Supplemental Material

CBE—Life Sciences Education

Farrar *et al*.

Supplemental Materials

Supplemental Table 1. Number of students missing demographic variable information, by number of variables missing and as a percent of the total working sample size.

Demographic variable missing	*N*	*Number of demographic variables missing*	*N (Number missing 0, 1, 2, or 3 variables)*	*Percent of sample size (%)*
None	10,893	0	10,893	82.6
Race/ethnicity only	150	1	1,959	14.9
Socioeconomic status only	1,686			
First-Generation status only	123			
Race/ethnicity & SES status	61	2	297	2.3
Race/ethnicity & first-generation	2			
SES status & first-generation status	234			
All three variables missing	35	3	35	0.3

Supplemental Table 2. Model output for the best-fit model including interactions (Model II), run only in the subset of students that had full demographic variable data (i.e., did not have any missing demographic data, n = 10,893; see Supplemental Table 1).

Variable	*Estimate (β)*	*Std.Error*	*p*
(Intercept)	-0.810	0.076	**<0.001**
PriorGPA	1.187	0.024	**<0.001**
Gender	-0.419	0.090	**<0.001**
PEER	-0.042	0.017	**0.012**
FirstGen	-0.079	0.015	**<0.001**
LowSES	-0.009	0.015	0.556
PriorGPA*Gender	0.078	0.029	**0.007**

Supplemental Table 3. Percent of students in the data set with specific ethnicities.

Ethnicity	*Percent*
White/Caucasian	28.0
Chinese American/Chinese	19.4
Vietnamese	9.5
East Indian/Pakistani	8.1
Filipino American/Filipino	5.2
Other Asian	4.0
Korean American/Korean	2.8
Japanese American/Japanese	1.8
Other Pacific Islander	0.4
Persons Excluded because of Ethnicity or Race (PEER)[a]	*(19.2)*
Latinx/Chicanx	15.3
Black/African American	2.7
American Indian/Alaska Native/Indigenous	0.8
Other	0.4

[a] PEER students are defined as Black/African American, Latinx or Chicanx, American Indian/Indigenous or multiracial ("Other") (Asai 2020).

Supplemental Table 4. Comparison of models using prior overall GPA and course prerequisites.

Fixed effect included[1]	*df*	*AIC*	*dAIC*	*BIC*	*dBIC*	*logLik*	*dLogLik*
PriorGPA	**8**	**16526.0**	**0.0**	**16581.9**	**0.0**	**-8255.0**	**0.0**
Introductory Biology 1 Grade	8	18416.6	1890.6	18472.6	1890.6	-9200.3	-945.3
Introductory Biology 2 Grade	8	18329.7	1803.7	18385.7	1803.7	-9156.9	-901.9
Introductory Biology 3 Grade	8	17586.7	1060.8	17642.7	1060.8	-8785.4	-530.4
Introductory Chemistry 1 Grade	8	19377.3	2851.3	19433.3	2851.3	-9680.7	-1425.7

[1] Fixed-effect only models (which also included Gender, PEER, FirstGen and Low SES as fixed effects and course offering as a random effect) were compared to see which indicator of students' prior performance best fit our dataset. For the subset of students which had prior GPA data and grades in all course prerequisites (n = 8063), we compared models that included either prior GPA, introductory biology grades (a three-part series typically taken each quarter of the first year) and the first introductory chemistry course grade. The best-fit model (in bold) for all information criteria included PriorGPA, thus, we moved forward with using prior GPA as the best proxy for students' prior course performance.

Supplemental Table 5. Top 5 models for our dataset, ranked by AIC.

