Development of the Mentoring in Undergraduate Research Survey

Lisa B. Limeri,[†] Nathan T. Carter,[‡] Riley A. Hess,[§] Trevor T. Tuma,[‖] Isabelle Koscik,[¶] Alexander J. Morrison,[¶] Briana Outlaw,[¶] Kathren Sage Royston,[¶] Benjamin H. T. Bridges,[¶] and Erin L. Dolan[¶]*

[†]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409; [‡]Department of Psychology, Michigan State University, East Lansing, MI 48824; [§]Department of Psychology, [‖]Warnell School of Forestry and Natural Resources, and [¶]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602

ABSTRACT

Here we present the development of the Mentoring in Undergraduate Research Survey (MURS) as a measure of a range of mentoring experienced by undergraduate science researchers. We drafted items based on qualitative research and refined the items through cognitive interviews and expert sorting. We used one national dataset to evaluate the internal structure of the measure and a second national dataset to examine how responses on the MURS related to theoretically relevant constructs and student characteristics. Our factor analytic results indicate seven lower order forms of mentoring experiences: abusive supervision, accessibility, technical support, psychosocial support, interpersonal mismatch, sexual harassment, and unfair treatment. These forms of mentoring mapped onto two higher-order factors: supportive and destructive mentoring experiences. Although most undergraduates reported experiencing supportive mentoring, some reported experiencing absence of supportive as well as destructive experiences. Undergraduates who experienced less supportive and more destructive mentoring also experienced lower scientific integration and a dampening of their beliefs about the value of research. The MURS should be useful for investigating the effects of mentoring experienced by undergraduate researchers and for testing interventions aimed at fostering supportive experiences and reducing or preventing destructive experiences and their impacts.

INTRODUCTION

Proponents of undergraduate STEM education reform advocate for widespread involvement of undergraduates in research because of the potential for undergraduates to experience professional, academic, and personal benefits (American Association for the Advancement of Science, 2011; Byars-Winston *et al.*, 2015; Estrada *et al.*, 2018). Undergraduate research experiences (UREs) are also increasingly recognized for their capacity to promote the integration of students into the scientific community, especially students from marginalized or minoritized backgrounds (Estrada *et al.*, 2011; Hernandez *et al.*, 2017; Hernandez, Woodcock *et al.*, 2018). Multiple qualitative and quantitative studies have shown that mentoring plays a critical role in STEM undergraduate researchers' personal and professional development (Thiry and Laursen, 2011; Aikens *et al.*, 2016; Estrada *et al.*, 2018; Hernandez, Hopkins *et al.*, 2018; Joshi *et al.*, 2019). Studies examining mentoring in UREs have generally focused on positive mentoring that undergraduates experience (Aikens *et al.*, 2016, 2017; Hernandez *et al.*, 2017). However, studies from both workplace and academic settings have shown that not all mentoring experiences are positive.

Mentoring, like any interpersonal relationship, can include dysfunctional elements or problematic events, which are collectively referred to as negative mentoring (Kram, 1983; Scandura, 1998; Eby *et al.*, 2000; Simon and Eby, 2003). Workplace mentees

Ross Nehm, *Monitoring Editor*
Submitted Jul 26, 2023; Revised Apr 29, 2024; Accepted May 1, 2024
CBE Life Sci Educ June 1, 2024 23:ar26
DOI:10.1187/cbe.23-07-0141

*Address correspondence to: Erin L. Dolan (eldolan@uga.edu).

report problems with mentors such as personality mismatches, mentor neglect, and mentor sabotage, as well as mentors lacking expertise (Eby *et al.*, 2000, 2004; Simon and Eby, 2003). Doctoral students in the life sciences also report negative mentoring experiences with their research advisors, including inaccessibility, deceit, and problematic supervisory styles such as micromanagement (Tuma *et al.*, 2021). A few studies of mentoring in UREs have noted variation in the quality of mentoring, such as absenteeism, unrealistic expectations, or insufficient guidance from mentors (Bernier *et al.*, 2005; Dolan and Johnson, 2010; Thiry and Laursen, 2011). To define negative mentoring in undergraduate research, we previously conducted a qualitative study of negative mentoring experienced by undergraduate researchers in the life sciences (Limeri *et al.*, 2019a). We identified seven types of negative mentoring experiences: absenteeism, abuse of power, interpersonal mismatch, lack of technical support, lack of psychosocial support, misaligned expectations, and unequal treatment. Although this research characterized negative mentoring experiences among undergraduate researchers, it did not provide direct evidence of the effects of these experiences.

Studies from the workplace suggest that negative mentoring harms mentees, decreasing their job satisfaction and increasing their stress as well as their intentions to leave their jobs (Eby and Allen, 2002). One study indicated that workplace negative mentoring may be so damaging that mentees who experience it may be worse off than if they had no mentor at all (Ragins *et al.*, 2000). Negative mentoring is most strongly associated with negative mentee outcomes when the mentoring relationships are assigned rather than formed organically (Eby and Allen, 2002). This is concerning because formal assignment is often how mentoring relationships are formed in UREs; either a faculty member assigns an undergraduate to a graduate or postdoctoral mentor or an undergraduate is assigned to a faculty member's research group (Dolan and Johnson, 2009; Limeri *et al.*, 2019b; Erickson *et al.*, 2022).

Evidence indicates that quality mentorship during UREs is especially beneficial for students from marginalized or minoritized backgrounds because UREs promote a sense of fit with the scientific community (Hurtado *et al.*, 2009; Estrada *et al.*, 2011, 2018; Hernandez *et al.*, 2017). These findings raise concerns that negative mentoring experiences may prevent rather than promote a sense of belonging. Such experiences may disproportionately harm students already facing barriers to their integration in STEM and exacerbate inequities.

Measuring the Range of Mentoring Experiences

Given the widespread recommendations to involve undergraduate STEM students in research and the potential for negative mentoring to cause harm (Gentile *et al.*, 2017), it is critical to understand the mentoring that undergraduate researchers experience and how it affects them. Accomplishing this requires an instrument with strong evidence of its utility for measuring the range of mentoring experienced by undergraduate researchers. Here we report the development of such a measure: the Mentoring in Undergraduate Research Survey (MURS). We opted to develop a new measure because we were unable to identify existing tools suitable for measuring the diversity of mentoring experienced by undergraduate researchers that we observed in our qualitative work (Limeri *et al.*, 2019a).

Several instruments have been used to measure mentorship quality (reviewed in Byars-Winston and Dahlberg, 2019); yet the majority lack validity evidence of response processes, internal structure, or relations to other variables (Hernandez, 2018). Few if any have been designed or used to assess mentorship quality at the undergraduate level. Furthermore, most have been designed to assess positive mentorship, and thus are likely to fall short of capturing key elements of negative mentoring experiences. For example, the mentoring competency assessment (MCA) was developed to evaluate research mentors' skills before and after a mentoring training program (Fleming *et al.*, 2013). The scale asks mentees to rate their mentor's ability on 26 skills associated with six mentor competencies: maintaining effective communication, aligning expectations, assessing understanding, addressing diversity, fostering independence, and promoting professional development. These competencies align with some but not all of the negative mentoring experienced by undergraduate researchers in our prior work. For instance, we found that students reported mismatches with their mentor's personality or work style (Limeri *et al.*, 2019a), which are unrelated to mentor skills per se. In addition, undergraduates most often reported mentor absenteeism as a form of negative mentoring, which mentees often attributed to mentors being overcommitted rather than unskilled. In sum, a new measure is needed to investigate how negative mentoring experiences affect undergraduates and the outcomes they realize from participating in UREs.

Measurement Validity Framework

Here we report the development of the MURS as a measure of mentoring experiences for use with undergraduate science researchers. To guide the development process, we adopted Kane's argument-based approach to measurement validity (Kane *et al.*, 1999). In this framework, validity is not an inherent property of a measurement instrument, but rather an argument for a proposed interpretation of responses to an instrument, which must be supported by evidence. The interpretive argument in this case is that undergraduate researchers' responses to the items on the MURS are indicative of students' mentoring experiences, such that students with higher scores experienced more of that form of mentoring. The process of building the validity argument involves identifying and providing evidence in support of the assumptions underlying this argument. Here we provide evidence to support this and other assumptions to build a validity argument for the MURS.

METHODS AND RESULTS

Studies of mentoring experiences as well as related experiences of abusive supervision (i.e., supervisor display of sustained, hostile verbal and nonverbal behaviors) and workplace incivility (i.e., mild but consistently rude or impolite behavior of coworkers) have primarily operationalized these phenomena in terms of recipients' perceptions (Eby *et al.*, 2013; Schilpzand *et al.*, 2016; Tepper *et al.*, 2017). Although perceptions have been criticized for their lack of objectivity (Linn *et al.*, 2015; Tepper, 2000), we have chosen to use this same approach here for multiple reasons. First, negative mentoring may not always be visible, and directly observing mentoring would be intrusive and impractical. Second, mentors may not be aware that their behaviors are problematic and may not be willing to report less-than-ideal behavior, making mentor reports of negative mentoring equally subjective. Finally, mentee perceptions of mentoring have been shown to fundamentally alter these relationships and to have long-term effects on mentee outcomes (Scandura, 1998; Eby and Allen, 2002; Eby *et al.*, 2008, 2010). Thus, our intent is to measure undergraduate researchers' perceptions of their mentoring experiences.

We carried out the process of developing and collecting validity evidence for the MURS over three phases: substantive, structural, and external (Benson, 1998). All phases of the study were reviewed and determined to be exempt by the University of Georgia Institutional Review Board (STUDY00004954). For ease of reading, we present the methods and results together for each phase. In the final phase, we also begin to investigate how mentoring experiences influence undergraduate researchers' integration into the scientific community.

Substantive Phase

Our aim with this phase was to collect evidence that the MURS aligned with construct we intended to measure and that respondents interpreted MURS items as intended (Messick, 1995). We started by defining and characterizing negative mentoring by identifying observations that reflect the construct (Benson, 1998). Specifically, we carried out a qualitative characterization of negative mentoring experienced by undergraduate researchers to define the content domain of the construct (Limeri *et al.*, 2019a). This work was a useful foundation for capturing a range of mentoring experiences because undergraduates reported both the absence of supportive experiences and mentor behaviors, characteristics, or interactions they experienced as actively harmful or destructive. For comprehensibility and ease of comparison, we present our methods and results using a common, negative valence such that supportive experiences are described in terms of their absence and destructive experiences are described in terms of their presence.

We drafted 107 survey items that corresponded to the seven dimensions of mentoring experiences identified by Limeri and colleagues (2019a): absenteeism, which we renamed inaccessibility (13 items); abuse of power, which we renamed abusive supervision (25 items); interpersonal mismatch (13 items); insufficient technical support (12 items); insufficient psychosocial support (14 items); misaligned expectations (18 items); and unequal treatment (12 items). We also adapted five items to represent an eighth dimension, sexual harassment, resulting in 112 items altogether. The sexual harassment items were preceded with a content warning and based on items previously used to measure undergraduates' experiences with sexual harassment in academic settings (Aycock *et al.*, 2019).

We pilot tested the 112 items by conducting cognitive interviews with undergraduate researchers, which provided evidence that students understood and responded to the items as intended. Using a screening survey (see Supplemental Materials), we recruited 32 participants from 14 institutions who had experienced a range of mentoring quality, from mostly positive to mostly negative. Of these, we selected 15 participants from a diverse group of 11 institutions: six very high research activity, one high research activity, two master's-granting institutions, one community college, and one research institute; three of the institutions were classified as minority-serving (Indiana University Center for Postsecondary Research, n.d.). Participants were compensated with a $25 gift card. Because of the large number of items, each participant reviewed only a subset (one or two dimensions for a total of 15–25 items) such that each item was reviewed by three or four participants. Based on the cognitive interviews (questions provided in Supplemental Materials), we refined and revised the items. We ultimately selected 57 items that were most clearly and consistently interpreted and best represented the range of the construct.

