Supplemental Material

CBE—Life Sciences Education

Wilton *et al*.

Supplemental Materials

This document is comprised of two sections:

1. Supplemental Figures and Tables
2. Qualitative and Rasch Methodology and Results that summarize our research approach. Further detailed information can be found in Clairmont, 2020.

Section 1: Supplemental Figures and Tables

Supplemental Table 1. Descriptive Data of the 2014-2016 Biology Cohorts

	Fall Year 1	Fall Year 2 (% retained)
Biology major students (n)	3008	2354 (78.25%)
Female	1961	1493 (76.13%)
PEERs (URM)	1045	758 (72.5%)
EOP	1123	848 (75.5%)
First-Generation	1363	1025 (75.2%)
Students <C- in CHEM 1A	727	471 (65%)

Table describes demographic information of interest.
Biology Major Students - students in the biology major overall, not just BIOME.
Female = Female Coded as 1 (Male not presented).
PEERs = Persons Excluded because of Ethnicity or Race;
URM = Underrepresented Minorities
EOP = Educational Opportunity Program - A university-calculated student classification based on factors such as parent income, parent education, student background characteristics
First Generation - Student first in family to go to college
Students <C- in CHEM 1A = Whether a student got less than a passing grade in the first Chemistry course necessary advancement in the biology major.

Supplemental Table 2. This course schedule is for a 10-week quarter system. The far-right column describes both the subject area and survey item category associated with that topic.

Week	Topics and Themes Discussed	Peer Mentorship Framework Item(s)
1	Identification of mentors and how to use mentorship. Introduction to near-peer Mentors. How to enroll in Campus Learning Assistance Services (CLAS).	Instrumental and Academic support
2	Effective time-management: How to block and schedule time. Start time-log assignment. How to effectively use office hours.	Instrumental and Academic support

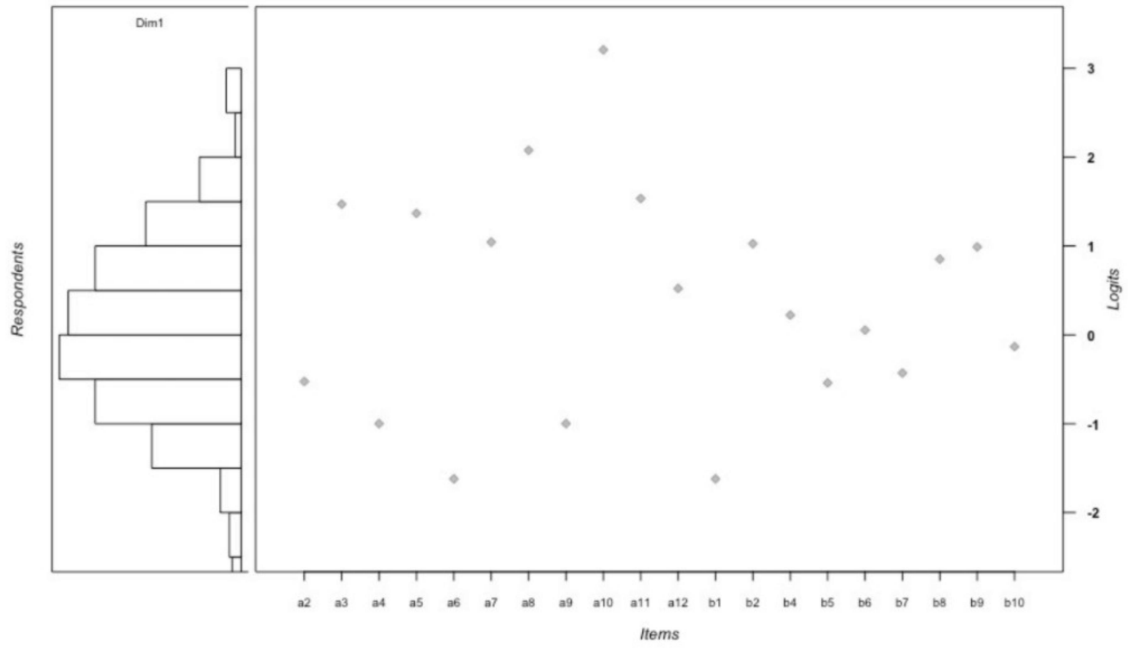
3	Reflect on time-log assignment. Growth mindset reading and intervention.	Psychosocial and Academic support
4	Chemistry midterm exam-wrapper: How to correct and learn from your completed *General Chemistry* exam.	Psychosocial and Academic support
5	Setting up SMART goals: The importance of setting academic goals. Write out, upload, and discuss goals with peers and mentors.	Psychosocial and Instrumental support
6	Study Groups: how to find peers and effectively study as a group.	Academic support
7	Science Podcast: discussion of current topics in biology (e.g. HeLa cells).	Psychosocial support
8	Science as a Career: how to find internships, research lab positions, scholarships, and write a resume. Meet the STEM and Health Sciences counselors.	Psychosocial and Instrumental support
9	Holiday week/optional class meeting: online assignment of finding an internship of interest through university portal.	Instrumental support
10	Paying it forward: Advice to future first-year biology students? Reflective assignment and discussion on what academic habits were successful - what to change for subsequent quarters.	Psychosocial, Instrumental, and Academic support