Prior GPA	*Gender*	*PEER*	*FirstGen*	*LowSES*	*Transfer*	*Quarter*	*ESL*	*df*	*AIC*	*BIC*	*logLik*	*dAIC*	*weight*
X	X	X	X	X	X	-	-	9	28662.3	28729.7	-14322.1	0.00	0.16
X	**X**	**X**	**X**	**X**	**-**	**-**	**-**	**8**	**28662.5**	**28722.4**	**-14323.3**	**0.2**	**0.14**
X	X	X	X	X	-	X	-	10	28662.7	28737.5	-14321.3	0.4	0.13
X	X	X	X	X	-	X	-	9	28663.0	28730.4	-14322.5	0.7	0.11
X	X	X	X	X	X	-	X	10	28664.3	28739.1	-14322.1	2.0	0.06

Note: After evaluating models containing all possible combinations of fixed effects from the global model, we found 5 models that had a delta AIC < 2 from the model with the lowest AIC value. The fixed effects included in each of these models are shown, with "-" indicating the variable was not included in the model, and "X" indicating that the variable was included. All models shown included a random effect of the specific course offering. Delta AIC is reported relative to the best-fit model (top row). We selected the model in bold as the "best fit" model as it had the lowest BIC and log-likelihood and did not include any variables that were not significant in the average of the top models (see Methods for details).

Supplemental Table 6. Model averaged coefficients for the top 5 models.

Variable	*Estimate (β)*	*Std.Error*	*p*
(Intercept)	-0.827	0.065	**<0.001**
PriorGPA	1.193	0.013	**<0.001**
Gender	-0.198	0.013	**<0.001**
FirstGen	-0.075	0.014	**<0.001**
PEER	-0.069	0.016	**<0.001**
LowSES	-0.031	0.015	**0.044**
Transfer	-0.012	0.015	0.412
Quarter	-0.013	0.023	0.567
ESL	-0.000	0.005	0.974

Note: After evaluating models containing all possible combinations of fixed effects from the global model, we found 5 models that had a delta AIC < 2 from the model with the lowest AIC and BIC value ("best fit" model) (see Supplemental Table 3). Here, we show the model averaged coefficients for each fixed effect. The coefficient is averaged across all models, where each separate model's coefficient for that fixed effect is multiplied by the Akaike weight of the model (Akaike weights can be found in Supplemental Table 3). Fixed effect coefficients are set to 0 if they were not included in the model (full model average). *p*-values reflect whether the coefficients differ significantly from zero (*z*-test; alpha = 0.05).

Supplemental Table 7. Comparison of standard multilevel model estimates and those from robust estimation for the best-fit model including interactions as determined by model selection (Model II).

	Standard model estimates			*Robust model estimates*		
Model I						
CourseGrade ~ PriorGPA + Gender + PEER + FirstGen + LowSES + (1|Offering)						
Variable	*Estimate(β)*	*SE*	*p*	*Estimate(β)*	*SE*	*p*
(Intercept)	-0.859	0.049	**<0.001**	-0.932	0.043	**<0.001**
PriorGPA	1.195	0.013	**<0.001**	1.123	0.012	**<0.001**
Gender	-0.197	0.013	**<0.001**	-0.189	0.012	**<0.001**
PEER	-0.069	0.016	**<0.001**	-0.049	0.015	**<0.001**
FirstGen	-0.076	0.014	**<0.001**	-0.070	0.013	**<0.001**
LowSES	-0.030	0.015	**0.049**	-0.021	0.014	0.145
Model II						
*CourseGrade ~ PriorGPA + Gender + PEER + FirstGen + LowSES + PriorGPA*Gender + (1|Offering)*						
Variable	*Estimate(β)*	*SE*	*p*	*Estimate(β)*	*SE*	*p*
(Intercept)	-0.621	0.073	**<0.001**	-0.708	0.066	**<0.001**
PriorGPA	1.117	0.022	**<0.001**	1.158	0.020	**<0.001**
Gender	-0.570	0.085	**<0.001**	-0.540	0.079	**<0.001**
PEER	-0.069	0.016	**<0.001**	-0.048	0.015	**0.002**
FirstGen	-0.074	0.014	**<0.001**	-0.068	0.013	**<0.001**
LowSES	-0.030	0.015	**0.045**	-0.021	0.014	0.130
PriorGPA*Gender	0.121	0.027	**<0.001**	0.113	0.025	**<0.001**

Note: As our original best-fit model did not meet assumptions of normality and homoscedasticity of residuals due to outliers, we also ran the same linear model using robust estimation using the R package `robustlmm` (Koller 2016). Robust model estimation differentially weights residuals from outliers to address these deviations from assumptions. We show that estimates and significance are relatively similar between standard mixed model and robust estimation. Robust estimates shown here are the same as those shown in Table 5 in the main text.