As a final step in the substantive phase, we conducted a sorting activity with nine individuals to provide further evidence of the dimensionality of the MURS (Nahm *et al.*, 2002). These individuals were selected based on their expertise in mentoring research or extensive experience mentoring undergraduate researchers. Specifically, we provided the experts with a list of the 57 items in random order, along with definitions for each of the eight dimensions. We then asked the experts to assign each of the items into one of the eight dimensions they thought the item fit best. We set 70% agreement among the experts as a threshold for retaining the item as is; 40 items passed this threshold. We also asked the experts to indicate their confidence in their sorting of each item and to offer their expert judgment of the relevance of the item to the assigned dimension. For the 40 items with high agreement, the associated certainty and relevance ratings were high (i.e., 70% threshold for ratings of "high" relevance and certainty was reached for all of these items), indicating that we were capturing the main ideas underpinning each dimension. We reworded the remaining 17 items to address ambiguities, producing 57 items that reflected the eight dimensions. Because we used expert feedback to reword the remaining items to better reflect the intended dimensions, we did not subject them to the sorting task again.

Structural Phase

Our aim with this phase was to generate evidence of the internal structure of the MURS (Messick, 1995). To accomplish this, we examined the extent to which the observed variables (i.e., item responses) covaried among themselves and we compared that structure with the theorized seven dimensions of undergraduate mentoring experiences plus the eighth dimension of sexual harassment (Benson, 1998). We also collected personality data based on the Big Five model of personality traits, namely openness, conscientiousness, extraversion, agreeableness, and neuroticism, which is the dominant model of personality structure in psychological research (John, 2021). We reasoned that undergraduates might vary in their perceptions or reporting of negative experiences as result of their personality traits. For instance, the trait of neuroticism includes the tendency toward negative feelings. Individuals high on neuroticism can interpret ordinary situations as threatening (Widiger and Oltmanns, 2017) and show heightened sensitivity to social cues and relationship conflict (Denissen and Penke, 2008). Thus, undergraduates with elevated levels of neuroticism may experience interactions with mentors more negatively. Individuals high on conscientiousness, or the tendency to be diligent and take obligations seriously, tend to be perceived as more engaged in their jobs and their work is more highly rated (Bakker *et al.*, 2012). Thus, undergraduates high on conscientiousness may garner more accolades and support from mentors, reducing their likelihood of reporting negative experiences. By analyzing relationships between responses on the MURS and personality traits, we sought to explore whether the MURS was measuring facets of personality that might make individuals more or less likely to report negative experiences with mentors.

Recruitment and Data Collection

We recruited by email a national sample of undergraduates who had indicated they had completed at least one term (quarter, semester, summer) of mentored research within the past year to respond to the 57 MURS items (Table 1). We received

TABLE 1: Demographic information for participants surveyed in the structural and external phases[a]

Demographic	Structural phase participants (n = 521)	External phase participants (n = 348)
Institution type	*Reported by 518 (99%) at 60 institutions*	*Reported by 348 (100%) at 32 institutions*
Very high research activity	347 (67%) at 22 institutions	284 (82%) at 22 institutions
High research activity	57 (9%) at 6 institutions	53 (15%) at 4 institutions
Doctoral universities	0	1 (0.3%) at 1 institution
Masters-Granting	30 (5.8%) at 9 institutions	7 (2.0%) at 2 institutions
Primarily undergraduate	60 (12%) at 12 institutions	3 (0.9%) at 3 institutions
Community college	24 (4.6%) at 11 institutions	0
Race/Ethnicity	*Reported by 507 (97%)*	*Reported by 345 (99%)*
American Indian or Alaskan Native	6 (1.2%)	6 (1.7%)
Black or African American	23 (4.5%)	21 (6.1%)
Hispanic or Latinx	53 (11%)	55 (16%)
East Asian	102 (20%)	65 (19%)
South Asian	96 (19%)	54 (16%)
Middle Eastern or North African	16 (3.2%)	14 (4.1%)
Native Hawaiian or Pacific Islander	8 (1.6%)	3 (0.9%)
White	254 (50%)	192 (56%)
Gender	*Reported by 516 (99%)*	*Reported by 343 (99%)*
Man	149 (28.9%)	89 (26%)
Woman	362 (70.2%)	254 (74%)
Another gender identity	3 nonbinary; 2 gender fluid (1.0%)	1 nonbinary; 1 gender nonconforming (0.6%)
Parental education	*516 (99%)*	*343 (99%)*
No parents with a 4-year degree	116 (22.5%)	75 (22%)
At least one parent with a 4-year degree	146 (28.3%)	91 (27%)
At least one parent with a graduate degree	254 (49.2%)	182 (53%)
Discipline	*Reported by 511 (98%)*	*Reported by 345 (99%)*
Life Sciences	356 (69.7%)	245 (71%)
Chemistry	53 (10.4%)	35 (10%)
Engineering/Computer Science	51 (10.0%)	23 (6.7%)
Physics	9 (1.8%)	11 (3.1%)
Geosciences	7 (1.4%)	8 (2.3%)
Social Sciences	16 (3.1%)	14 (4.1%)
Allied Health	8 (1.6%)	3 (0.9%)
Interdisciplinary STEM	9 (1.8%)	4 (1.2%)
Math	2 (0.4%)	2 (0.6%)
Mentor's position	*Reported by 518 (99%)*	*Reported by 343 (99%)*
Faculty	325 (62.7%)	197 (57%)
Postdoctoral associate	68 (13.1%)	38 (11%)
Graduate student	84 (16.2%)	84 (24%)
Undergraduate student	26 (5.0%)	24 (7.0%)
Prior research experience	*Reported by 515 (99%)*	NA
None	18 (3.5%)	
1 term	108 (21.0%)	
2 terms	122 (23.7%)	
3 terms	66 (12.8%)	
More than 3 terms	201 (39.0%)	

[a]Institution type was determined using the Carnegie Classification of Institutions (Indiana University Center for Postsecondary Research, n.d.). Racial/ethnic identity counts may not add up to 100% because participants could select multiple racial/ethnic identities.

573 survey responses in total, of which 16 did not consent to be included in the study and thus were removed from the analysis. We included two attention checks (items that directed respondents to select a particular response, e.g., “This is a control question, please select ‘strongly agree’”) to screen out responses that reflected insufficient attention (DeSimone *et al.*, 2015). Respondents had to respond to both attention checks accurately to be included in the analysis; 36 responses were excluded because they did not pass one or both attention checks. Thus, the final analytic sample was $n = 521$. Students took an average of 15 min to complete the survey and were compensated with a $10 gift card. Our survey included our 57 MURS items as well as a 20-item measure of the five-factor model of personality (mini-IPIP; Donnellan *et al.*, 2006) and a series of demographic questions (see Supplemental Materials).

Factor Analysis

To examine the internal structure of the MURS, we estimated an eight-factor confirmatory factor model for ordinal indicators using diagonally weighted least squares (DWLS) estimation in the “lavaan” package (Rosseel, 2012) in R statistical software (R Core Team, 2021). Confirmatory factor analysis (CFA) is appropriate for established measures or when there is a theoretically-grounded reason to hypothesize about the factor structure. In contrast, exploratory factor analysis is appropriate when the researchers do not have a priori hypotheses about the factor structure of the items. Because we had theorized dimensions during our qualitative study and the substantive phase, we had a priori expectations about the factor structure. Therefore, CFA is a more useful analytic strategy because it allowed us to test our hypothesized factor structure. For more on this, see Knekta *et al.* (2019). We evaluated model fit holistically by considering both absolute and incremental indicators of fit: root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), Tucker–Lewis index (TLI), and comparative fit index (CFI). We evaluated model-data fit using criteria recommended by Hu and Bentler (1999): CFA > 0.95, TLI > 0.95, SRMR < 0.08, and RMSEA < 0.06. We also used χ^2 tests to evaluate the difference between the hypothesized and the actual observed results and compare alternative models.

Given that *inaccessibility*, *insufficient technical support*, *unclear expectations,* and *insufficient psychosocial support* were all measured by positively-phrased items, we reverse-scored them for analysis. We avoided having items with opposite valences within the same dimension to avoid introducing construct-irrelevant variance (e.g., error due to respondents misreading a negatively-worded item) (Roszkowski and Soven, 2010). Thus, the response scale was 1 to 5, with higher values indicating more negative mentoring experiences (1 = strongly disagree to 5 = strongly agree with the negative version of the statement). We also recoded all items belonging to the *sexual harassment* factor to be dichotomous (0 = Never; 1 = Any frequency greater than “Never”) due to very low endorsement rates (95–98% of respondents chose “Never” for these items).

This model showed good fit to the data, χ^2 (1511) = 2,868.95, $p < 0.001$, RMSEA = 0.044 (95% confidence interval [CI]: 0.042–0.047), CFI = 0.97, TLI = 0.97, SRMR = 0.075, but resulted in two nonadmissible solution problems. First, the correlation between the *technical support* and *clear expectations* factors was estimated at 0.96, causing the latent variable covariance matrix to be not positive definite (i.e., at least one factor in the model could be fully explained by a linear combination of the other factors). Second, the loading for one *sexual harassment* item (“My mentor made sexual comments about me”) was estimated at 1.02, resulting in a negative error variance. To resolve these problems, we collapsed the *technical support* and *clear expectations* factors into a single factor, and deleted the item with a loading greater than 1.0. The resulting seven-factor model showed similar model-data fit, χ^2(1463) = 2,922.69, $p <$ 0.001, RMSEA = 0.047 (95% CI: 0.044–0.049), CFI = 0.97, TLI = 0.96, SRMR = 0.074. One item from the *abusive treatment* factor was found to have a loading less than 0.40 (l = 0.36), and this item was deleted. After deletion, fit was similar, χ^2(1409) = 2810.51, $p < 0.001$, RMSEA = 0.046 (95% CI: 0.044–0.049), CFI = 0.97, TLI = 0.97, SRMR = 0.068.

The addition of the expected higher-order factors, supportive and destructive, worsened model-data fit, χ^2(1422) = 3402.56, $p < 0.001$, RMSEA = 0.055 (95% CI: 0.053–0.057), CFI = 0.95, TLI = 0.95, SRMR = 0.083. To further inspect this issue, we analyzed the covariance matrix of the second-order latent variables using exploratory methods. The scree plot (see Supplemental Materials) suggested either two or three factors, as did other indicators of factor structure. Specifically, the Very Simple Structure (VSS) statistics, Velicer’s minimum average partial (MAP) test, and empirical Bayesian information criterion (BIC) suggested two factors whereas the sample-size-adjusted BIC suggested three[1]. Therefore, we examined both solutions. The two-factor model suggested the hypothesized two-factor structure was largely supported with one exception: the *interpersonal mismatch* factor cross-loaded onto both higher-order factors. The three-factor model suggested that *interpersonal mismatch* was a higher-order factor unto itself. To get more precise estimates we respecified our higher-confirmatory factor model for ordinal responses to include the cross-loading (Model A), and another model which specified *interpersonal mismatch* as its own factor (Model B). Model A showed good fit to the data, χ^2(1421) = 3039.87, $p < 0.001$, RMSEA = 0.050 (95% CI: 0.047–0.052), CFI = 0.96, TLI = 0.96, SRMR = 0.078, suggesting similar fit as the seven-factor model, but with a simpler model. The *interpersonal mismatch* factor loaded 0.58 on the **destructive** factor, and 0.44 onto the **supportive** factor. Model B yielded identical fit and *df* as in Model A but suggested a 0.91 correlation between the **destructive** and *interpersonal mismatch* factors, and therefore we proceed with the model including cross-loadings for the two higher-order factors as our final model. The higher-order structure of Model A, for which loadings were generally high, is shown in Figure 1. First-order loadings ranged from 0.52 to 0.99, with a mean loading of

[1] VSS is an index that assesses the degree to which the loading pattern reflects simple structure (items have a high loading on a single factor, and near-zero loadings on all other factors); factor solutions with simple structure are preferred; Velicer’s MAP aims to find the solution that minimizes the average residual covariances after systematic factor variance is controlled for; empirical BIC assesses the likelihood of the model given the data controlling for the number of parameters in the model based on the solution’s χ^2 and *df*, and—all else equal—prefers simpler models over more complex ones; the sample sized adjusted BIC is similar but is based on the model log-likelihood rather than the χ^2 and the number of parameters rather than the *df*; in addition, it is adjusted for any differences in sample size (which was not an issue here).