Supplemental Figure 1A. Item difficulties, left column, are in logits, and represent the ability level at which a person would have a 50% chance of endorsing an item. Items are dichotomous - students were asked “how many times in the last two weeks have you…” and then presented with options to select. A selected item for a student was coded as 1, otherwise, 0. No standard errors of measurement are presented, here, hence, there may be some potential item overlap.

Ranked Rasch Difficulty	Survey Item
-2.05	Worked on practice questions that won't be graded
-1.63	Asked another student a question about schoolwork
-1.26	Planned my social time around my study schedule
-0.98	Sought information about an internship or lab position

-0.49	Started studying for a test or quiz more than three days in advance
-0.47	Studied with another student in my class
-0.43	Marked problems or concepts to study again later
-0.36	Double-checked my work before turning it in
-0.31	Attended a tutoring session, such as CLAS
-0.07	Reworked problems that I missed on previous assignments
0.21	Delayed a reward for myself until after I met my academic goals for the day
0.55	Chosen not to spend time with people who keep me from getting my work done
0.58	Practiced for tests or quizzes using a timer
0.73	Answered a question asked by another student about school work
0.85	Created and followed a study schedule
0.9	Taken a step back from my work to judge my overall understanding
0.92	Planned ahead to take a relaxing break before a test or quiz
1.21	Made use of campus-based support programs such as Student Health, CAPS, or the AS Food Bank
1.34	Emailed a TA or faculty member directly
1.44	Talked to a TA or faculty member outside of class, such as office hours
2.09	Sought information about a professor of a class I want to take
2.95	Participated in activities with an academic society or academically oriented sorority or fraternity

Supplemental Figure 1B. A Wright Map shows the distribution of item difficulties compared to the estimates of person abilities. The observed distribution of person abilities is on the left, and on the right, the observed distribution of item difficulties. This helps describe targeting of the items to person to check whether items can provide information on students across the full range of academic habit abilities - that is, there are hard items to get information about students of higher abilities and easier items to get information about students of lower ability.



Supplemental Table 3. Item text, Item Difficulty, and Differential Item Functioning

Item Difficulty and Fit in Academic Habit Complexity Scale

	Delta	Outfit MNSQ	Infit MNSQ
Attended a tutoring session, such as CLAS	-0.31	1.155	1.111
Studied with another student in my class	-0.47	0.988	0.996
Talked to a TA or faculty member outside of class, such as office hours	1.44	0.873	0.937
Planned my social time around my study schedule	-1.26	0.939	0.967