Supplemental Table 8. Robustness weights for the best fit model determined by model selection.

Robustness Weights						
	Model I CourseGrade ~ PriorGPA + Gender + PEER + FirstGen + LowSES + (1|Offering)			*Model II* CourseGrade ~ PriorGPA + Gender + PEER + FirstGen + LowSES + PriorGPA*Gender + (1|Offering)		
	Observations (%)	*Mean weight*	*Median weight*	*Observations (%)*	*Mean weight*	*Median weight*
Residuals weights not ≅ 1	2713 (20.6%)	0.761	0.799	2713 (20.6%)	0.760	0.799
Random effects weights not ≅ 1	6 (17.1%)	0.613	0.654	6 (17.1%)	0.613	0.654

Note: Robustness weights detail the number of observations that had residuals re-weighted by the robust estimation (presented as both N and as a percentage of overall sample size), and the mean weighting for those residuals that were not assigned a weight of 1. The same number of observations were set to be re-weighted in robust estimation for both Model I and Model II, and thus were run on the same weighted data. Output produced by the `robustlmm` package in R (Koller, 2016).

Supplemental Table 9. Model output for the best-fit model including interactions (Model II), run only in the subset of students that responded to the introductory surveys.

Variable	*Estimate (β)*	*Std.Error*	*p*
(Intercept)	-0.168	0.300	0.576
PriorGPA	0.983	0.093	**<0.001**
Gender	-1.156	0.343	**0.001**
PEER	-0.110	0.062	0.078
FirstGen	0.006	0.060	0.922
LowSES	-0.016	0.062	0.792
PriorGPA*Gender	0.325	0.108	**0.003**

Note: We ran the best-fit model (see Tables 4 and 5; Methods) modeling the effects of demographic and academic factors on course grades using only the subset of students that responded to the introductory surveys (n = 896; data from fall 2018, winter and spring 2019). Course offering was included as a random effect. Model estimates and standard errors are shown, along with whether an estimate was significantly different from 0 (*z*-test).

Supplemental Table 10. Predicted course grades for a range of theoretical prior GPAs from a simple model including only PriorGPA + Gender + PriorGPA*Gender + (1|Section).

Prior GPA	*Predicted course grade*		*Difference in predicted course grades (Women – Men)*
	Man	*Woman*	
1.00	0.37	-0.07	-0.44
2.00	1.54	1.22	-0.32
2.50	2.13	1.86	-0.27
3.00	**2.72**	**2.52**	**-0.20**
3.5	3.31	3.16	-0.15
4.00	3.90	3.81	-0.09

Note: This table highlights the predicted differences in human physiology grade outcomes between men and women with the same incoming prior GPA. The values shown are generated from a simple model, specified by CourseGrade ~ -0.74 + $\text{ß}_{\text{PriorGPA}}$ [1.13] + ß_{Gender} [-0.59] + $\text{ß}_{\text{GPA*Gender}}$ [0.13]. Students receive letter grades rather than grade points and cannot receive negative grade point values. The bold row highlights expected outcomes for a prior GPA around 3.0, which approximates the average for both men and women in the course (3.09 versus 3.07 for men and women, on average, respectively).

Supplemental Table 11. Models including main effects only with and without prior GPA.