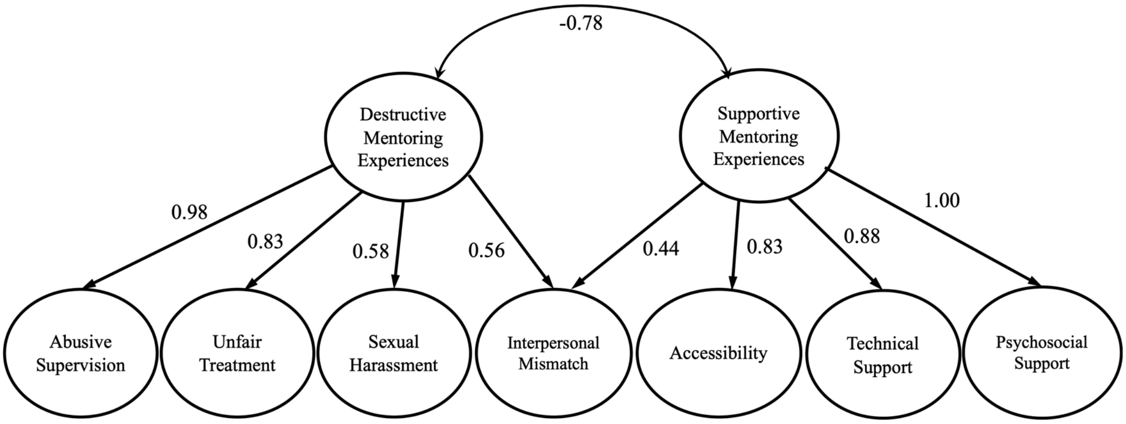

FIGURE 1: Final factor model for the MURS. Ovals represent latent factors and straight lines represent factor loadings. The final model includes seven first-order factors and two second-order factors. The first-order (item level) loadings are not pictured due to space, and can be found in Table 3. The negative correlation between the second order factors reflects the positive valence of supportive mentoring experiences and the negative valence of destructive mentoring experiences.

0.85, *SD* = 0.11 (Table 2). The cross-loading of interpersonal mismatch on both higher-order factors indicates that the MURS is best used either as a total score or using scores at the facet or first-order factor level, rather than as scores for the two higher-order factors of supportive or destructive mentoring experiences.

We conducted item response theory (IRT) analyses to further refine the MURS subscales by removing items that did not contribute to the reliability of the measure. For each subscale, we estimated the graded response model (Samejima, 1968). Global (i.e., scale-level) model-data fit was evaluated using the family of M_2-based goodness of fit statistics (see Maydeu-Olivares and Joe, 2006; Cai and Hansen, 2013; Cai and Monroe, 2014). The M_2 statistic is statistically equivalent to the χ^2 used in structural equation modeling; its properties allow for calculating fit indices such as the RMSEA, CFI, and TLI. At the item-level, model-data fit was assessed using the $S\text{-}\chi^2$ statistic (Orlando and Thissen, 2000), which examines the degree to which observed item responses deviate from expectations across the distribution of the latent variable. These analyses were conducted in the Multidimensional Item Response Theory (MIRT) Package for use in R statistical software program (Chalmers, 2012). Table 3 shows the results of the model-data fit analyses and IRT-based marginal reliability for each scale. These results suggested good model-data fit, which indicates item parameters and the resulting information functions are stable and interpretable.

Given the good fit, we continued to examine each subscale and remove items that provided very low IRT information (Embretson and Reise, 2000). For *abusive supervision*, we found that five of 14 items had very low information across the construct continuum (My mentor gossiped about people in the lab; scolded people in the lab; invaded my privacy; discussed topics that were too personal; and took credit for my work). The *inaccessibility* measure showed none of its five items that required removal. The *technical support* scale showed four of 12 items that had very low information (My mentor explained how my work fit into the bigger picture; was clear about when I was expected to be working; expected me to work reasonable hours; and my mentor and I talked about my career aspirations). Only one item in the *psychosocial support* scale showed low information (My mentor thought the work I did was important). For the *interpersonal mismatch* scale, only one item (My mentor and I had incompatible work styles) showed low information. Of the eight items for *unfair treatment*, one was found to have extremely high misfit (My mentor treated people unfairly based on their career interests) and was removed prior to estimating global model-data fit. The remaining seven item measure showed low information for three items (My mentor treated people unfairly based on their major; was biased against certain groups of people and had favorites in the lab).

Given that only three items were included in the *sexual harassment* measure, we added "My mentor made sexual comments about me" back to the item set to achieve model identification. The fit was very good, but the marginal reliability was very low at 0.16; this is due to the fact that it measures such extreme and rare behavior (e.g., touching without permission, sexual remarks) that it only has high reliability. In other words, at 2 SD above the mean the IRT reliability of the scale is 0.96, but the measure is low in reliability for those near the mean. The second and third items are repetitive and are extremely highly correlated, and thus only one could be used (the item "My mentor made sexual comments about me" caused a loading greater than 1.0 in our initial model for the same reason), as their content is highly similar (i.e., making sexual remarks vs. making sexual remarks about the respondent specifically). The final MURS items and their standardized factor loadings are presented in Table 2.

Personality and Negative Mentoring. We examined correlations between the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) and the MURS factors. Our aim was to ensure that the MURS was not measuring facets of personality that might make individuals more or less likely to report negative experiences with mentors. Openness was the only personality trait that significantly related to negative mentoring, showing a weak negative association ($r = -0.14$, $p = 0.001$). Openness also exhibited significant but small negative associations with most dimensions of

TABLE 2: Standardized first-order factor loadings for final MURS item set

Dimension	Item	Standardized loading (λ)
Abusive supervision		
1	My mentor was rude to me.	0.908
2	My mentor belittled me.	0.885
3	My mentor created an intimidating environment.	0.873
4	My mentor was too harsh with their criticism.	0.88
5	My mentor was passive aggressive.	0.879
6	My mentor made me do excessive grunt work.	0.712
7	My mentor made inappropriate comments about my personal life.	0.852
8	My mentor was condescending.	0.923
9	My mentor blamed me for their mistakes.	0.860
Accessibility (reverse-scored)		
1	My mentor gave me the attention I needed.	0.946
2	My mentor was available when I needed them.	0.884
3	My mentor was around to answer questions.	0.870
4	My mentor made time to meet with me.	0.817
5	My mentor responded when I contacted them.	0.715
Technical support (reverse-scored)		
1	My mentor helped me understand the purpose of research tasks.	0.763
2	My mentor gave me work that was the right level of difficulty for me.	0.702
3	My mentor was clear about how my performance was being evaluated.	0.657
4	My mentor gave me the right amount of work.	0.763
5	My mentor made sure I was prepared to do research tasks.	0.824
6	My mentor gave me enough guidance in my research.	0.840
7	My mentor gave me useful feedback on my work.	0.837
8	My mentor was clear about what they wanted me to do.	0.806
Psychosocial support (reverse-scored)		
1	My mentor was friendly.	0.87
2	My mentor respected me.	0.913
3	My mentor had faith in me.	0.842
4	My mentor valued my contributions to the research.	0.832
5	My mentor cared about me as a person.	0.847
6	My mentor encouraged me.	0.847
Interpersonal mismatch		
1	My mentor and I had a tense relationship.	0.899
2	My mentor and I had incompatible personalities.	0.872
3	My mentor and I worked poorly together.	0.92
4	My mentor and I had difficulty getting along.	0.934
5	My mentor and I had incompatible communication styles.	0.876
Sexual harassment		
1	My mentor touched me without my permission.	0.980
2	My mentor made sexual remarks.	0.917
3	My mentor made sexual jokes.	0.987
Unfair treatment		
1	My mentor treated people unfairly based on their race/ethnicity	0.89
2	My mentor treated people unfairly based on their gender/sex	0.892
3	My mentor treated people unfairly based on their religion	0.934
4	My mentor treated people unfairly based on their sexual orientation	0.958

negative mentoring: *abusive supervision* ($r = -0.12$, $p = 0.006$), *inaccessibility* ($r = -0.12$, $p = 0.008$), *insufficient technical support* ($r = -0.13$, $p = 0.004$), *insufficient psychosocial support* ($r = -0.10$, $p = 0.02$), *interpersonal mismatch* ($r = -0.14$, $p = 0.001$), *sexual harassment* ($r = -0.08$, $p = 0.06$), *unfair treatment* ($r = -0.06$, $p = 0.19$). These results suggested that students' perceptions of their mentoring experiences may be broadly influenced by their level of openness. To account for this, we opted to measure openness in the next phase of data collection so that we could ensure that it did not influence the outcomes of interest and confound our ability to estimate the impact of negative mentoring experiences. We also chose to measure neuroticism

TABLE 3: IRT model-data fit and reliability statistics[a]

					90% CI						Misfit
Dimension	M_2	*df*	*p*	RMSEA	Low	High	TLI	CFI	SRMR	$\rho_{xx'}$	rate
Abusive supervision	96.33	35	<0.001	0.059	0.045	0.074	0.95	0.96	0.066	0.86	0/14
Inaccessibility	32.05	5	<0.001	0.102	0.070	0.137	0.97	0.98	0.042	0.87	0/5
Insufficient technical support	232.69	54	<0.001	0.080	0.069	0.091	0.96	0.97	0.054	0.91	0/12
Insufficient psychosocial support	229.47	14	<0.001	0.174	0.154	0.194	0.94	0.92	0.08	0.87	0/7
Interpersonal mismatch	68.26	9	<0.001	0.113	0.088	0.139	0.97	0.98	0.064	0.85	0/6
Sexual harassment	5.73	2	0.057	0.060	0.000	0.121	0.98	0.99	0.083	0.16*	n/a
Unfair treatment	120.72	14	<0.001	0.123	0.103	0.144	0.96	0.97	0.059	0.72	1/8+

[a] Model-data fit for each dimension has to be done one dimension at a time for IRT. M_2 is the theoretical equivalent of χ^2 for IRT global fit; it is a χ^2 type statistic for nominal data, as opposed to continuous data. Rho xx' is IRT-based marginal reliability. Misfit rate is the number of items showing significant misfit out of the total number of items retained from the CFA. RMSEA = root mean square error of approximation; TLI = Tucker–Lewis index; CFI = comparative fit index; SRMR = standardized root mean residual; ρxx' = IRT-based marginal reliability; Misfit rate = The number of items with significant misfit according to the s-x2 item fit statistic out of the total number of items. + Indicates the misfitting item was removed prior to estimating global model-data fit. *Although the overall marginal reliability of the sexual harassment scale was low, its reliability at high levels of the construct was acceptable. Global fit and marginal reliability were estimated prior to items being deleted unless noted otherwise.

to ensure the replicability of the lack of association with negative mentoring experiences.

External Phase

In our final phase of data collection and analysis, we aimed to characterize relationships between responses on the MURS with variables we hypothesized would relate to negative mentoring experiences, or its nomological network (Cronbach and Meehl, 1955). In other words, we sought to interpret the meaning of the MURS scores in relation to theoretically-relevant constructs and outcomes (Benson, 1998). To accomplish this, we collected evidence to test whether scores on the MURS correlated as expected with measures of related constructs, did not correlate with measures of unrelated constructs, and were predictive of theoretically-related and practically-relevant outcomes. For ease of reading, we present the methods for data collection and analysis first. Then, we describe our hypotheses of how responses on the MURS relate to student outcomes, covariates, and other measures of mentoring quality along with our results characterizing these relationships. We continue to present methods and results using a common, negative valence for ease of reading and comparison.

Data Collection

To carry out the external phase, we collected and analyzed a second national dataset. We recruited undergraduates at 32 institutions who were about to do research for the *first time* to avoid selection bias in the sample (i.e., students staying or leaving research experiences because of their mentoring experiences). We did not include any selective programs (i.e., programs that had an application process or selected students based on academic standing, such as honors programs and REU programs) to mitigate bias in our sample. We used a presurvey/postsurvey design to evaluate how students' negative mentoring experiences related to changes they may or may not realize from participating in undergraduate research.