Emailed a TA or faculty member directly	1.34	1.013	1.006
Asked another student a question about school work	-1.63	0.765	0.872
Answered a question asked by another student about school work	0.73	1.008	1.010
Sought information about a professor of a class I want to take	2.09	1.197	1.057
Sought information about an internship or lab position	-0.98	0.825	0.876
Participated in activities with an academic society or academically-oriented sorority or fraternity	2.95	1.032	0.987
Made use of campus-based support programs such as Student Health, CAPS, or the AS Food Bank	1.21	1.255	1.110
Chosen not to spend time with people who keep me from getting my work done	0.55	1.034	1.028
Worked on practice questions that won't be graded	-2.05	0.877	0.928
Created and followed a study schedule	0.85	1.176	1.110
Practiced for tests or quizzes using a timer	0.58	1.162	1.101
Delayed a reward for myself until after I met my academic goals for the day	0.21	0.973	0.981
Started studying for a test or quiz more than three days in advance	-0.49	0.915	0.952

Reworked problems that I missed on previous assignments	-0.07	0.902	0.918
Marked problems or concepts to study again later	-0.43	0.917	0.932
Taken a step back from my work to judge my overall understanding	0.90	0.945	0.947
Planned ahead to take a relaxing break before a test or quiz	0.92	1.151	1.089
Double-checked my work before turning it in	-0.36	1.050	1.048

Sample size adjusted critical range for MNSQ statistics is 84-1.16. M = men, F = women, PEER = underrepresented minority, nPEER = Whites and Asians, reference category. The magnitude of significant DIF approaching .5 logits and greater is reported in logits beside the group that the DIF favors.

Supplemental Table 4. Baseline models with the full data set with CHEM 1A Grades (GPA Points) as the outcome of interest with standard errors below each regression coefficient.

	Baseline model with covariates (standard errors on the second line)	Baseline model with covariates and interactions (standard errors on the second line)

Reference Group: Not BIOME		
BIOME	0.185	0.229
	-0.046	-0.086
	p = 0.0001***	p = 0.009***
Admit Quarter		
Reference Group: 2017 Cohort		
2018 Cohort	-0.209	-0.211
	-0.043	-0.043
	p < 0.001***	p = 0.001***
2019 Cohort	-0.336	-0.335
	-0.045	-0.045
	p = 0.000***	p = 0.000***
Gender		
Reference Group: Female		
Male	0.122	0.12
	-0.036	-0.036
	p = 0.001***	p = 0.001***
SAT Math Score (divided by 100)		
	0.702	0.704
	-0.031	-0.031
	p = 0.000***	p = 0.000***
SAT Verbal Score (divided by 100)	-0.021	-0.02
	-0.048	-0.048
	p = 0.662	p = 0.676

Standardized SAT Writing Score	0.077	0.074
	-0.035	-0.035
	p = 0.026**	p = 0.033**
Ethnicity		
Reference Group: Caucasian		
Asian	-0.032	-0.012
	-0.043	-0.046
	p = 0.451	p = 0.801
International	0.258	0.209
	-0.118	-0.142
	p = 0.029**	p = 0.140
Unknown Ethnicity	-0.193	-0.401
	-0.243	-0.281
	p = 0.429	p = 0.154
Underrepresented Minority	-0.184	-0.182
	-0.046	-0.049
	p = 0.0001***	p = 0.0002***
High school gpa	0.794	0.792
	-0.067	-0.067
	p = 0.000***	p = 0.000***
Parent Education		
Reference Group: 2-Year College Graduate		
4-Year College Graduate	-0.021	-0.018

	-0.078	-0.078
	p = 0.788	p = 0.818
High School Graduate	-0.118	-0.116
	-0.081	-0.081
	p = 0.147	p = 0.155
Missing Information	-0.07	-0.074
	-0.184	-0.184
	p = 0.703	p = 0.689
No High School	-0.184	-0.184
	-0.097	-0.097
	p = 0.058*	p = 0.058*
Post Graduate Study	0.004	0.006
	-0.078	-0.078
	p = 0.958	p = 0.942
Some College	-0.163	-0.161
	-0.086	-0.086
	p = 0.057*	p = 0.061*
Some High School	-0.078	-0.075
	-0.096	-0.096
	p = 0.418	p = 0.437
Parent Income (log scale)	0.009	0.009
	-0.012	-0.012
	p = 0.469	p = 0.466
BIOME*Ethnicity Interactions		
BIOME*Asian		-0.138