	Robust model estimates (Model I)			*Robust estimates without PriorGPA*		
Variable	*β*	*SE*	*p*	*β*	*SE*	*p*
(Intercept)	-0.93	0.04	**<0.001**	3.02	0.02	**<0.001**
PriorGPA	1.12	0.01	**<0.001**	-	-	-
Gender	-0.19	0.01	**<0.001**	-0.22	0.02	**<0.001**
PEER	-0.05	0.02	**<0.001**	-0.22	0.02	**<0.001**
FirstGen	-0.07	0.01	**<0.001**	-0.22	0.02	**<0.001**
LowSES	-0.02	0.01	0.145	-0.14	0.02	**<0.001**

Note: Gender, PEER status, first generation status, low-socioeconomic status all negatively relate to course outcomes, and model estimates are more negative without controls for prior academic performance (i.e., *PriorGPA*).

Supplemental Table 12. Predicted course grades for a range of theoretical prior GPAs based upon the estimates for the best model including interactions (Model II).

Prior GPA	*Predicted course grade*		*Difference in predicted course grades (*W_{PFGLS} – M*)*
	M	W_{PFGLS}	
1.00	0.45	-0.11	-0.56
2.00	1.61	1.16	-0.45
2.50	2.19	1.79	-0.39
3.00	**2.77**	**2.43**	**-0.34**
3.50	3.35	3.07	-0.28
4.00	3.92	3.70	-0.22

Note: This table highlights the predicted differences in human physiology grade outcomes between a man with access to all other systemic advantages measured (student M; i.e., non-PEER, non-first-generation, non-low-socio economic status) and a woman without those systemic advantages (student W_{PFGLI}; i.e., PEER, first-generation, low-socioeconomic status) with the same incoming prior GPA (see *Results*). Model estimates used to calculate these theoretical values can be found in Table 5.

Note that these are theoretical values; students receive letter grades rather than grade points and cannot receive negative grades point values. The bold row highlights expected outcomes for a prior GPA around 3.0, which approximates the average for both men and women in the course (3.09 versus 3.07 for men and women, on average, respectively).

Supplemental Table 13. Top 7 models for the 2018-2019 dataset for affective survey data, ranked by AICc.

Prior GPA	*Gender*	*Science Identity*	*Science Self-Efficacy (SSE)*	*Course Anxiety*	*PriorGPA*Gender*	*Gender*ScienceIdentity*	*Gender*SSE*	*Gender*Anxiety*	*df*	*AICc*	*logLik*	*dAIC*	*weight*
X	**X**	**X**	**X**	**X**	**X**	**-**	**-**	**X**	**9**	**1643.7**	**-812.8**	**0.0**	**0.25**
X	X	X	X	X	X	-	-	-	8	1644.3	-814.1	0.6	0.18
X	X	X	X	X	X	-	X	X	10	1645.0	-812.4	1.3	0.13
X	X	X	X	X	X	X	-	X	10	1645.0	-812.4	1.3	0.13
X	X	X	X	-	X	-	-	-	7	1645.3	-815.6	1.5	0.12
X	X	X	X	X	X	X	X	X	11	1645.7	-811.7	1.9	0.10
X	X	X	X	X	X	X	-	-	9	1645.7	-813.7	2.0	0.09

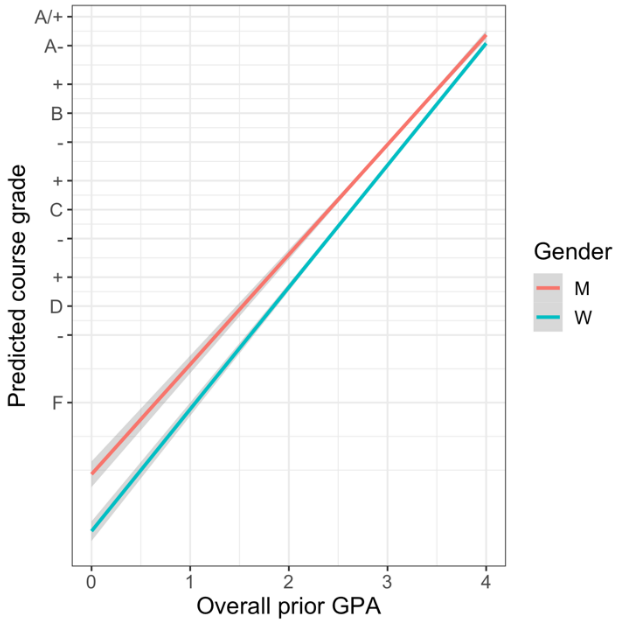
Note: After evaluating models containing main effects of gender, Prior GPA, and each affective factor, as well as interactions between gender and prior GPA and interactions between gender and all affective factors, seven models emerged that had a delta AICc < 2 from the model with the lowest AICc value. The fixed effects included in each of these models are shown, with “-“indicated the variable was not included in the model, and “X” indicating that the variable was included. Delta AICc is reported relative to the best-fit model (top row). AICc was used in lieu of AIC due to the smaller sample size of the survey dataset (three offerings, n = 896).