Prior to the start of their research experience, we emailed participants a presurvey with measures of our constructs of interest as well as items to measure student demographics (see full item set with references and description of validity evidence in Supplemental Materials). At the end of one term of research (quarter, semester, summer), we emailed them the postsurvey, which included the MURS items along with measures of our constructs of interest (see full item set with references and description of validity evidence in Supplemental Materials). Students were compensated $25 total for their participation: $10 for the presurvey, $15 for the postsurvey. We received 359 responses to both the presurvey and postsurvey; 11 of which did not pass all attention checks. Thus, the final sample size was *n* = 348 (Table 1).

Scoring Mentoring Experiences

Given the results of the internal phase analysis, we scored students' mentoring experiences for the MURS in its entirety and at the dimension level by calculating means. Thus, the loadings of individual items on first-order factors and cross-loading of interpersonal mismatch did not influence score calculations. Related to this, it is good measurement practice to evaluate the fit of a measurement model after any modifications and with each independent dataset. We attempted to fit a CFA with our second national dataset to further assess our final factor structure. However, the model estimation resulted in both a not positive-definite covariance matrix and estimated negative covariances. We attempted to determine the source of these issues and suspect that multicollinearity among the dimensions of MURS may be the issue. We attempted an alternative approach of fitting the data using a multigroup model with both the first and second dataset, and then cross validating the measurement model by assessing measurement invariance. The model ran but the imbalanced sample sizes made this approach unfeasible. Future research using the MURS should continue to assess model fit.

Base Rate of Mentoring Experiences

We plotted histograms (Figure 2) and calculated means and SDs (Table 4) to gain insight into prevalence of mentoring experiences. Undergraduates reported the highest absence of *technical support* (*M* = 1.55, SD = 0.72) compared with all other forms of negative mentoring, although its base rate was still low. Undergraduates reported lower levels of *abusive supervision* (*M* = 1.23, SD = 0.49), *inaccessibility* (*M* = 1.32, SD = 0.61),

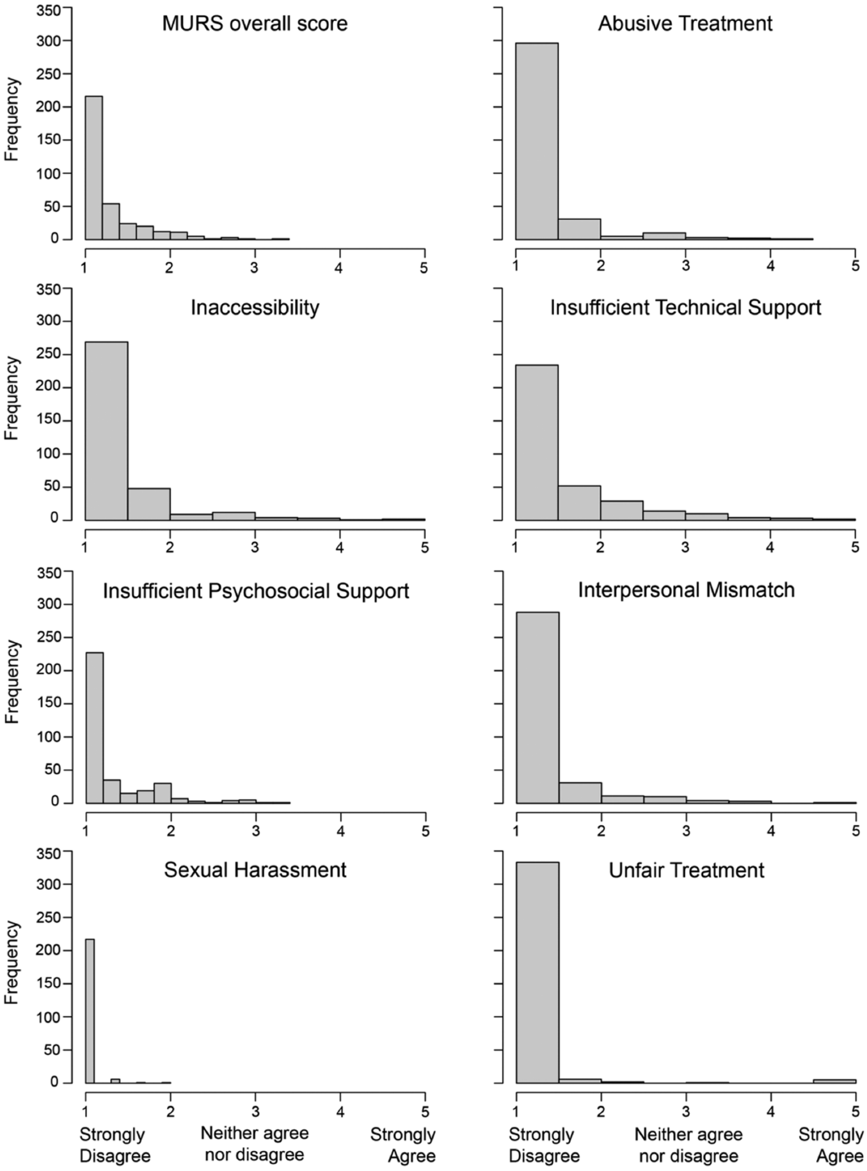

FIGURE 2: Histograms of the MURS and its seven dimensions. For comprehensibility and ease of comparison, histograms reflect our use of a common, negative valence such that supportive experience results are reverse-scored and presented in terms of their absence (inaccessibility, insufficient technical support, insufficient psychosocial support) and destructive experiences are presented in terms of their presence. The overall MURS score reflects negative mentoring experiences.

insufficient psychosocial support ($M = 1.30$, SD = 0.45), *interpersonal mismatch* ($M = 1.27$, SD = 0.57) and the lowest levels of *sexual harassment* ($M = 1.02$, SD = 0.10) and *unfair treatment* ($M = 1.09$, SD = 0.51), indicating that these forms of negative mentoring were quite uncommon in our sample.

We looked for differences in students' reports of mentoring experiences based on their personal characteristics using Wilcoxon rank-sum tests (i.e., nonparametric *t* tests) and Kruskal–Wallis tests (i.e., nonparametric ANOVAs). We found no differences in any dimensions of mentoring by race/ethnicity when comparing the experiences of students who identified as Asian, White, or from a minoritized race or ethnicity[2]. We report base rates of negative mentoring experiences for each racial/ethnic group in Table 5. It is theoretically possible for base rates to be similar across racial/ethnic groups, and for effects of negative mentoring experiences to differ among groups. To test this, we estimated a set of regression models adding an interaction term between race/ethnicity and MURS predicting each outcome (outcomes are described in detail below). The interaction terms were not significant in all cases. We also found no differences by generation in college, or by mentor rank (faculty or not faculty). We only observed one difference by student gender: men reported more *abusive supervision* than women (men: $n = 89$, $M = 1.35$, SD 0.63; women: $n = 254$, $M = 1.19$, SD = 0.43, $W = 13{,}392$, $p = 0.003$).

Discriminant and Convergent Evidence

We evaluated whether scores on the MURS were associated as expected (or not) to theoretically related constructs by making a priori predictions about how mentoring experiences would relate to students' personality traits, attachment styles, emotions about research, and other measures of mentoring. We then evaluated these relations by examining bivariate correlations (Table 4). We use a value of $p < 0.05$ to determine significance and we report all p values for readers to make their own assessments given the exploratory nature of the work.

Personality Traits. We hypothesized that mentoring experiences would be unrelated to any personality traits. Based on results from the structural phase analysis, we sought to rule out the hypothesis that an undergraduate's reports of negative mentoring experiences were due to their level of openness. Prior research indicates that individuals high on openness are likely to judge experiences as less negative and more likely to respond to abusive supervision with coping strategies that mitigate the emotional labor associated with such experiences (Steel *et al.*, 2008;

[2] Although we make use of the broad category of "Asian," we recognize that students who identify as Asian have a spectrum of experiences and more careful disaggregation by specific cultural or national identity is needed to understand these experiences. We make use of the broad category of "minoritized" to include students who identify as American Indian/Alaskan Native, African American or Black, Native Hawaiian/Pacific Islander, and Hispanic/Latine. Again, we recognize there are important differences between these groups and students have a range of experiences within and across racial and ethnic groups. Our intention with using these broad categories is explore whether there are any patterns shared across these groups.

TABLE 4: Descriptive statistics and correlations for external phase data[a]

Construct and variable		*N*	*M*	SD	Abusive supervision	Inaccessibility	Insufficient technical support	Insufficient psychosocial support	Interpersonal mismatch	Sexual harassment	Unfair treatment	MURS overall
Negative mentoring	Abusive supervision	348	1.23	0.49								
	Inaccessibility	348	1.32	0.61	**0.44*****							
	Insufficient technical support	348	1.55	0.72	**0.39*****	**0.70*****						
	Insufficient psychosocial support	348	1.30	0.45	**0.58*****	**0.60*****	**0.67*****					
	Interpersonal mismatch	348	1.27	0.57	**0.72*****	**0.53*****	**0.54*****	**0.64*****				
	Sexual harassment	225	1.02	0.10	0.08	–0.01	**0.14***	0.06	0.03			
	Unfair treatment	347	1.09	0.51	0.07	0.03	0.03	0.04	0.10	0.08		
Personality	Neuroticism	348	3.34	0.96	–0.01	–0.05	–0.04	–0.03	0.01	–0.03	–0.01	–0.03
	Openness	348	3.85	0.87	–0.04	–0.02	0.00	–0.06	–0.01	0.13	0.00	–0.03
Attachment style	Avoidant	348	2.52	0.89	0.02	0.06	0.10	0.09	0.09	0.07	0.06	0.10
	Anxious	348	2.72	0.90	0.06	**0.16****	**0.17****	**0.12***	**0.16****	**0.16***	–0.06	**0.15****
	Secure	348	4.28	0.68	–0.09	–0.06	–0.08	**–0.12***	–0.08	0.11	0.00	–0.10
Emotions	Positive affect toward research	341	4.02	0.88	**–0.17****	**–0.31*****	**–0.46*****	**–0.42*****	**–0.28*****	–0.12	–0.02	**–0.39****
	Negative affect toward research	341	2.12	0.70	**0.43*****	**0.24*****	**0.33*****	**0.30*****	**0.45*****	0.05	0.08	**0.42*****

Construct and variable		*N*	*M*	SD	Abusive supervision	Inaccessibility	Insufficient technical support	Insufficient psychosocial support	Interpersonal mismatch	Sexual harassment	Unfair treatment	MURS overall	Mentoring competence
Mentoring competence		347	3.64	0.48	**–0.46*****	**–0.62*****	**–0.79*****	**–0.70*****	**–0.56*****	–0.10	–0.08	**–0.75*****	
Mentoring relationship quality		340	4.47	0.80	**–0.45*****	**–0.73*****	**–0.77*****	**–0.72*****	**–0.59*****	–0.06	0.00	**–0.76*****	**0.78*****
Scientific self-efficacy	Pre	348	3.50	0.68	0.05	–0.08	**–0.11***	–0.10	–0.07	0.01	**–0.14***	–0.10	
	Post	340	4.12	0.64	**–0.13***	**–0.31*****	**–0.39****	**–0.37*****	**–0.23*****	–0.02	–0.06	**–0.34*****	
Scientific identity	Pre	348	4.06	0.67	–0.03	**–0.16****	**–0.23*****	**–0.21*****	**–0.12***	0.01	–0.05	**–0.19*****	
	Post	340	4.24	0.76	**–0.12***	**–0.27****	**–0.38*****	**–0.38*****	**–0.21*****	–0.05	0	**–0.32*****	
Research beliefs	Pre	348	3.98	0.48	**–0.18*****	–0.08	**–0.22*****	**–0.18*****	**–0.23*****	–0.04	–0.04	**–0.23*****	
	Post	341	3.98	0.59	**–0.25*****	**–0.22*****	**–0.33*****	**–0.27*****	**–0.29*****	–0.07	–0.09	**–0.34*****	
Intentions	Pre	347	4.19	0.68	**–0.15****	–0.08	**–0.18*****	**–0.16****	**–0.14****	–0.09	0.03	**–0.16****	
	Post	341	4.15	0.84	**–0.15****	**–0.19*****	**–0.23*****	**–0.17****	**–0.16****	–0.09	–0.03	**–0.22*****	

[a] Relationships that were predicted a priori are indicated with the following emphases: shading indicate a predicted negative relationship and a double border || indicates a predicted positive relationship. Significant relationships are **bolded** with *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

TABLE 5: Base rate of mentoring experiences by racial/ethnic groups

	Abusive treatment		Inaccessibility		Insufficient psychosocial support		Insufficient technical support		Interpersonal mismatch		Sexual harassment		Unfair treatment	
Race/Ethnicity	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Asian ($n = 99$)	1.26	0.5	1.34	0.63	1.29	0.39	1.50	0.71	1.28	0.54	1.01	0.04	1.15	0.7
Minoritized Race/Ethnicity ($n = 97$)	1.25	0.44	1.32	0.6	1.30	0.47	1.52	0.73	1.30	0.52	1.02	0.08	1.14	0.13
White ($n = 146$)	1.19	0.55	1.32	0.6	1.30	0.48	1.61	0.71	1.24	0.66	1.02	0.08	1.02	0.63

Wu and Hu, 2013). Given that the personality trait of neuroticism reflects a tendency toward negative feelings, we also sought to rule out the hypothesis that an undergraduate's reports of negative mentoring experiences were due to their level of neuroticism. As predicted, all correlations between the seven negative mentoring dimensions and both *neuroticism* and *openness* were near-zero and nonsignificant (Table 4). These results suggest that the MURS is unlikely to be measuring personality traits per se.

