, the national high-school exams were cancelled. Thus, these students had not experienced in-person high-stakes exams and therefore lacked practice in this type of assessment. Starting university with limited (or no) experience with high-stakes summative assessments may have exacerbated student test anxiety, a possibility that can be evaluated through follow-up work on test anxiety and performance.

5 Research limitations

This research has several limitations. First, in our qualitative prompt, we gave several examples of assessment practices (e.g., mandatory activities, quizzes). While we included this wording to stimulate student response, it also could have constrained student responses to only those suggestions. Furthermore, we recognize that student experiences with assessment practices both at university and prior to attending university likely impacted how students responded to our qualitative prompt. In Norway, schooling before university is characterized by low-stakes homework and tests with very few high-stakes exams. We were also limited by small sample sizes, especially with respect to student grades, which are generally difficult to access in Norway. Furthermore, these grades are reported as letters only, on an ordinal scale; a continuous scale, using either total points or percentages, might have given us greater resolution for interpreting student performance as a correlate of test anxiety. We also observed a gender and generation bias in our sample, with more women than men and more continuing-generation than first-generation students. Given our small and biased sample, we urge caution in drawing strong conclusions from this study alone and instead call for readers to interpret our results in context of the broader literature discussed. Finally, when we categorized students' preferred types of assessments, we did not distinguish between graded and non-graded assessments, even though students may see a distinction, and accountability through grading ensures more student engagement (Freeman et al., 2007). We also acknowledge that students likely experience test anxiety - often driven by fear of negative evaluation - on graded, versus ungraded, assessments.

6 Conclusions

Our findings, alongside the body of existing work on equitable assessment, provide instructors with some simple suggestions for assessing students in ways that are fair and support student learning. Instructors can provide multiple low-stakes, formative assessments (alone or in combination with higher-stakes exams) that allow students to monitor their own learning while also reducing anxiety (von der Embse et al., 2018) and possibly eliminating anxiety-driven performance deficits. These formative assessments do not all need to be *graded* to be effective. For example, they can be part of a pass/fail or competency-based grading system (Harlap et al., 2022). Alternatively, they can include peer assessment to minimize the grading burden. Instructors can also introduce tests in ways that minimize anxiety, either through simple writing prompts (Ramirez & Beilock, 2011) or reappraisal exercises (Jamieson et al., 2016) that allow students to combat some of the psychological threat that arises during high-stakes testing situations. Our results furthermore suggest that instructors allow students some agency in how they are assessed–either by choosing what assessments contribute to their grade in a course (contract grading), having some voice in the grading scale, or electing pass/fail grading over A-E grading. Investigating creative possibilities for equitable, student-centred assessment should be a priority for future work in STEM higher education – in Norway and around the globe.

Acknowledgements

We would like to thank the participants of the Emerging Research in STEM Education Mini-Symposium at the University of Bergen, the bioBERG Research Group, the Ballen lab, and the Auburn DBER group for feedback on our data analysis. Special thanks to Kristin Holtermann for accessing and de-identifying student information. Support for RAC was provided by a Mobility Grant from the Norwegian Research Council awarded to John Arvid Grytnes.

References

Ballen, C. J., & Holmegaard, H. T. (2019). With big data comes big responsibilities for science equity research. *Journal of Microbiology & Biology Education, 20*(1), 1–10. https://doi.org/10.1128/jmbe.v20i1.1643

Ballen, C. J., & Salehi, S. (2021). Mediation analysis in discipline-based education research using structural equation modeling: beyond "what works" to understand how it works, and for whom. *Journal of Microbiology & Biology Education*, *22*(2), e00108-21. https://doi.org/10.1128/jmbe.00108-21

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLOS ONE*, *12*(10), 1–14. https://doi.org/10.1371/journal.pone.0186419

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., & Zamudio, K. R. (2017). Enhancing diversity in undergraduate science: self-efficacy drives performance gains with active learning. *CBE—Life Sciences Education*, *16*(4), ar56. https://doi.org/10.1187/cbe.16-12-0344

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

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364.