Supplemental Table 14. Model averaged coefficients for the top 7 models for affective survey data.

Variable	*Estimate (β)*	*Std.Error*	*p*
(Intercept)	-0.241	0.325	0.458
PriorGPA	0.972	0.103	**0.000**
Gender	-0.991	0.376	**0.008**
CourseAnxiety	-0.001	0.041	0.987
ScienceIdentity	0.092	0.055	0.094
SSE	0.079	0.036	**0.027**
PriorGPA*Gender	0.291	0.118	**0.014**
Gender*CourseAnxiety	-0.049	0.055	0.371
Gender*SSE	-0.013	0.035	0.721
Gender*ScienceIdentity	0.024	0.056	0.676

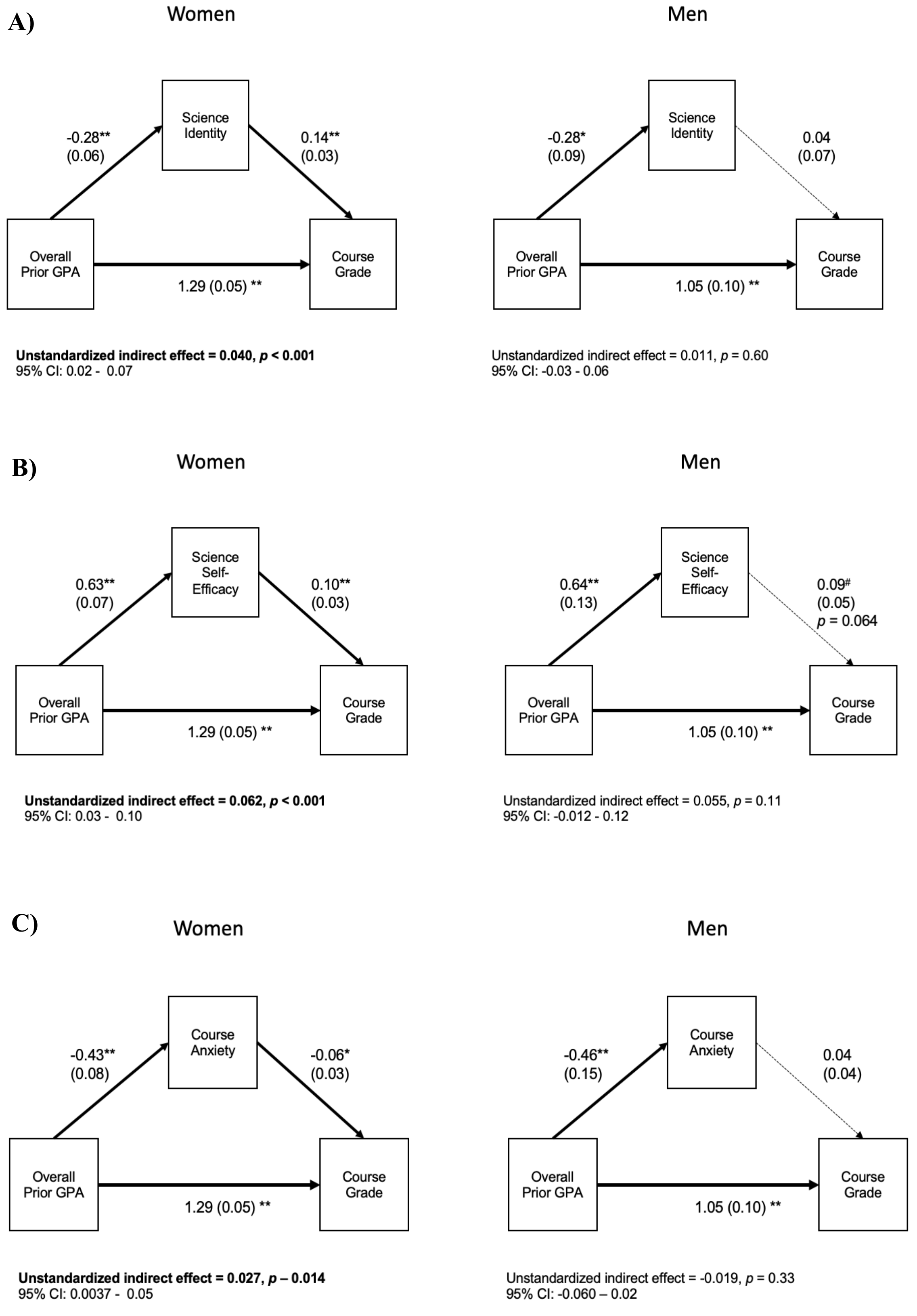
Note: After evaluating models containing main effects of gender, Prior GPA, and each affective factor, as well as interactions between gender and prior GPA and interactions between gender and all affective factors, seven models emerged that had a delta AICc < 2 from the model with the lowest AICc value (see Supplemental Table 10). Here, we show the model averaged coefficients for each fixed effect. The coefficient is averaged across all models, where each separate model's coefficient for that fixed effect is multiplied by the Akaike weight of the model (weights can be found in Supplemental Table 10). Fixed effect coefficients are set to 0 if they were not included in the model (full model average). *P*-values reflect whether the coefficients differ significantly from zero (z test).