Attachment Styles. Attachment styles are stable patterns of emotions and behaviors exhibited in close relationships, which are thought to develop through early interactions between infants and their caregivers (Bowlby, 1979; Ainsworth, 1989; Bowlby and Ainsworth, 2013). Researchers have described two main forms of attachment: *secure,* in which the infant perceives their caregiver as a source of comfort and strength, and insecure or anxious attachment. Forms of insecure attachment include *anxious* attachment, in which the infant perceives their caregiver as an unreliable—sometimes offering support and other times not, and *avoidant* attachment, in which the infant has learned the caregiver is not a reliable source of support and thus does not expect or seek comfort from them (Carver, 1997). These early experiences are thought to shape an individual's internal working model of relationships and thus influence how adults think, feel, and behave in close relationships (Hazan and Shaver, 1994), including supervisory relationships (Fitch *et al.*, 2010). Thus, we explored whether and how undergraduate researchers' attachment styles related to their negative mentoring experiences.

We hypothesized that undergraduates' mentoring experiences would relate to their attachment styles, but that the magnitude of the correlations would be small to moderate such that mentoring experiences are not redundant with attachment style. We focused on avoidant and anxious attachment styles because we hypothesized that these attachment styles would influence an undergraduate researcher's expectations for their relationships with their mentors. Specifically, we predicted that *avoidant* attachment style would negatively relate to both *inaccessibility* and *insufficient psychosocial support* because individuals who are avoidant would expect less attention and support from their mentors having learned to not expect such support from their caregivers. Thus, they would be less likely to report dissatisfaction when their mentor was unavailable to them or did not provide psychosocial support. However, we found that undergraduates' levels of *avoidant* attachment were not associated with their ratings of mentor *inaccessibility* ($r = 0.06$, $p = 0.29$) or *insufficient psychosocial support* ($r = 0.09$, $p = 0.093$) (Table 4).

Research indicates that anxious attachment includes an individual's fear of abandonment in relationships as well as the tendency to want to have a closer relationship than their relational partner (Carver, 1997). We hypothesized that undergraduates' levels of anxious attachment would positively relate to *inaccessibility* and *insufficient psychosocial support* because individuals with anxious attachment styles desire a higher level of attention and support and thus may be more distressed by these forms of negative mentoring. Surprisingly, undergraduates who indicated an *anxious* attachment style reported slightly higher levels of most forms of negative mentoring, except *abusive supervision* and *unfair treatment* (Table 4). We did not have a priori hypotheses about secure attachment style and negative mentoring experiences. Yet, we observed a small but significant relationship between undergraduates reporting a secure attachment style and lower levels of *insufficient psychosocial support* ($r = -0.12$, $p < 0.05$) (Table 4). Altogether, these results indicate that undergraduate researchers who have an *anxious* attachment style may be slightly more susceptible to negative mentoring experiences. Furthermore, undergraduates with a secure attachment style might perceive more psychosocial support or require less psychosocial support to thrive. Collectively, however, these effects were modest (r values from $|0.10|$ to $|0.17|$; Table 4), which indicates that the MURS is unlikely to simply be measuring attachment styles.

Emotions about Research. Emotions are responses, including feelings, actions, and physiological changes, to situations that garner an individual's attention (Gross and Thompson, 2007). Appraisal theory indicates that emotions arise when an individual positively or negatively appraises a situation that is personally significant to them (Scherer, 1999). Prior research shows that students' emotions can have substantial effects on their academic engagement and performance (Pekrun and Linnenbrink-Garcia, 2012). In addition, negative behaviors in the workplace are associated with employees experiencing toxic emotions and emotional exhaustion (e.g., Porath and Pearson, 2012; Henle and Gross, 2014; Han *et al.*, 2017). Thus, we hypothesized that negative mentoring experiences would impact whether students have positive or negative emotions about their research experience. Specifically, we hypothesized that students' positive emotions about research (e.g., excitement, accomplishment) would be negatively related to experiencing *insufficient technical support* and *insufficient psychosocial support* because we postulated that students who experience these forms of support are more likely to feel positively about themselves and their work. Indeed, undergraduates who reported higher levels of *insufficient technical support* and *insufficient psychosocial support*, as well as all other forms of negative mentoring except *sexual harassment* and *unfair treatment*, reported significantly lower levels of *positive emotions* (r values from −0.17 to −0.46; Table 4). We also hypothesized that

TABLE 6: Regression analysis results

Model and predictors		Scientific self-efficacy postscore		Scientific identity postscore		Research beliefs postscore		Intentions postscore	
		β	Variance explained	β	Variance explained	β	Variance explained	β	Variance explained
Baseline	Prescore	0.39***	$R^2 = 0.15$	0.58***	$R^2 = 0.33$	0.52***	$R^2 = 0.27$	0.67***	$R^2 = 0.45$
MURS	Prescore	0.36***	$R^2 = 0.25$	0.54***	$R^2 = 0.38$	0.47***	$R^2 = 0.32$	0.65***	$R^2 = 0.46$
	MURS	−0.31***		−0.22***		−0.23***		−0.13**	
MCA	Prescore	0.36***	$R^2 = 0.30$	0.51***	$R^2 = 0.43$	0.48***	$R^2 = 0.33$	0.65***	$R^2 = 0.47$
	MCA	0.38***		0.31***		0.24***		0.14***	
MRQ	Prescore	0.36***	$R^2 = 0.27$	0.51***	$R^2 = 0.45$	0.49***	$R^2 = 0.31$	0.65***	$R^2 = 0.46$
	MRQ	0.35***		0.34***		0.20***		0.13**	

β = standardized estimate. **, $p < 0.01$; ***, $p < 0.001$.

students' *negative emotions* about research (e.g., stress, apathy) would be positively related to all forms of negative mentoring because all of these experiences are likely to generate mentee distress. As expected, undergraduates' *negative emotions* about research were significantly correlated with all forms of negative mentoring experiences except *sexual harassment* and *unfair treatment* (r values ranged from 0.24 to 0.45; Table 4).

Other Measures of Mentoring. If the MURS is measuring the range of mentoring undergraduates experience, responses on the MURS should relate to the perceived quality of their mentoring relationships (Allen and Eby, 2003). Responses on the MURS should also relate to measures of perceived mentoring competency, including a mentor's abilities to communicate effectively with their mentee, align their expectations with those of their mentee, and foster their mentee's independence (Fleming *et al.*, 2013). Specifically, we predicted that:

- *Mentoring relationship quality* will negatively relate to the overall MURS score and to *all seven dimensions* of negative mentoring because negative mentoring experiences should undermine the development and maintenance of a quality relationship.
- *Mentoring competence* (MCA) will negatively relate to the overall MURS score and to *all seven dimensions* of negative mentoring because MCA is needed to prevent negative mentoring experiences.

Undergraduates who reported lower levels of mentoring relationship quality reported significantly higher levels of negative mentoring experience overall ($r = –0.75$, $p < 0.001$) and of all dimensions of negative mentoring except *sexual harassment* and *unfair treatment* (r values ranged from –0.46 to –0.79; Table 4). Undergraduates who reported lower levels of mentor competence also reported significantly higher levels of negative mentoring overall ($r = –0.76$, $p < 0.001$) and of all dimensions of negative mentoring, except *sexual harassment* and *unfair treatment* (r values from –0.45 to –0.77; Table 4). Collectively, these results indicate that the MURS is measuring aspects of mentoring relationships that relate to mentoring quality and mentor competence, without being completely redundant with these measures.

Predictive Evidence

Finally, we examined how mentoring experiences measured by MURS related to outcomes undergraduates typically experience from participating in research. Research experiences are widely accepted as formative experiences in which undergraduates grow in their belief that they can be successful in science (i.e., scientific self-efficacy) and their view of themselves as a "science person" (i.e., scientific identity) (Kardash, 2000; Hunter *et al.*, 2007; Estrada *et al.*, 2011; Robnett *et al.*, 2015; Gentile *et al.*, 2017). Furthermore, expectancy value theory postulates that one is motivated to engage in a task, such as pursuing a science research career, if one believes they can be successful (i.e., science self-efficacy) and that the task has value (e.g., the benefits of doing science research outweigh the costs) (Wigfield and Eccles, 2000; Barron and Hulleman, 2015). Based on this research and theory, we formulated a series of hypotheses regarding how experiencing negative mentoring would relate to undergraduate researchers' development of *scientific self-efficacy* and *scientific identity* as well as their beliefs about the value of research (*research beliefs*) and their intentions to continue in science and in research (*intentions*). We evaluated these relationships by fitting a series of linear regression models using mean scores for relevant scales, as in this example model: Outcome_t2 ~ Outcome_t1 + MURS We sought to determine whether MURS explained variance in undergraduates' postresearch self-efficacy, identity, beliefs, and intentions above and beyond their preresearch ratings (Table 6).

Scientific Self-Efficacy. Research indicates that social persuasion, meaning encouraging feedback from influential individuals, such as mentors, functions as a source of self-efficacy (Usher and Pajares, 2008). Undergraduate researchers may experience less development of their *scientific self-efficacy* if they do not experience social persuasion because their mentors are inaccessible or are not providing psychosocial support. Mastery experiences, or tackling and ultimately succeeding at a challenging task, function as another critical source of self-efficacy development (Usher and Pajares, 2008). Undergraduate researchers may have fewer mastery experiences if they receive insufficient technical support to be successful, their tasks are not at the right level of challenge, or they perceive themselves to be unsuccessful. Thus, we hypothesized that students' development of *scientific self-efficacy* during research would be limited by experiencing negative mentoring. As expected, undergraduates who reported experiencing more negative mentoring also reported significantly lower postresearch self-efficacy after accounting for their preresearch self-efficacy ($\beta = –0.31$, $p < 0.001$; Table 6).

Scientific identity. Typically, undergraduate research experiences positively influence students' *scientific identity*, making them feel like more of a "science person" (Estrada *et al.*, 2011; Robnett *et al.*, 2015; Gentile *et al.*, 2017). Identity development or lack thereof, is influenced by recognition from members of the community, such as mentors (Carlone and Johnson, 2007; Hazari *et al.*, 2010). Thus, we hypothesized that students' development of a *scientific identity* during research would be limited by experiencing negative mentoring. Indeed, undergraduates who reported experiencing more negative mentoring also reported lower postresearch scientific identity after controlling for preresearch identity ($\beta = -0.22$, $p < 0.001$; Table 6).

