Supplemental Material

CBE—Life Sciences Education

Eddy *et al.*

Supplementary Materials:

A. The Classes and Student Population: Supplemental Table

Suppl Table 1. Pearson's correlation coefficients between Total Exam Points Earned and other measures of student competency.

	Term A	Term B	Term C
Exam 1	0.80	0.85	0.81
Exam 2	0.86	0.87	0.81
Exam 3	0.89	0.81	0.81
Exam 4	0.84	0.78	0.83
Cumulative College GPA	NA	0.73	0.72
Grade in Prior Biology Class	NA	0.73	0.79

B. Study 1 Supplemental Tables: What roles do students prefer to play in peer discussions?

A combination of grounded theory and content analysis was used to code students' open-ended responses to the question "What role do you prefer to play in peer discussions?" (Glaser and Strauss, 2009; Strauss and Corbin, 1990). We identified four categories: leader, collaborator, listener, and recorder and sample quotes for each of these is in Suppl Table 1. Two independent reviewers coded the responses and came to consensus when they disagreed.

Suppl Table 2. Themes identified from the open-ended question 'What role do you prefer to play in peer discussions?"

Theme	Percentage	Sample quotes	
Collaborator	44.1%	I like being able to help explain answers and I also enjoy learning from others	Listening and contributing
Leader	27.2%%	Generally, in most situations I obtain the role of the leader, but this is dependent upon my knowledge of the material.	Ruler. (Just kidding, an all round role I would say, asking questions about things I am not sure of, explaining my reasoning on things and asking others besides my group on why they put an answer)
Listener	11.1%	listening	to listen to others'

			input
Recorder	5.0%	Working on worksheets in class, I learn best if I am in charge of writing the answers.	writer

Suppl Table 3. Preferred Roles in Groupwork. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student characteristics on the roles they play in peer discussions. Friend = Friend in Group (Y/N); Ethn = Race/Ethnicity/Nationality; RBC = Relative Biology Competency

Rank	Model	AICc	Δi	ωi
1	Gender + Friend + Ethn	798.32	0	0.29
2	Gender + Friend	798.82	0.50	0.22
3	Gender + Ethn	799.91	1.59	0.13
4	Gender + Friend + RBC	800.57	2.25	0.09
5	Gender	801.27	2.95	0.07
6	Gender + Friend + Ethn + RBC	802.64	4.32	0.03
7	Gender + RBC	802.71	4.39	0.03
8	Gender + Friend + Ethn + Gender*Friend	802.83	4.51	0.03
9	Gender + Friend + Gender*Friend	803.25	4.93	0.02
10	Gender + Ethn+ RBC	803.75	5.43	0.02
11	Gender + Friend + Ethn + Gender*Ethn	804.02	5.43	0.02

Supp Table 4. Model-averaged multinomial regression coefficients Note: The reference level is always listed first in the comparison. A negative number indicates the student is more likely to prefer the role that is the reference level. A positive number indicates the student is more likely to prefer the role that is the comparison. **** $p \leq 0.0001$, *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, † $p \leq 0.01$