Burnette, J. L., Hoyt, C. L., Russell, V. M., Lawson, B., Dweck, C. S., & Finkel, E. (2020). A growth mind-set intervention improves interest but not academic performance in the field of computer science. *Social Psychological and Personality Science, 11*(1), 107–116. https://doi.org/10.1177/1948550619841631

Büyüköztürk, S., Akgün, Ö. E., Özkahveci, Ö., & Demirel, F. (2004). The validity and reliability study of the Turkish version of the motivated strategies for learning questionnaire. *Educational Sciences: Theory & Practice, 4*(2), 231–237.

Canning, E. A., LaCosse, J., Kroeper, K. M., & Murphy, M. C. (2020). Feeling like an imposter: the effect of perceived classroom competition on the daily psychological experiences of first-generation college students. *Social Psychological and Personality Science*, *11*(5), 647–657. https://doi.org/10.1177/1948550619882032

Carey, R. L. (2014). A cultural analysis of the achievement gap discourse: challenging the language and labels used in the work of school reform. *Urban Education*, *49*(4), 440–468. https://doi.org/10.1177/0042085913507459

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: a theoretically based approach. *Journal of personality and social psychology, 56*(2), 267–283. https://doi.org/10.1037/0022-3514.56.2.267

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*(6), 569–592. https://doi.org/10.1016/j.learninstruc.2004.09.002

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

Caviola, S., Toffalini, E., Giofrè, D., Ruiz, J. M., Szűcs, D., & Mammarella, I. C. (2021). Math performance and academic anxiety forms, from sociodemographic to cognitive aspects: a meta-analysis on 906,311 participants. *Educational Psychology Review, 34*, 363–399. https://doi.org/10.1007/s10648-021-09618-5

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, *97*(2), 268–274. https://doi.org/10.1037/0022-0663.97.2.268

Chase, M. K. (2020). Student Voice in STEM Classroom Assessment Practice: A Pilot Intervention. Research & Practice in Assessment, 15(2), n2.

Clack, A., & Dommett, E. J. (2021). Student learning approaches: Beyond assessment type to feedback and student choice. Education Sciences, 11(9), 468.

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable?. *PLoS One*, *12*(12), e0189610. https://doi.org/10.1371/journal.pone.0189610

Cotner, S., Jeno, L. M., Walker, J. D., Jørgensen, C., & Vandvik, V. (2020). Gender gaps in the performance of Norwegian biology students: the roles of test anxiety and science confidence. *International Journal of STEM Education*, *7*, 55. https://doi.org/10.1186/s40594-020-00252-1

Culler R. E., & Holahan C. J. (1980). Test anxiety and academic performance: the effects of study-related behaviors. *Journal of Educational Psychology, 72*(1), 16–20. https://doi.org/10.1037/0022-0663.72.1.16

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory into Practice, 55*(2), 153–159. https://doi.org/10.1080/00405841.2016.1148989

Driessen, E. P., Beatty, A. E., & Ballen, C. J. (2022). Evaluating open-note exams: Student perceptions and preparation methods in an undergraduate biology class. *Plos one, 17*(8), e0273185. https://doi.org/10.1371/journal.pone.0273185

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research*, *12*(2), 020106.

Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. *Journal of anxiety disorders, 15*(3), 231–245.

Ewell, S.N., Driessen, E.P., Grogan, W., Johnston, Q., Ferdous, S., Mehari, Y., Peart, A., Seibenhener, M., & Ballen, C.J. (2022). A Comparison of study behaviors and metacognitive evaluation used by lower-level and upper-level biology students. In revision.

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: the processing efficiency theory. *Cognition & Emotion, 6*(6), 409–434. https://doi.org/10.1080/02699939208409696

Feiz, P., & Hooman, H. A. (2013). Assessing the Motivated Strategies for Learning Questionnaire (MSLQ) in Iranian students: construct validity and reliability. *Procedia-Social and Behavioral Sciences, 84*, 1820–1825. https://doi.org/10.1016/j.sbspro.2013.07.041

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., ... & Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education, 6*(2), 132-139.

Freire, P. (1970). Pedagogy of the oppressed. London: Penguin Random House.

Gardner, J., Harlen, W., Hayward, L., Stobart, G., & Montgomery, M. (2010). *EBOOK: Developing Teacher Assessment*. McGraw-Hill Education (UK).

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open book? A comparison of the effects of exam types on performance, retention, and anxiety. *Online Submission, 2*(8), 469–478.