Research Beliefs. We hypothesized that undergraduates who experience more negative mentoring would perceive research as less beneficial and more costly (Gaspard, Dicke, Flunger, Brisson *et al.*, 2015; Gaspard, Dicke, Flunger, Schreier *et al.*, 2015; Ceyhan and Tillotson, 2020). We focused on measuring students' beliefs about the *intrinsic value* of research (i.e., how interesting or enjoyable research is), the *communal value* of research (i.e., potential for research to benefit a broader community or society), and the *opportunity costs* of research (i.e., sacrifices students perceive they would have to make to engage in research) (Barron and Hulleman, 2015; Brown *et al.*, 2015). We hypothesized that, collectively, students' postresearch beliefs would be hampered by negative mentoring experiences (increased perceptions of opportunity costs and decreased perceptions of intrinsic and communal value). As expected, undergraduates who reported experiencing more negative mentoring also reported lower postresearch *beliefs* after controlling for preresearch beliefs ($\beta = -0.23$, $p < 0.001$; Table 6).

Graduate and Career Intentions. Undergraduates who participate in research clarify their career choices and, as a result, can change their *intentions* to pursue graduate education and careers in science (Estrada *et al.*, 2011, 2018; Gentile *et al.*, 2017). We hypothesized that students' *intentions* would negatively relate to experiencing more negative mentoring because, if students do not receive sufficient support, perceive a mismatch with more experienced researchers, or are treated poorly or unfairly, they are more likely to opt out of science or research paths. Indeed, undergraduates who reported experiencing more negative mentoring also reported lower postresearch intentions after controlling for their preresearch intentions, although the effect was more modest than observed for other outcomes ($\beta = -0.13$, $p < 0.01$; Table 6).

We next examined correlations between dimensions of the MURS and premeasures and postmeasures of each outcome. As expected, all of the dimensions of MURS were negatively related to students' postresearch ratings of their self-efficacy, identity, research beliefs, and intentions, except for sexual harassment and unfair treatment. Surprisingly, students' preresearch ratings of their identity, beliefs, and intentions were also negatively related to responses on the MURS, although these relationships were more modest (r values from –0.12 to –0.23). It may be that students who identify less as a scientist, who hold more skeptical beliefs about the value of doing research, or who do not have strong intentions to continue in science or research have greater mentoring needs and thus report less favorable mentoring experiences. Alternatively, mentors may be consciously or unconsciously sensing that their mentees are less integrated into the scientific community and proffering less favorable mentoring.

To compare the explanatory values of the MURS versus measures of mentoring relationship quality and MCA, we fit a similar series of linear regression models using the mentoring relationship quality (MRQ) scale or the MCA, as in this example: Outcome _t2 ~ Outcome_t1 + MRQ. The standardized estimates for the MURS, MRQ, and MCA were quite similar for all of the outcomes we examined (Table 6). In addition, the variance in outcomes explained by mentoring was similar, regardless of whether negative mentoring (MURS), MRQ, or MCA was the focus of analysis. In other words, all three mentoring measures explained variance in postresearch self-efficacy, identity, beliefs, and intentions beyond preresearch ratings, but the variance explained was similar.

DISCUSSION

Results from the MURS indicate that most undergraduate researchers experience high levels of supportive forms of mentoring and low levels of destructive forms of mentoring. Yet, undergraduates in both of our samples reported experiencing the absence of supportive mentoring as well as destructive mentoring, and these negative experiences were associated with less favorable outcomes of participating in research. Substantial time and resources are invested in undergraduate STEM research experiences (Eagan *et al.*, 2013; Gentile *et al.*, 2017), but negative experiences with mentors appear to be limiting students' growth. Additional action is needed, both locally and nationally, to incentivize, support, and reward quality undergraduate research mentoring and more fully realize the benefits of these investments.

Given that *insufficient technical support* was the most prevalent form of negative mentoring experienced by undergraduates in our study, it is surprising that we found no association with mentor position type (faculty vs. non-faculty). This suggests that *insufficient technical support* may be unrelated to mentor expertise and experience. Instead, undergraduates may hesitate to ask for guidance if they feel like they should already know or be able to do aspects of their research, and thus miss opportunities to elicit sufficient technical support. Alternatively, undergraduates may feel like their mentors provide technical support that is not sufficiently aligned with their current understanding to help them learn and make research progress. Another possibility is that research is a particularly unstructured environment compared with other learning experiences, such as coursework. Undergraduates may use courses as reference point when evaluating whether URE expectations and mentor feedback are clear, finding UREs and research mentors lacking in comparison. Future research could identify which, if any, of these factors are driving undergraduates' ratings of technical support, ultimately informing the design and selection of suitable interventions.

Several components of the mentoring professional development curriculum, *Entering Mentoring,* are designed to help mentors develop strategies for providing sufficient technical support (Pfund *et al.*, 2015). For example, lessons on communicating effectively are designed to help mentors set a tone that asking questions is normal and expected. Lessons on assessing understanding are also designed to help mentors learn to gauge mentee knowledge and skills, for instance by asking mentees to explain aloud or in writing any techniques, protocols, or concepts they are learning. Presumably, mentors who are successful in establishing a culture where asking questions is encouraged and who are skilled in assessing mentee understanding are likely to have better insights into what mentees know and can do, and can provide support accordingly. Future research could test the effectiveness of these lessons in particular and *Entering Mentoring* in general for increasing undergraduates' perceptions of sufficient technical support.

It is noteworthy that *inaccessibility* was highly correlated with *insufficient technical support* and to a lesser extent with *insufficient psychosocial support*. Although not surprising, these results indicate that mentors must be available to mentees to provide sufficient support. In addition, *inaccessibility*, *insufficient psychosocial support*, and *insufficient technical support* had the largest negative effect on students' postresearch scientific self-efficacy, which is a well-documented outcome of UREs and an important predictor of continuing in a science research-related career path (Estrada *et al.*, 2011; Adedokun *et al.*, 2013; Robnett *et al.*, 2015; Frantz *et al.*, 2017; Hess *et al.*, 2023). These findings raise questions of whether traditional dyadic mentoring is the most effective mentoring structure for undergraduates who are likely to need more support than mentees who are further along in their education and development. Prior research indicates that a triadic model of mentoring, where undergraduate researchers are mentored by both a graduate student or postdoctoral associate and a faculty member, offers benefits over dyadic structures (Aikens *et al.*, 2016; Joshi *et al.*, 2019). Team and peer mentoring structures can also be advantageous because they enable peers to help one another and offer additional role models from whom mentees can learn (Gentile *et al.*, 2017; Sonnenberg-Klein *et al.*, 2017). Future research could compare the effects of different mentoring structures on undergraduates' perceptions of mentor accessibility and sufficient technical and psychosocial support.

Longitudinal research using the MURS should be useful for examining mentoring in undergraduate research as a recursive process. For example, our results indicated that *interpersonal mismatch* was related to absence of supportive mentoring and to destructive mentoring experiences. It may be that negative mentoring experiences reflect a developmental process in which mentees feel well-matched with their mentor if they experience supportive mentoring, fair treatment, and effective supervision, and mismatched if they experience abusive supervision, unfair treatment, or insufficient support. Our results also revealed negative relationships between *insufficient technical and psychosocial support* and undergraduates' positive emotions about research, as well as positive relationships between *abusive supervision* and undergraduates' negative emotions about research. Longitudinal studies could yield insight into whether certain forms of mentoring foster positive or negative emotions, which in turn prompt undergraduates to continue in or exit from research paths. Based on the evidence presented here, the MURS could be used in its entirety to measure mentoring experiences collectively or by dimension to gain mechanistic insights about the influence of specific types of mentoring experiences on undergraduate researchers' career motivations and decisions.

Research on mentoring highlights the importance of shared beliefs, values, and interests between mentors and mentees (Turban and Jones, 1988). This "psychological similarity" is associated with higher levels of psychosocial support, relationship quality, and commitment to STEM careers (Hernandez *et al.*, 2017; Pedersen *et al.*, 2022). It may be that interpersonal mismatch and psychological similarity are two ends of the same continuum (i.e., one construct) or two distinct constructs. For instance, mentees may feel they are similar to their mentors (or not), without feeling mismatched, or they may perceive the absence of similarity as an indicator of mismatch. Given the positive effects of psychological similarity reported elsewhere and the negative effects of mismatch observed here, future research should examine how these constructs relate as well as how they function to influence undergraduate researchers' professional growth and career pursuits. For instance, research has shown that a "birds of a feather" intervention (Gehlbach *et al.*, 2016; Robinson *et al.*, 2019), which highlights a dyad's shared interests, can promote psychological similarity and relationship quality between undergraduate STEM mentees and their mentors (Hernandez *et al.*, 2020, 2023). Such an intervention may set undergraduate researchers on a path toward developing quality relationships with research mentors and buffer against the perception of interpersonal mismatch.

Finally, the MURS—either in its entirety or with a focus on specific dimensions—could be used by programs and institutions to monitor the quality of undergraduate research mentoring and evaluate the effectiveness of efforts to improve undergraduate research mentorship over time. Such efforts should be designed and implemented in ways that protect student confidentiality and avoid making judgments about individual mentors based on limited student responses. Instead, the focus should be on using the results to make informed, structural changes that enable mentors and mentees to work together effectively and that maximize the benefits and minimize the costs of undergraduate research. The MURS could be administered by programs or units to identify which dimensions are rated least positively by students. If MURS results reveal that mentor *inaccessibility* is an issue, the unit could discuss reasonable expectations for mentor accessibility and strategies for feasibly meeting these expectations. Options could include offering more course-based research options to reduce demand for undergraduate research internships (Dolan and Weaver, 2021), pilot testing team-based approaches to undergraduate research (Strachan *et al.*, 2019), or rethinking unit-wide faculty workload allocation such that undergraduate research mentoring is sufficiently incentivized, supported, and rewarded (O'Meara *et al.*, 2018). If MURS results reveal that students are not receiving sufficient technical support (i.e., clear expectations, sufficient preparation, and feedback), the unit could provide a template for mentors and mentees to communicate about expectations (Pfund *et al.*, 2015) or offer skill-building sessions to provide additional preparation undergraduate researchers (e.g., how to read and apply scientific literature, how to write a research paper, how to prepare and present a research poser) (Branchaw *et al.*, 2020; Dolan and Weaver, 2021). Future research could assess the effects of these or other interventions aimed at fostering positive mentoring relationships (Lee *et al.*, 2015; Pfund *et al.*, 2015).

Limitations

Given the potential for measurement tools to shape future research, we took several steps to maximize the quality and utility of MURS as a measure of mentoring experiences. We collected data from a diverse group of undergraduate researchers at a variety of types of institutions who varied in their gender, racial, and other identities, which bolsters the potential applicability of our results to diverse student groups. Yet, we were unable to collect sufficient responses during the external phase to further assess our final factor structure. As noted above, it is plausible that multicollinearity among the MURS dimensions in the external phase sample may be the issue. The dimensions may be less related in the internal phase dataset because we intentionally sampled across the continuum of mentoring experience, while the external phase dataset did not. In addition, some research indicates that using the DWLS estimator has the potential to inflate fit indices (Xia and Yang, 2018; Shi and Maydeu-Olivares, 2020), which we used to assess the goodness of fit of our measurement model. Thus, future research should continue to assess the internal structure of the MURS, ideally using SEM and accounting for the influence of ordered categorical data on fit indices.

We were also unable to collect sufficient responses to allow for examination of the experiences of particular groups of students (e.g., individuals identifying as Native American, Black, nonbinary). Given that prior research has shown that access to and quality of mentorship varies across sociodemographic groups (reviewed in Byars-Winston and Dahlberg [2019] and National Academies of Sciences [2018]), this limitation may be the reason our results show no differences in mentoring experiences among students based on their sociodemographics (i.e., race/ethnicity, gender, first-generation college status). Future research should examine whether students who identify with particular groups respond differently to the MURS, experience negative mentoring at different rates, or are differentially affected by it.