		-0.117
		p = 0.239
Biome*International		0.105
		-0.241
		p = 0.664
BIOME*Unknown		0.805
		-0.562
		p = 0.153
BIOME*Underrepresented Minority		-0.024
		-0.116
		p = 0.836
Constant	-5.268	-5.285
	-0.436	-0.436
	p = 0.000***	p = 0.000***
Observations	2,613	2,613
R2	0.484	0.484
Adjusted R2	0.48	0.48
Residual Std. Error	0.830 (df = 2592)	0.830 (df = 2588)
F Statistic	121.321*** (df = 20; 2592)	101.308*** (df = 24; 2588)

BIOME = In BIOME, coded 1; Not in BIOME = Coded 0;
Parent Education variables - Highest degree attained of either student parent;
Parent Income (on log scale) - Parent income on the natural log (LN) scale. When parent income is listed as 0, as small value (.01) was added to keep it defined.
*p<0.1; **p<0.05; ***p<0.01

Supplementary Table 5. Non-Propensity score matched sample regression results of CHEM 1B on time course enrollment on BIOME

Variable	Model 1 On Time Course Taking (Baseline)			Model 2: On Time Course Taking (with Group Interactions)		
	Odds Ratio	Standard Error	p-value	Odds Ratio	Standard Error	p-value
BIOME						
Reference Group: Not BIOME	_	_	_	_	_	_
BIOME_1	1.92	0.15	<0.001	2.37	0.325	0.008
Admit Quarter						
Reference Group: 2017 Cohort	_	_	_	_	_	_
2018 Cohort	0.68	0.132	0.003	0.68	0.132	0.003
2019 Cohort	0.45	0.139	<0.001	0.45	0.139	<0.001
Gender						
Reference Group: Female	_	_	_	_	_	_
Male	1.16	0.112	0.2	1.16	0.112	0.2
Ethnicity						
Reference Group: Caucasian	_	_	_	_	_	_
Asian	1.16	0.14	0.3	1.17	0.147	0.3
International	1.27	0.457	0.6	1.14	0.507	0.8
Unknown	0.4	0.715	0.2	0.3	0.787	0.13
Underrepresented Minority	0.82	0.135	0.15	0.86	0.141	0.3
Parent Education						
2-YEAR COLLEGE GRADUATE	_	_	_	_	_	_
4-YEAR COLLEGE GRADUATE	1.25	0.225	0.3	1.24	0.225	0.3
HIGH SCHOOL GRADUATE	1.24	0.231	0.4	1.24	0.231	0.4
MISSING	1.57	0.603	0.5	1.57	0.604	0.5
NO HIGH SCHOOL	0.97	0.267	>0.9	0.98	0.267	>0.9
POST GRADUATE STUDY	0.89	0.227	0.6	0.89	0.228	0.6
SOME COLLEGE	1.19	0.245	0.5	1.18	0.245	0.5
SOME HIGH SCHOOL	1.05	0.268	0.9	1.06	0.268	0.8
SAT Math Score (divided by 100)	2.93	0.096	<0.001	2.93	0.096	<0.001
SAT Verbal Score (divided by 100)	0.99	0.148	>0.9	1	0.148	>0.9
Standardized SAT Writing Score	1.2	0.104	0.076	1.2	0.104	0.086
High School GPA	4.4	0.203	<0.001	4.35	0.203	<0.001
Parent Income (on log scale)	1.01	0.028	0.6	1.01	0.028	0.6
Interaction: BIOME * Ethnicity						
BIOME * Asian	_	_	_	0.93	0.446	0.9

BIOME * International	_	_	_	1.42	1.19	0.8
BIOME * Unknown	_	_	_	162,259	300	>0.9
BIOME * Underrepresented Minority	_	_	_	0.65	0.388	0.3

Ethnicity Unknown - Students with explicitly listed unknown ethnicities
Parent Education variables - Highest degree attained of either student parent;
Parent Income (on log scale) - Parent income on the natural log (LN) scale. When parent income is listed as 0, as small value (.01) was added to keep it defined.