Comparisons:	**Intercept** β±SE (p-value)	**Gender: Female (ref: Male)** β±SE (p-value)	**Friend in Group:Yes (ref: No)** β±SE (p-value)	**Race/Ethnicity/Nationality:** **Ethn: Asian (ref:White)** β±SE (p-value)	**Ethn: International (ref: White)** β±SE (p-value)	**Ethn: Underserved (ref: White)** β±SE (p-value)	**Relative Biology Competency** β±SE (p-value)	**Term** β±SE (p-value)
Leader vs. Listener	-1.73±0.765 (0.023)*	1.27±0.460 (0.005)**	-0.82±0.461 (0.076)†	1.20±0.579 (0.037)*	2.01±0.800 (0.012)*	1.52±0.806 (0.060)†	-0.49±0.22 (0.0270)*	0.44±0.47 (0.337)
Leader vs. Collaborator	-0.28±0.337 (0.401)	1.37±0.301 (0.0001)****	0.09±0.349 (0.800)	0.07±0.317 (0.817)	-0.52±0.698 (0.460)	-0.63±0.625 (0.309)	-0.14±0.158 (0.352)	0.08±0.340 (0.821)
Leader vs. Recorder	-3.24±0.933 (0.003)**	2.42±0.799 (0.0025)**	0.41±0.723 (0.571)	-0.94±0.668 (0.159)	-13.31±362.829 (0.971)	-0.24±0.934 (0.801)	0.18±0.321 (0.570)	-1.5±1.08 (0.162)
Collaborator vs. Listener	-1.45±0.773 (0.061)†	-0.07 ± 0.439 (0.858)	-0.91 ± 0.424 (0.0322)*	1.13 ± 0.553 (0.041)*	2.53 ± 0.774 (0.0011)**	2.15 ± 0.767 (0.005)**	-0.35±0.203 (0.088)†	0.37±0.426 (0.3857)
Collaborator vs. Recorder	-2.96 ± 0.918 (0.0013)**	1.04 ± 0.788 (0.184)	0.32 ± 0.692 (0.643)	-1.01 ± 0.637 (0.111)	-12.8 ± 658.137 (0.984)	0.40 ± 0.889 (0.653)	0.33±0.305 (0.280)	-1.59±1.06 (0.135)
Listener vs. Recorder	-1.51 ± 1.27 (0.233)	1.12 ± 0.864 (0.192)	1.23 ± 0.766 (0.109)	-2.15 ± 0.807 (0.0078)**	-14.8 ± 493.706 (0.976)	-1.8 ± 1.03 (0.090)†	0.68±0.347 (0.0512)†	-1.96±1.111 (0.078)†
Leader vs. Other	-0.58 ± 0.515 (0.256)	0.85 ± 0.416 (0.039)*	-0.92 ± 0.432 (0.032)*	-0.72 ± 0.452 (0.112)	-1.27 ± 1.141 (0.265)	-0.37 ± 0.452 (0.113)	-0.04±0.219 (0.861)	0.18±0.459 (0.702)
Collaborator vs. Other	0.30 ± 0.523 (0.564)	-0.52 ± 0.396 (0.194)	-1.01 ± 0.398 (0.011)*	-0.79 ± 0.424 (0.063)†	-0.76± 1.134 (0.5043)	0.26 ± 0.752 (0.728)	0.12±0.203 (0.280)	0.10±0.427 (0.8165)
Listener vs. Other	1.15 ± 0.857 (0.181)	-0.44 ± 0.522 (0.403)	-0.11 ± 0.507 (0.835)	-1.92 ± 0.634 (0.024)*	-3.28 ± 1.192 (0.0059)**	-1.89 ± 0.894 (0.0344)*	0.455±0.465 (0.338)	-0.27±0.526 (0.0607)†
Recorder vs. Other	2.66 ± 1.02 (0.009)**	-1.56 ± 0.838 (0.062)†	-0.95 ± 0.869 (0.276)	0.11 ± 0.530 (0.829)	5.85 ± 5.793 (0.312)	-0.071 ± 0.733 (0.923)	-0.045±0.178 (0.802)	0.16±0.603 (0.789)

Relative Variable Importance	NA	1	0.71	0.51	0.20	0.10

C. *Study 2 Supplemental Tables*: Do students indicate that groupmates limit the ability of others to participate in peer discussions?

1. Codes from student responses to "What is the worst part of groupwork?"

Suppl. Table 5. Themes identified from the open-ended question 'What is the worst part of groupwork?"

Theme	Percentage	Sample quotes	
Lack of knowledge	24.1%	No one has an idea about what we're learning.	Very confusing questions that no one knows how to tackle.
Groupmates limiting participation of other groupmates	12.5%	sometime some of the members can be dominating	Just the fact that sometimes since we are on a time crunch, not everyone gets to share their ideas. Also, some people have more influence when they share their ideas that others for some reason.
Groupmate Deficit – blaming groupmates for not participating	12.6%	Some people haven't done the reading and try to back up an answer that makes absolutely no sense and do not provide any new insight.	It annoys when me someone is being shy. International students tend to be shy and withdrawn, or only want to interact with other international students. I don't think it's even a language barrier, I think a lot of the Asian students come from a culture where they're not socialized to be as loud and outgoing as we are in America, so they don't want to share their ideas even when they're very

			smart.
Conflicting ideas	9.9%	When every single member believes the answer is different from the other, and the explanations sound plausible.	too many opinions sometimes
Uncomfortable participating	7.1%	i get intimidating when someone ask question that I dont have answers for.	Not being able to contribute when I don't understand the material, not knowing if my group and I are correct/have the correct answers
Other	33.8%		

2. Model Selection Results for whether or not a student reports a groupmate limiting participation.

Suppl. Table 6. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student level characteristics on whether a student reported a groupmate limiting other's participation as the worst part of groupwork.