It is important to note that our study lacks evidence related to the consequences of testing (American Educational Research Association *et al.*, 2014). We urge caution in using scores on the MURS to pass judgments on individual faculty members as "negative mentors" or to determine who can and cannot mentor undergraduate researchers. Mentorship is inherently dyadic and embedded in a context, such as a program or degree plan that exerts additional influence on the mentoring relationship. Mentees themselves and other contextual factors could contribute to dysfunction in mentoring relationships (Eby and McManus, 2004). Furthermore, mentees differ in their goals, interests, experiences, and aspirations and thus require different investments of time, training, and support from mentors. Thus, a mentor who may be a poor fit with one mentee may be an excellent fit with another. Finally, mentees themselves may be biased against particular mentors based on their identities, as has been observed in student end-of-course evaluations of instruction (MacNell *et al.*, 2015; Goos and Salomons, 2017; Fan *et al.*, 2019; Esarey and Valdes, 2020). Future research needs to examine and study potential unintended negative consequences of the MURS for mentees, mentors, and programs.

Our study had limitations beyond those associated with measurement development. First, we conducted multiple tests, which may have resulted in false positives. The relationships reported here should continue to be investigated in future research. Second, we failed to find the relationships between the sexual harassment and unfair treatment dimensions of MURS and almost all other constructs we hypothesized to be related. These negative results could be due to insufficient measurement of this dimension, or due to sexual harassment and unfair treatment being virtually absent in our samples. We recommend collecting larger or more targeted samples when focusing on these scales due to their low incidence.

ACKNOWLEDGMENTS

We are grateful to Angela Byars-Winston, Cheryl Dickter, Lillian Eby, Paul Hernandez, and Justin Lavner for their guidance, advice, and feedback. We thank David Esparza, Levon Esters, Matthew Holtz, Laura Lunsford, Christine Pfund, Zachary Wood, and Michelle Ziadie for providing additional review of our items. We are also grateful to all of our institutional partners who helped us recruit participants, our participants for sharing their time and experiences, and Social Psychology of Research Experiences and Education group members for feedback on drafts of this manuscript. This work was supported in part by National Science Foundation Improving Undergraduate STEM Education Grant 1841061 and the Georgia Athletic Association Professorship for Innovative Science Education. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding organizations.

REFERENCES

Adedokun, O. A., Bessenbacher, A. B., Parker, L. C., Kirkham, L. L., & Burgess, W. D. (2013). Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy. *Journal of Research in Science Teaching*, *50*(8), 940–951. https://doi.org/10.1002/tea.21102

Aikens, M. L., Robertson, M. M., Sadselia, S., Watkins, K., Evans, M., Runyon, C. R., ... & Dolan, E. L. (2017). Race and gender differences in undergraduate research mentoring structures and research outcomes. *CBE—Life Sciences Education*, *16*(2), ar34. https://doi.org/10.1187/cbe.16-07-0211

Aikens, M. L., Sadselia, S., Watkins, K., Evans, M., Eby, L. T., & Dolan, E. L. (2016). A social capital perspective on the mentoring of undergraduate life science researchers: An empirical study of undergraduate–postgraduate–faculty triads. *CBE—Life Sciences Education*, *15*(2), ar16. https://doi.org/10.1187/cbe.15-10-0208

Ainsworth, M. D. (1989). Attachments beyond infancy. *The American Psychologist*, *44*(4), 709–716. https://doi.org/10.1037/0003-066X.44.4.709

Allen, T., & Eby, L. T. (2003). Relationship effectiveness for mentors: Factors associated with learning and quality. *Journal of Management*, *29*(4), 469–486. https://doi.org/10.1016/S0149-2063(03)00021-7

American Association for the Advancement of Science (2011). *Vision and change in undergraduate biology education: A call to action*. Retrieved November 28, 2015, from http://visionandchange.org/finalreport/

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Aycock, L. M., Hazari, Z., Brewe, E., Clancy, K. B. H., Hodapp, T., & Goertzen, R. M. (2019). Sexual harassment reported by undergraduate female physicists. *Physical Review Physics Education Research*, *15*(1), 010121. https://doi.org/10.1103/PhysRevPhysEducRes.15.010121

Bakker, A. B., Demerouti, E., & ten Brummelhuis, L. L. (2012). Work engagement, performance, and active learning: The role of conscientiousness. *Journal of Vocational Behavior*, *80*(2), 555–564. https://doi.org/10.1016/j.jvb.2011.08.008

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-value-cost model of motivation. *Psychology*, *84*, 261–271.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*(1), 10–17. https://doi.org/10.1111/j.1745-3992.1998.tb00616.x

Bernier, A., Larose, S., & Soucy, N. (2005). Academic mentoring in college: The interactive role of student's and mentor's interpersonal dispositions. *Research in Higher Education*, *46*(1), 29–51. https://doi.org/10.1007/s11162-004-6288-5

Bowlby, J. (1979). The Bowlby-Ainsworth attachment theory. *Behavioral and Brain Sciences*, *2*(4), 637–638. https://doi.org/10.1017/S0140525X00064955

Bowlby, J., & Ainsworth, M. (2013). The origins of attachment theory. *Attachment Theory: Social, Developmental, and Clinical Perspectives*, *45*(28), 759–775.

Branchaw, J. L., Butz, A. R., & Smith, A. R. (2020). Evaluation of the second edition of entering research: A customizable curriculum for apprentice-style undergraduate and graduate research training programs and courses. *CBE—Life Sciences Education*, *19*(1), ar11.

Brown, E. R., Thoman, D. B., Smith, J. L., & Diekman, A. B. (2015). Closing the communal gap: The importance of communal affordances in science career motivation. *Journal of Applied Social Psychology*, *45*(12), 662–673. https://doi.org/10.1111/jasp.12327

Byars-Winston, A., & Dahlberg, M. (eds.). (2019). *The Science of effective mentorship in STEMM*. Washington, DC: National Academies Press. https://doi.org/10.17226/25568

Byars-Winston, A. M., Branchaw, J., Pfund, C., Leverett, P., & Newton, J. (2015). Culturally diverse undergraduate researchers' academic outcomes and perceptions of their research mentoring relationships. *International Journal of Science Education*, *37*(15), 2533–2554. https://doi.org/10.1080/09500693.2015.1085133

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data. CRESST Report 839*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Carlone, H. B., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching*, *44*(8), 1187–1218. https://doi.org/10.1002/tea.20237

Carver, C. S. (1997). Adult attachment and personality: Converging evidence and a new measure. *Personality and Social Psychology Bulletin*, *23*(8), 865–883. https://doi.org/10.1177/0146167297238007

Ceyhan, G. D., & Tillotson, J. W. (2020). Early year undergraduate researchers' reflections on the values and perceived costs of their research experience. *International Journal of STEM Education*, *7*(1), 1–19. https://doi.org/10.1186/s40594-020-00248-x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Denissen, J. J. A., & Penke, L. (2008). Neuroticism predicts reactions to cues of social inclusion. *European Journal of Personality*, *22*(6), 497–517. https://doi.org/10.1002/per.682

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Dolan, E., & Johnson, D. (2009). Toward a holistic view of undergraduate research experiences: An exploratory study of impact on graduate/postdoctoral mentors. *Journal of Science Education and Technology*, *18*(6), 487. https://doi.org/10.1007/s10956-009-9165-3

Dolan, E. L., & Johnson, D. (2010). The undergraduate–postgraduate–faculty triad: Unique functions and tensions associated with undergraduate research experiences at research universities. *CBE—Life Sciences Education*, *9*(4), 543–553.

Dolan, E. L., & Weaver, G. C. (2021). *A guide to course-based undergraduate research*, 1st ed., New York, NY: Macmillan Higher Education.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Eagan, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., & Garibay, J. C. (2013). Making a difference in science education: The impact of undergraduate research programs. *American Educational Research Journal*, *50*(4), 683–713. https://doi.org/10.3102/0002831213482038

Eby, L. T., & Allen, T. (2002). Further investigation of protégés' negative mentoring experiences: Patterns and outcomes. *Group & Organization Management*, *27*(4), 456–479. https://doi.org/10.1177/1059601102238357

Eby, L. T., Allen, T. D., Hoffman, B. J., Baranik, L. E., Sauer, J. B., Baldwin, S., ... & Evans, S. C. (2013). An interdisciplinary meta-analysis of the potential antecedents, correlates, and consequences of protégé perceptions of mentoring. *Psychological Bulletin*, *139*(2), 441–476. https://doi.org/10.1037/a0029279

Eby, L. T., Butts, M., Lockwood, A., & Simon, S. A. (2004). Proteges' negative mentoring experiences construct development and nomological validation. *Personnel Psychology*, *57*(2), 441–447.

Eby, L. T., Butts, M. M., Durley, J., & Ragins, B. R. (2010). Are bad experiences stronger than good ones in mentoring relationships? Evidence from the protégé and mentor perspective. *Journal of Vocational Behavior*, *77*(1), 81–92. https://doi.org/10.1016/j.jvb.2010.02.010

Eby, L. T., Durley, J. R., Evans, S. C., & Ragins, B. R. (2008). Mentors' perceptions of negative mentoring experiences: Scale development and nomological validation. *Journal of Applied Psychology*, *93*(2), 358–373. https://doi.org/10.1037/0021-9010.93.2.358

Eby, L. T., & McManus, S. E. (2004). The protégé's role in negative mentoring experiences. *Journal of Vocational Behavior*, *65*(2), 255–275. https://doi.org/10.1016/j.jvb.2003.07.001

Eby, L. T., McManus, S. E., Simon, S. A., & Russell, J. E. A. (2000). The Protege's perspective regarding negative mentoring experiences: The development of a taxonomy. *Journal of Vocational Behavior*, *57*(1), 1–21. https://doi.org/10.1006/jvbe.1999.1726

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. New York, NY: Psychology Press. https://doi.org/10.4324/9781410605269

Erickson, O. A., Cole, R. B., Isaacs, J. M., Alvarez-Clare, S., Arnold, J., Augustus-Wallace, A., ... & Burgio, K. R. (2022). "How do we do this at a distance?!" A descriptive study of remote undergraduate research programs during COVID-19. *CBE—Life Sciences Education*, *21*(1), ar1.

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106–1120. https://doi.org/10.1080/02602938.2020.1724875

Estrada, M., Hernandez, P. R., & Schultz, P. W. (2018). A longitudinal study of how quality mentorship and research experience integrate underrepresented minorities into STEM careers. *CBE—Life Sciences Education*, *17*(1), ar9. https://doi.org/10.1187/cbe.17-04-0066

Estrada, M., Woodcock, A., Hernandez, P. R., & Schultz, P. W. (2011). Toward a model of social influence that explains minority student integration into the scientific community. *Journal of Educational Psychology*, *103*(1), 206–222. https://doi.org/10.1037/a0020743

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS One*, *14*(2). https://doi.org/10.1371/journal.pone.0209749

Fitch, J. C., Pistole, M. C., & Gunn, J. E. (2010). The bonds of development: An attachment-caregiving model of supervision. *The Clinical Supervisor*, *29*(1), 20–34. https://doi.org/10.1080/07325221003730319

Fleming, M., House, S., Hanson, V. S., Yu, L., Garbutt, J., McGee, R., ... & Rubio, D. M. (2013). The mentoring competency assessment: Validation of a new instrument to evaluate skills of research mentors. *Academic Medicine*, *88*(7), 1002–1008. https://doi.org/10.1097/ACM.0b013e318295e298

Frantz, K. J., Demetrikopoulos, M. K., Britner, S. L., Carruth, L. L., Williams, B. A., Pecore, J. L., ... & Goode, C. T. (2017). A comparison of internal dispositions and career trajectories after collaborative versus apprenticed research experiences for undergraduates. *CBE—Life Sciences Education*, *16*(1), ar1.