Section 2: Qualitative and Rasch Methodology and Results

Qualitative Assessment of BIOME

Employing the survey-development variant of an exploratory sequential design (Cresswell & Plano Clark, 2011), a construct for measurement was selected during the qualitative portion of the research project so that an instrument could be developed. Among several candidates for measurement, academic behavior complexity was selected as a focal construct on the basis of both its interest to students and its plausible connection to program outcomes of promoting student academic success in *General Chemistry*. Observation statements from field notes were selected as sources for items exemplifying academic behavior complexity and then refined through cognitive interviews. This approach allows for full traceability between the final content of items and the concrete particular situations in which they are actually observed. Items are shown to reference real behaviors and their wording is made more authentic by attention to the form and context of how students talk about them.

Ethnographic classroom observation revealed that student mentees tended to respond to BIOME by increasing the complexity of their academic habits. Students were observed seeking advice about how to increase the number and types of academic habits in which they took part, and verbally reporting back about their efforts to program and research staff. While traditionally, qualitative evidence of this kind has not been treated as evidence that a causal process is underway, qualitative researchers have compellingly argued that such evidence should be at least as persuasive as statistical evidence (Maxwell, 2004). Qualitative work enabled us to critically reflect on our generative approach; this ensured our quantitative approaches were measuring what we intended to and that we can rely on our self-report survey instrument to measure academic habit complexity. For instance, focus group discussions enabled us to understand if students were merely responding in a way that they knew would make the program head happy (a process termed “satisficing,” see Barge & Gehlbach, 2012), while also reaffirming the relevance of the survey items and the reasons students are responding the way they are.

Focus groups and cognitive interviews provided additional evidence about the extent of the domain of this construct. In one focus group design, for example, students were presented with a randomly chosen list of half of the items intended for inclusion on our measure of habit complexity and prompted to write "other things that could go on this list." The results of this activity demonstrated that the topics of masked items were reintroduced to the list by students and that students were able to conceptualize academic behaviors as a coherent domain. A 22-item survey instrument was authored using this process (Supplemental Figure 2A). Analysis of survey item difficulty and student ability is presented as a Wright Map (Supplemental Figure 2B), while an in-depth discussion of the process of fitting the Rasch model to the data and validating the scale is available (Clairmont, 2020).

Constructing the Academic Habit Complexity Instrument

To initiate characterization of the impacts of our program, we sought to identify putative mechanisms that could potentially explain how the BIOME course influenced the academic success of first-year biology student mentees. To do so, an evaluator external to the program was recruited to qualitatively assess BIOME documents and materials, while conducting in-person observations of the program in action (conducted by A.C.). The evaluator recorded

ethnographic field notes for the duration of the 10-week course, supplemented by six focus groups (comprising ~six students each) of BIOME student mentees conducted near the beginning, middle, and end of the course. Using a goal-free evaluation approach (Scriven, 1973), several potential constructs for further inquiry were nominated for study. The research team evaluated each construct according to a set of transparent selection criteria that included assessing the depth of evidence that each construct played in the active part of student talk in focus groups and behaviors observed in the BIOME classroom.

Once construct selection was complete, items composing the instrument were authored by re-analyzing ethnographic field notes to locate talk and behavior pertaining to academic habits. These observations were rendered into first-person statements about behavior in the last two weeks, for example, “In the past two weeks, I have: Double-checked my work before turning it in.” Cognitive interviews were conducted with members of the target population (n = 27) to determine whether the construct was understood as intended, and no participants exhibited difficulty in the comprehension phase of the response process (Tourangeau et al., 2000). The resulting instrument is a 22-item checklist which participants completed by selecting the statements that applied to them. The construction of the instrument is described in a high level of detail in Clairmont, 2020. Ethnographic observations and focus group discussions enabled us to generate several indicators, or items for a survey instrument of academic habit complexity (Supplemental Figure 1A). Student total scores on this instrument were used in later analyses.