Global gender gap report. (2022). *World Economic Forum*. https://www.weforum.org/reports/global-gender-gap-report-2022. Accessed 23 December 2022.

Guadier-Diaz, M. M., Sinisterra, M., & Muscatell, K. A. (2019). Motivation, belongingness, and anxiety in neuroscience undergraduates: emphasizing first-generation college students. *Journal of Undergraduate Neuroscience Education, 17*(2), A145–A152.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology, 106*(2), 375–389. https://doi.org/10.1037/a0034679

Harlap, Y., Jørgensen, C., & Cotner, S. (2022). Maintaining quality assessment practices in Norwegian higher education after the two-evaluator law. *Nordic Journal of STEM Education, 6*(1). https://doi.org/10.5324/njsteme.v6i1.4873

Hammarlund, S. P., Scott, C., Binning, K. R., & Cotner, S. (2022). Context matters: how an ecological-belonging intervention can reduce inequities in STEM. *BioScience, 72*(4), 387–396. https://doi.org/10.1093/biosci/biab146

Hansen, M. N. (2010). Utdanningspolitikk og ulikhet. *Tidsskrift for Samfunnsforskning*, *51*(1), 101–133. https://doi.org/10.18261/issn1504-291x-2010-01-05

Harris, R. B., Grunspan, D. Z., Pelch, M. A., Fernandes, G., Ramirez, G., & Freeman, S. (2019). Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course?. *CBE—Life Sciences Education, 18*(3), ar35. https://doi.org/10.1187/cbe.18-05-0083

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Hill, C., Corbett, C., & St. Rose, A. (2009). *Why so few? women in science, technology, engineering, and mathematics*. Washington, DC: American Association of University Women.

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: a test of different models. *Anxiety, Stress & Coping*, *10*(3), 219–244. https://doi.org/10.1080/10615809708249302

Holseter, A. M. R. (2021). Foreldrenes utdanningsnivå betyr fremdeles mye for gjennomføring. *Statistisk Sentralbyrå, 13*, 22.

Hovdhaugen, E. (2009). Transfer and dropout: different forms of student departure in Norway. *Studies in Higher Education*, *34*(1), 1–17. https://doi.org/10.1080/03075070802457009

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Jakešová, J., & Hrbáčková, K. (2014). The Czech adaptation of motivated strategies for learning questionnaire (MSLQ). *Asian Social Science*.

Jamieson, J. P., Peters, B. J., Greenwood, E. J., & Altose, A. J. (2016). Reappraising stress arousal improves performance and reduces evaluation anxiety in classroom exam situations. *Social Psychological and Personality Science, 7*(6), 579–587. https://doi.org/10.1177/1948550616644656

Jenifer, J. B., Rozek, C. S., Levine, S. C., & Beilock, S. L. (2022). Effort (less) exam preparation: math anxiety predicts the avoidance of effortful study strategies. *Journal of Experimental Psychology: General, 151*(10), 2534–2541. https://doi.org/10.1037/xge0001202.

Kader, A. A. (2016). Debilitating and facilitating test anxiety and student motivation and achievement in principles of microeconomics. *International Review of Economics Education, 23*, 40–46. https://doi.org/10.1016/j.iree.2016.07.002

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kirkland, K., & Hollandsworth, J. G. (1980). Effective test taking: skills-acquisition versus anxiety-reduction techniques. *Journal of Consulting and Clinical Psychology, 48*(4), 431–439. https://doi.org/10.1037/0022-006X.48.4.431.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1), rm1. https://doi.org/10.1187/cbe.18-04-0064

Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of gendered performance difference in introductory STEM courses. *arXiv preprint arXiv:1608.07565*.

Larsen, M. H. (2018). Foreldrenes utdanning gir store utslag i ungdomsskolekarakterene. https://forskning.no/skole-og-utdanning-ntb-utdanning/foreldrenes-utdanning-gir-store-utslag-i-ungdomsskolekarakterene/1221455 Accessed 23 December 23 2022

Lenth, R. V. (2022). emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.3. https://CRAN.R-project.org/package=emmeans

Lowe, P. A. (2015). Should test anxiety be measured differently for males and females? examination of measurement bias across gender on measures of test anxiety for middle and high school, and college students. *Journal of Psychoeducational Assessment*, *33*(3), 238–246. https://doi.org/10.1177/0734282914549428