Gaspard, H., Dicke, A.-L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226. https://doi.org/10.1037/dev0000028

Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663. https://doi.org/10.1037/edu0000003

Gehlbach, H., Brinkworth, M. E., King, A. M., Hsu, L. M., McIntyre, J., & Rogers, T. (2016). Creating birds of similar feathers: Leveraging similarity to improve teacher–student relationships and academic achievement. *Journal of Educational Psychology*, *108*(3), 342–352. https://doi.org/10.1037/edu0000042

Gentile, J., Brenner, K., & Stephens, A. (eds.). (2017). *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*. Washington, DC: National Academies Press. Retrieved May 17, 2017, from https://www.nap.edu/catalog/24622/undergraduate-research-experiences-for-stem-students-successes-challenges-and-opportunities

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341–364. https://doi.org/10.1007/s11162-016-9429-8

Gross, J. J., & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In: *Handbook of emotion regulation*, New York, NY: The Guilford Press, 3–24.

Han, G. H., Harms, P. D., & Bai, Y. (2017). Nightmare bosses: The impact of abusive supervision on employees' sleep, emotions, and creativity. *Journal of Business Ethics*, *145*(1), 21–31. https://doi.org/10.1007/s10551-015-2859-y

Hazan, C., & Shaver, P. R. (1994). Attachment as an organizational framework for research on close relationships. *Psychological Inquiry*, *5*(1), 1–22. https://doi.org/10.1207/s15327965pli0501_1

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, *47*(8), 978–1003. https://doi.org/10.1002/tea.20363

Henle, C. A., & Gross, M. A. (2014). What have I done to deserve this? Effects of employee personality and emotion on abusive supervision. *Journal of Business Ethics*, *122*(3), 461–474. https://doi.org/10.1007/s10551-013-1771-6

Hernandez, P. R. (2018). Landscape of assessments of mentoring relationship processes in postsecondary STEMM contexts: A synthesis of validity evidence from mentee, mentor, and institutional/programmatic perspectives. *Commissioned Paper Prepared for the NASEM Committee on the Science of Effective Mentoring in Science, Technology, Engineering, Medicine, and Mathematics (STEMM)*. Retrieved August 5. 2022, from https://nap.nationalacademies.org/resource/25568/Hernandez%20-%20Landscape%20of%20Assessments%20of%20Mentoring.pdf

Hernandez, P. R., Adams, A. S., Barnes, R. T., Bloodhart, B., Burt, M., Clinton, S. M., … & Fischer, E. V. (2020). Inspiration, inoculation, and introductions are all critical to successful mentorship for undergraduate women pursuing geoscience careers. *Communications Earth & Environment*, *1*(1). https://doi.org/10.1038/s43247-020-0005-y

Hernandez, P. R., Estrada, M., Woodcock, A., & Schultz, P. W. (2017). Protégé perceptions of high mentorship quality depend on shared values more than on demographic match. *The Journal of Experimental Education*, *85*(3), 450–468. https://doi.org/10.1080/00220973.2016.1246405

Hernandez, P. R., Ferguson, C. F., Pedersen, R., Richards-Babb, M., Quedado, K., & Shook, N. J. (2023). Research apprenticeship training promotes faculty-student psychological similarity and high-quality mentoring: A longitudinal quasi-experiment. *Mentoring & Tutoring: Partnership in Learning*, *31*(1), 163–183. https://doi.org/10.1080/13611267.2023.2164973

Hernandez, P. R., Hopkins, P. D., Masters, K., Holland, L., Mei, B. M., Richards-Babb, M., … & Shook, N. J. (2018). Student integration into STEM careers and culture: A longitudinal examination of summer faculty mentors and project ownership. *CBE—Life Sciences Education*, *17*(3), ar50. https://doi.org/10.1187/cbe.18-02-0022

Hernandez, P. R., Woodcock, A., Estrada, M., & Schultz, P. W. (2018). Undergraduate research experiences broaden diversity in the scientific workforce. *BioScience*, *68*(3), 204–211. https://doi.org/10.1093/biosci/bix163

Hess, R. A., Erickson, O. A., Cole, R. B., Isaacs, J. M., Alvarez-Clare, S., Arnold, J., … & Dolan, E. L. (2023). Virtually the same? Evaluating the effectiveness of remote undergraduate research experiences. *CBE—Life Sciences Education*, *22*(2), ar25. https://doi.org/10.1187/cbe.22-01-0001

Hu, L., & 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. https://doi.org/10.1080/10705519909540118

Hunter, A.-B., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education*, *91*(1), 36–74. https://doi.org/10.1002/sce.20173

Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L., & Espinosa, L. L. (2009). Diversifying science: Underrepresented student experiences in structured research programs. *Research in Higher Education*, *50*(2), 189–214. https://doi.org/10.1007/s11162-008-9114-7

Indiana University Center for Postsecondary Research (n.d.). *Carnegie classification of institutions of higher education*. Retrieved December 18, 2018, from http://carnegieclassifications.iu.edu/

John, O. P. (2021). History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures. In *Handbook of personality: Theory and research*, 4th ed. (pp. 35–82) New York, NY: The Guilford Press.

Joshi, M., Aikens, M. L., & Dolan, E. L. (2019). Direct ties to a faculty mentor related to positive outcomes for undergraduate researchers. *BioScience*, *69*(5), 389–397. https://doi.org/10.1093/biosci/biz039

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Kardash, C. M. (2000). Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology*, *92*(1), 191–201. https://doi.org/10.1037/0022-0663.92.1.191

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.

Kram, K. E. (1983). Phases of the mentor relationship. *Academy of Management Journal*, *26*(4), 608–625. https://doi.org/10.2307/255910

Lee, S. P., McGee, R., Pfund, C., & Branchaw, J. (2015). Mentoring up: Learning to manage your mentoring relationships. *The Mentoring continuum: From graduate school through tenure*. Syracuse, NY: Graduate School Press of Syracuse University.

Limeri, L. B., Asif, M. Z., Bridges, B. H. T., Esparza, D., Tuma, T. T., Sanders, D., … & Dolan, E. L. (2019a). "Where's my mentor?!" Characterizing negative mentoring experiences in undergraduate life science research. *CBE—Life Sciences Education*, *18*(4), ar61. https://doi.org/10.1187/cbe.19-02-0036

Limeri, L. B., Asif, M. Z., & Dolan, E. L. (2019b). Volunteered or Voluntold? The motivations and perceived outcomes of graduate and postdoctoral mentors of undergraduate researchers. *CBE—Life Sciences Education*, *18*(2), ar13. https://doi.org/10.1187/cbe.18-10-0219

Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*(6222), 1261757. https://doi.org/10.1126/science.1261757

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nahm, A. Y., Rao, S. S., Solis-Galvan, L. E., & Ragu-Nathan, T. S. (2002). The Q-sort method: Assessing reliability and construct validity of questionnaire items at a pre-testing stage. *Journal of Modern Applied Statistical Methods*, *1*(1), 114–125. https://doi.org/10.22237/jmasm/1020255360

National Academies of Sciences, Engineering, and Medicine. (2018). *Graduate STEM education for the 21st century*. Washington, DC: National Academies Press. https://doi.org/10.17226/25038

O'Meara, K., Jaeger, A., Misra, J., Lennartz, C., & Kuvaeva, A. (2018). Undoing disparities in faculty workloads: A randomized trial experiment. *PloS One*, *13*(12), e0207316.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*,*24*(1),50–64.https://doi.org/10.1177/01466216000241003

Pedersen, R. M., Ferguson, C. F., Estrada, M., Schultz, P. W., Woodcock, A., & Hernandez, P. R. (2022). Similarity and contact frequency promote mentorship quality among Hispanic undergraduates in STEM. *CBE—Life Sciences Education*, *21*(2), ar27. https://doi.org/10.1187/cbe.21-10-0305

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In: *Handbook of research on student engagement*, ed. S. L. Christenson, A. L. Reschly, & C. Wylie, Boston, MA: Springer US, 259–282. https://doi.org/10.1007/978-1-4614-2018-7_12

Pfund, C., Branchaw, J. L., & Handelsman, J. (2015). *Entering mentoring*, 2nd ed., New York, NY: Macmillan. Retrieved March 21, 2019, from http://www.macmillanlearning.com/Catalog/product/enteringmentoring-revised-pfund

Porath, C. L., & Pearson, C. M. (2012). Emotional and behavioral responses to workplace incivility and the impact of hierarchical status. *Journal of Applied Social Psychology*, *42*(S1). https://doi.org/10.1111/j.1559-1816.2012.01020.x

R Core Team (2021). *R: A language and environment for statistical computing [Computer software]*. R Foundation for Statistical Computing. Retrieved June 23, 2023, from https://www.R-project.org/

Ragins, B. R., Cotton, J. L., & Miller, J. S. (2000). Marginal mentoring: The effects of type of mentor, quality of relationship, and program design on work and career attitudes. *The Academy of Management Journal*, *43*(6), 1177–1194. https://doi.org/10.2307/1556344

Robinson, C. D., Scott, W., & Gottfried, M. A. (2019). Taking it to the next level: A field experiment to improve instructor-student relationships in college. *AERA Open*, *5*(1), 2332858419839707. https://doi.org/10.1177/2332858419839707

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *Journal of Research in Science Teaching*, *52*(6), 847–867. https://doi.org/10.1002/tea.21221

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

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113–130. https://doi.org/10.1080/02602930802618344

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series*, *1968*(1). https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Scandura, T. A. (1998). Dysfunctional mentoring relationships and outcomes. *Journal of Management*, *24*(3), 449–467. https://doi.org/10.1016/S0149-2063(99)80068-3

Scherer, K. R. (1999). Appraisal theory. In: *Handbook of cognition and emotion*, New York, NY: John Wiley & Sons Ltd, 637–663. https://doi.org/10.1002/0470013494.ch30

Schilpzand, P., De Pater, I. E., & Erez, A. (2016). Workplace incivility: A review of the literature and agenda for future research. *Journal of Organizational Behavior*, *37*(S1), S57–S88. https://doi.org/10.1002/job.1976

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Simon, S. A., & Eby, L. T. (2003). A typology of negative mentoring experiences: A multidimensional scaling study. *Human Relations*, *56*(9), 1083–1106. https://doi.org/10.1177/0018726703569003

Sonnenberg-Klein, J., Abler, R. T., Coyle, E. J., & Ai, H. H. (2017). Multidisciplinary vertically integrated teams: Social network analysis of peer evaluations for Vertically Integrated Projects (VIP) program teams. *Paper presented at 2017 ASEE Annual Conference & Exposition, Columbus, Ohio*.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*(1), 138–161. https://doi.org/10.1037/0033-2909.134.1.138

Strachan, S. M., Marshall, S., Murray, P., Coyle, E. J., & Sonnenberg-Klein, J. (2019). Using vertically integrated projects to embed research-based education for sustainable development in undergraduate curricula. *International Journal of Sustainability in Higher Education*, *20*(8), 1313–1328.

Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal*, *43*(2), 178–190. https://doi.org/10.5465/1556375

Tepper, B. J., Simon, L., & Park, H. M. (2017). Abusive supervision. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 123–152.

Thiry, H., & Laursen, S. L. (2011). The role of student-advisor interactions in apprenticing undergraduate researchers into a scientific community of practice. *Journal of Science Education and Technology*, *20*(6), 771–784.

Tuma, T. T., Adams, J. D., Hultquist, B. C., & Dolan, E. L. (2021). The dark side of development: A systems characterization of the negative mentoring experiences of doctoral students. *CBE—Life Sciences Education*, *20*(2), ar16. https://doi.org/10.1187/cbe.20-10-0231

Turban, D. B., & Jones, A. P. (1988). Supervisor-subordinate similarity: Types, effects, and mechanisms. *Journal of Applied Psychology*, *73*(2), 228. https://doi.org/10.1037/0021-9010.73.2.228

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*(4), 751–796. https://doi.org/10.3102/0034654308321456

Widiger, T. A., & Oltmanns, J. R. (2017). Neuroticism is a fundamental domain of personality with enormous public health implications. *World Psychiatry*, *16*(2), 144–145. https://doi.org/10.1002/wps.20411

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wu, T.-Y., & Hu, C. (2013). Abusive supervision and subordinate emotional labor: The moderating role of openness personality. *Journal of Applied Social Psychology*, *43*(5), 956–970. https://doi.org/10.1111/jasp.12060

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data. *Multivariate Behavioral Research*, *53*(5), 731–755. https://doi.org/10.1080/00273171.2018.1480346