Rasch Analysis of Student Academic Habit Complexity

For assessment of whether BIOME influenced mentee academic habits differently than their non-enrolled peers, all first-year biology students were invited by email to complete the online 22-item academic habit complexity survey instrument (Supplemental Figure 1A) distributed by UCSB Institutional Research, Planning, and Assessment at the beginning (within the first two weeks) and end (last week) of the academic quarter. The survey instrument was deployed in both 2018 and 2019 cohorts gathering 392 and 346 individual responses, respectively. Entry into a random draw for small monetary gift cards was offered to all participants who completed both surveys. Those who completed the survey were given an option to opt out of their responses being used in this study. Those who selected to opt out of the belonging survey were removed from the data set but were still eligible for the gift card draw. Those participants who remained were anonymized through the removal of all identifiers by Institutional Research, Planning, and Assessment before analysis.

As we wanted to ensure that we could use the student scores derived from the student responses to the survey in further analyses to make invariant comparisons among student groups, we turned to the Rasch model for measure validation (Rasch 1960). Survey instruments that will be used to approximate interval-level measurement as we would like should be composed of items that are primarily sensitive to the construct of interest. In other words, we want items that are not influenced by other properties. We want survey items that are linearly related to one another and form a unidimensional construct and capture the range of the construct in in the focal population. To investigate these hypotheses for the academic habit complexity scale, a Rasch model was selected to be tested against the data (Bond & Fox, 2015;

Boone, 2016). The classic Rasch model -- a one-parameter logistic item response theory model -- is typically fit to data derived from dichotomous items (Rasch, 1960; Bond & Fox, 2015). Each item targets a location along the continuum of the hypothesized construct, e.g. low, medium, and high levels of the construct.

In Rasch measurement theory, items are sought out that meet the standards of the model. Questionable research practices such as the arbitrary deletion of items to improve the fit of the model or relaxing the assumptions of the model are carefully avoided. The suitability of the data to use for measurement is evaluated by the extent to which items fit the model (for instance, items that are hard are not answered correctly by students of lower ability more than students of higher ability), the extent to which persons fit the model, and the match between the range and distribution of item severity and person ability. We assessed the Rasch model fit, calculated estimates of person ability and item difficulty.
We then focused on item fit criteria (mean square fit statistics) and item coverage – that is, we checked to see whether item difficulties would be informative across the range of the sample ability distributions. We ensured that all infit values, commonly used in Rasch modeling, were between .8 and 1.2 (for a broader discussion – see Bond & Fox, 2015). Person-item targeting was visually inspected using a Wright Map (Supplemental Figure 1B). Next, person-separation reliability, similar to Cronbach's alpha, was determined to check adequacy. This index effectively communicates how well the instrument can be used to detect differences among persons, with higher values corresponding to increased reliability. Last, we ensured that item difficulties were consistent with the hypotheses embodied in the construct map (Clairmont, 2020).

Rasch Analysis of Student Academic Habit Complexity

To answer whether the academic habit complexity survey instrument accurately characterizes student academic habits, all first-year biology students were invited by email to complete the online academic habit complexity survey scale. Although the response rate was ~35-40% of the first-year biology cohorts, there were no significant demographic differences between BIOME and non-BIOME respondents. Rasch analysis of the academic habit complexity scale revealed that all items fit the hypothesized unidimensional Rasch model on the first attempt, the scale was found to target the range of persons adequately while also showing good person-separation reliability (Supplemental Figure 1B above; Clairmont, 2020). Mean-squared infit statistics were computed, and all items were within canonically acceptable ranges (0.876-1.11; Supplemental Table 3; Bond & Fox, 2015). The model explained 32% of the overall variance (Linacre, 2003). According to Reckase (1979), Rasch models explaining at least 20% of the variance are adequate for stable item difficulty and person ability estimates. Dimensionality was further investigated using a PCA of standardized residuals (Linacre, 1998) and comparison with a plausible multidimensional model using the BIC criterion - the results of both procedures suggesting that a unidimensional model is appropriate. Person-reliability of .77 indicates that the scale is sensitive enough to identify multiple, qualitatively distinct level of the construct among participants (Bond & Fox, 2015). Further, as hypothesized during the ethnography and focus group discussions, the difficulty range of items (SD = 1.21 logits) was sufficient to capture a

wide range of person ability (SD = .98 logits) and the mean item difficulty (.26 logits) was quite close to mean person ability (.05 logits). These comparisons suggest that the survey instrument was well-targeted to the population.