Lundeberg, M. A., Kang, H., Wolter, B., DelMas, R., Armstrong, N., Borsari, B., ... & Herreid, C. F. (2011). Context matters: increasing understanding with interactive clicker case studies. *Educational technology research and development*, *59*(5), 645-671. https://doi.org/10.1007/s11423-010-9182-1

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. https://doi.org/10.1111/asap.12033

Malespina, A., & Singh, C. (2022). Gender differences in test anxiety and self-efficacy: why instructors should emphasize low-stakes formative assessments in physics courses. *European Journal of Physics, 43*(3), 035701. https://doi.org/10.1088/1361-6404/ac51b1

Maloney, E. A., Sattizahn, J. R., & Beilock, S. L. (2014). Anxiety and cognition. *WIREs Cognitive Science*, *5*(4), 403–411. https://doi.org/10.1002/wcs.1299

McClendon, R. C. (1996). Motivation and cognition of preservice teachers: MSLQ. *Journal of Instructional Psychology, 23*(3), 216.

Milner IV, H. R. (2012). Beyond a test score: explaining opportunity gaps in educational practice. *Journal of Black Studies*, *43*(6), 693–718. https://doi.org/10.1177/0021934712442539

Mohamadi, M., Alishahi, Z., & Soleimani, N. (2014). A study on test anxiety and its relationship to test score and self-actualization of academic EFL students in Iran. *Procedia - Social and Behavioral Sciences*, *98*, 1156–1164. https://doi.org/10.1016/j.sbspro.2014.03.529

Núñez-Peña, M. I., Suárez-Pellicioni, M., & Bono, R. (2016). Gender differences in test anxiety and their impact on higher education students' academic achievement. *Procedia-Social and Behavioral Sciences, 228*, 154-160. https://doi.org/10.1016/j.sbspro.2016.07.023

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd edition). New York NY: MacGraw-Hill.

O'Brien, L. T., Bart, H. L., & Garcia, D. M. (2020). Why are there so few ethnic minorities in ecology and evolutionary biology? challenges to inclusion and the role of sense of belonging. *Social Psychology of Education, 23*(2), 449–477. https://doi.org/10.1007/s11218-019-09538-x

Odom, S., Boso, H., Bowling, S., Brownell, S., Cotner, S., Creech, C., Drake, A. G., Eddy, S., Fagbodun, S., Hebert, S., James, A. C., Just, J., St. Juliana, J. R., Shuster, M., Thompson, S. K., Whittington, R., Wills, B. D., Wilson, A. E., Zamudio, K. R., ... & Ballen, C. J. (2021). Meta-analysis of gender performance gaps in undergraduate natural science courses. *CBE—Life Sciences Education*, *20*(3), ar40. https://doi.org/10.1187/cbe.20-11-0260

Oludipe, B. (2009). Influence of test anxiety on performance levels on numerical tasks of secondary school physics students. A*cademic Leadership: The Online Journal, 7*(4), 19.

Piaget, J. (1972). *The psychology of the child*. New York NY: Basic Books.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. University of Michigan.

Pretorius, L., van Mourik, G. P., & Barratt, C. (2017). Student choice and higher-order thinking: Using a novel flexible assessment regime combined with critical thinking activities to encourage the development of higher order thinking. *International Journal of Teaching and Learning in Higher Education, 29*(2), 389-401.

Quinn, D. M. (2020). Experimental effects of "achievement gap" news reporting on viewers' racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher*, *49*(7), 482–492. https://doi.org/10.3102/0013189X20932469

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science, 331*(6014), 211–213. https://doi.org/10.1126/science.1199427

Regjeringen. (2001). Gjør din plikt – Krev din rett. *St. meld. nr. 27 (2000-2001)*. Kirke-, utdannings- og forskningsdepartementet. https://www.regjeringen.no/no/dokumenter/stmeld-nr-27-2000-2001-/

Regjeringen. (2021). *Prop. 111L (2020-2021)*. Kunnskapsdepartementet. https://www.regjeringen.no/no/dokumenter/prop.-111-l-20202021/

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387.

Rogers, C., & Freiberg, J. (1994). *Freedom to learn* (3rd edition). Columbus OH: Merrill.