Supplemental Figure 1. Predicted grades.



Supplemental Figure 1. Predicted course grades for a range of theoretical prior GPAs from a simple model including only PriorGPA + Gender + PriorGPA*Gender + (1|Section), where $ß_{PriorGPA} = 1.13$, $ß_{Gender} = -0.59$, and $ß_{GPA*Gender} = 0.13$.

Supplemental Figure 2. Mediation analyses for affective factors.



A) Women: **Unstandardized indirect effect = 0.040, *p* < 0.001**
95% CI: 0.02 - 0.07

A) Men: Unstandardized indirect effect = 0.011, *p* = 0.60
95% CI: -0.03 - 0.06

B) Women: **Unstandardized indirect effect = 0.062, *p* < 0.001**
95% CI: 0.03 - 0.10

B) Men: Unstandardized indirect effect = 0.055, *p* = 0.11
95% CI: -0.012 - 0.12

C) Women: **Unstandardized indirect effect = 0.027, *p* – 0.014**
95% CI: 0.0037 - 0.05

C) Men: Unstandardized indirect effect = -0.019, *p* = 0.33
95% CI: -0.060 – 0.02

Supplemental Figure 2. Partial mediation analyses show gender differences in the significant effects of affective factors on mediating the relationship between prior performance (prior GPA) and course grades in upper division physiology. In these partial mediation models, prior GPA both directly and indirectly through either science identity (**A**), science self-efficacy (**B**) or course anxiety (**C**) affects student course grades. Dark arrows representing significant relationships, and light/dashed arrows representing non-significant relationships. Estimates are shown on each arrow, with standard errors in parentheses. For each partial mediation, women are shown on the left and men on the right. Prior GPA significantly affected scores for all affective factors in both men and women, but these affective factor scores only significantly affected course grades for women, not for men. Mediation analyses were completed using the "mediation" package in R (v.4.5.0; Tingley et al., 2014).
*$p < 0.05$, **$p < 0.01$