Propensity Score Matching Strategy

Propensity scores represent the probability that a student opts into BIOME based on a set of observed covariates. These propensity scores are then used for matching students from the treatment group (those in BIOME) to students from the control group (those not in BIOME) in order to correct for student self-selection bias (Rosenbaum & Rubin, 1985). Estimated propensity scores, on a probability scale (though, one could also use logits) are then used for matching (from 0 to 1). This process was implemented via the MatchITpackage in R (Ho *et al*., 2007, 2011; Thoemmes, 2011; using RStudio: Integrated development environment for R (Version 1.3.959) (www.rstudio.com/). The propensity score generating model, a logistic regression, included the background demographic variables found in the baseline model above and were provided by UCSB Institutional Research, Planning, and Assessment. They were selected based on their relevance to student selection into BIOME from previous literature (see for example, Wilton et al., 2019). While there is some debate about which variables to choose for propensity score models, it is commonly held that the best options involve selecting variables that are theoretically informed (see for instance, Leite, 2016; Stuart, 2010) but, for instance, it may also be worth considering numerous interactions (Gelman, Hill, & Vehatari, 2020). Numerous variables and higher order terms were included in our propensity score generating model. The success of matching is judged by the extent to which the matched treatment and control propensity scores overlap. To check the success of a matching procedure we assessed the absolute standardized difference in mean propensity scores between treatment and control, visualized the distribution of propensity scores for treatment and control units of each variable in the model, and checked whether the variance of the distributions are similar between treatment (BIOME) and control (non-BIOME) groups (Leite, 2016; Rosenbaum and Rubin, 1985). Standardized mean differences between treatment and control group for each covariate of interest of within .25 standardized deviations have been deemed acceptable by the *What Works Clearinghouse Procedures and Standards Handbook* (U.S. Department of Education, Institute of Education Sciences, & What Works Clearinghouse, 2013)

An important caveat of using the matched data sample is that we are not estimating the average treatment effect (ATE) but the average treatment on the treated (ATT); in other words, we are estimating the effect of BIOME on those who are likely to be in the program (see, for example, Morgan & Winship, 2015). Another important limitation of propensity score matching more generally is that propensity scores, and hence the matched sample, are only as good as the included predictors of the treatment. Therefore, the propensity score may be affected by any unmeasured variables or unincluded interactions or higher order terms.

To compare final CHEM 1A grade differences between BIOME students and non-BIOME students as well as generate the propensity score-matched populations by estimating the ATT, we built a multilevel linear regression model using cohort year as a random intercept to allow for

variation in the student population across years (Theobald, 2018). However, since the interest is in the treatment effect, the emphasis was on the interpretation of the coefficient representing the treatment effect of BIOME. From this perspective, both the multilevel model and the single-level model yielded similar results.

References:

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories. Cambridge University Press, Cambridge.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*, *2*, 266-283.

Linacre, J. M. (2003). Data variance: Explained, modeled and empirical. *Rasch Measurement Transactions*, *17*(3), 942-943.

Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference. Cambridge University Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of educational statistics, 4(3), 207-230.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.

Stuart E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review. *Journal of the Institute of Mathematical Statistics*, 25(1), 1–21. https://doi.org/10.1214/09-STS313

Thoemmes, F. J. & Kim E.S. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences, *Multivariate Behavioral Research*, 46:1, 90-118, DOI: 10.1080/00273171.2011.540475

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

Wilton, M., Gonzalez-Nino, E., McPartlan, P., Terner, Z., Christoffersen, R.E., & Rothman, J.H. (2019). Improving Academic Performance, Belonging, and Retention through Increasing Structure of an Introductory Biology Course. CBE Life Sciences Education, 18(4).