Rosenberg, M. B., Hilton, M. L., & Dibner, K. A. (2018). *Indicators for monitoring undergraduate STEM education.* National Academies Press.

Rosseel, Y. (2012). lavaan: An *R* package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rozek, C. S., Ramirez, G., Fine, R. D., & Beilock, S. L. (2019). Reducing socioeconomic disparities in the STEM pipeline through student emotion regulation. *Proceedings of the National Academy of Sciences, 116*(5), 1553–1558. https://doi.org/10.1073/pnas.1808589116

Salehi, S., Berk, S. A., Brunelli, R., Cotner, S., Creech, C., Drake, A. G., ... & Ballen, C. J. (2021). Context matters: social psychological factors that underlie academic performance across seven institutions. *CBE—Life Sciences Education*, *20*(4), ar68. https://doi.org/10.1187/cbe.21-01-0012

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Harcombe, W., McGaugh, S., Wassenberg, D., Yonas, A., & Ballen, C. J. (2019). Gender performance gaps across different assessment methods and the underlying mechanisms: the case of incoming preparation and test anxiety. *Frontiers in Education*, *4*, 107. https://doi.org/10.3389/feduc.2019.00107

Salehi, S., Cotner, S., & Ballen, C. J. (2020). Variation in incoming academic preparation: consequences for minority and first-generation students. *Frontiers in Education, 5*, 552364. https://doi.org/10.3389/feduc.2020.552364

Shukla, S. Y., Theobald, E. J., Abraham, J. K., & Price, R. M. (2022). Reframing Educational Outcomes: moving beyond Achievement Gaps. *CBE—Life Sciences Education*, *21*(2), es2. https://doi.org/10.1187/cbe.21-05-0130

Stanger-Hall, K. F. (2017). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

Steele, C. M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066x.52.6.613

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

Tangney, S. (2014). Student-centered learning: a humanist perspective. *Teaching in Higher Education*, *19*(3), 266–275. https://doi.org/10.1080/13562517.2013.860099

Theobald, M., Breitwieser, J., & Brod, G. (2022). Test anxiety does not predict exam performance when knowledge is controlled for: strong evidence against the interference hypothesis of test anxiety. *Psychological Science, 33*(12), 2073–2083. https://doi.org/10.1177/09567976221119391

Thompson, S. K., Hebert, S., Berk, S., Brunelli, R., Creech, C., Drake, A. G., ... & Ballen, C. J. (2020). A call for data-driven networks to address equity in the context of undergraduate biology. *CBE—Life Sciences Education*, *19*(4), mr2. https://doi.org/10.1187/cbe.20-05-0085

Trapani, J., & Hale, K. (2022). Higher education in science and engineering. Science & engineering indicators 2022. NSB-2022-3. *National Science Foundation*.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: a 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Warburton, E. C., Bugarin, R., & Nunez, A.-M. (2001). Bridging the gap: academic preparation and postsecondary success of first-generation students. Statistical analysis report. *Postsecondary Education Descriptive Analysis Reports*. Washington, DC: NCES.

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: a meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment, 20*(1), 23–45. doi: 10.1080/10627197.2015.997614

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wiggins, G. P. (1998). *Educative assessment: designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Wine J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92–104. https://doi.org/10.1037/h0031332

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934. https://doi.org/10.1177/0013164413495237

Zeidner, M. (1995). Adaptive coping with test situations: a review of the literature. *Educational Psychologist, 30*(3), 123–133. https://doi.org/10.1207/s15326985ep3003_3

Zeidner, M. (1998). *Test anxiety: the state of the art*. New York: Plenum.

Zeidner, M. (2010). Test anxiety. *The Corsini Encyclopedia of Psychology*. https://doi.org/10.1002/9780470479216.corpsy0984

Zeidner, M. (2014). Test anxiety. *The Wiley Handbook of Anxiety Disorders*. https://doi.org/10.1002/9781118775349.ch28

Zhang, J., Zhao, N., & Kong, Q. P. (2019). The relationship between math anxiety and math performance: a Meta-analytic investigation. *Frontiers in Psychology*, *10*, 1613. https://doi.org/10.3389/fpsyg.2019.01613

Zoller, U., & Ben-Chaim, D. (1990). Gender differences in examination-type preferences, test anxiety, and academic achievements in college science education–a case study. *Science Education, 74*(6), 597–608.