Rank	Model	AICc	Δi	ωi
1	RBC	209.63	0.0	0.17
2	RBC + Gender	210.03	0.40	0.14
3	RBC + Term	211.23	1.60	0.08
4	RBC + Friend	211.60	1.97	0.06
5	RBC + Gender + Term	211.74	2.11	0.06
6	RBC + Gender + Friend	212.05	2.42	0.05
7	Gender	212.19	2.56	0.05
8	(NULL Model)	212.47	2.84	0.04
9	RBC + Ethn	212.79	3.17	0.03
10	RBC + Friend + Term	213.26	3.63	0.03
11	RBC + Gender + Ethn	213.44	3.81	0.02
12	RBC + Gender + Friend + Term	213.80	4.17	0.02
13	RBC + Gender + Friend + Gender x Friend	213.84	4.22	0.02
14	Gender + Term	213.93	4.30	0.02
15	Term	214.17	4.54	0.02
16	Gender + Friend	214.17	4.54	0.02
17	RBC + Ethn + Term	214.38	4.76	0.02
18	Friend	214.40	4.77	0.02
19	RBC + Friend + Ethn	214.55	4.92	0.01
20	Ethn	214.68	5.06	0.01
21	Gender + Ethn	214.76	5.14	0.01
22	RBC + Gender + Ethn + Term	215.17	5.54	0.01
23	RBC + Gender + Friend + Ethn	215.29	5.66	0.01

24	RBC + Gender + Friend + Term + Gender x Friend	215.58	5.95	0.01
25	Gender + Friend + Term	215.96	6.34	0.01
26	Gender + Friend + Gender x Friend	216.04	6.42	0.01
27	Friend + Term	216.16	6.53	0.01

3. Model selection results for Likert Scale 'Dominator in Group' Question.

Suppl. Table 7. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student level characteristics on whether a student reports a dominator in their group.

Rank	Model	AICc	Δi	ωi
1	RBC + Ethn + Term	593.17	0.0	0.39
2	RBC + Ethn + Term + Friend	594.40	1.23	0.21
3	RBC + Ethn + Term + Gender	595.03	1.86	0.16
4	RBC + Ethn + Term + Friend + Gender	596.31	3.14	0.08
5	RBC + Ethn + Term + Friend + Gender + Gender x Friend	598.42	5.25	0.03
6	RBC + Ethn + Term + Friend + Ethn x Friend	599.44	6.27	0.02
7	RBC + Ethn + Term + Gender + Gender x Ethn	599.73	6.56	0.01
8	RBC + Term	599.76	6.59	0.01
9	RBC + Ethn	599.89	6.72	0.01
10	RBC + Ethn + Friend	600.02	6.85	0.01
11	RBC + Term + Friend	600.84	7.67	0.01
12	RBC + Ethn + Term + Friend + Gender + Ethn x Friend	601.26	8.09	0.01

Suppl. Table 8. Model averaged coefficients for whether a student reports Lack of Access issues in their group and a dominator in their group. ǂ = p-value < 0.1, * <0.05, **<0.01, ***<0.001.Dominator in Group: The more negative the response the more likely the student is to feel there is a dominator.

Parameters	Lack of Access		Dominator in Group	
	Relative Variable Importance	Estimate ± Adjusted SE	Relative Variable Importance	Estimate ± Adjusted SE
Intercept	NA	-2.19 ± 0.423***	NA	*Multiple intercepts*
Term: *(reference level: Class A)* Class B	0.30	 0.24 ± 0.399	0.96	 -0.76 ± 0.264**
Relative Biology Competency:	0.76	-0.38 ± 0.182*	1	0.51 ± 0.125**
Gender: *(reference level: Male)* Female	0.48	 0.50 ± 0.467	0.32	 -0.13 ± 0.293
Race/Ethnicity/Nationality: *(reference level: White American)* Asian American Underserved American International	0.19	 0.61 ± 0.444 0.76 ± 0.730 -0.41 ± 1.116	0.97	 -0.52 ± 0.288ǂ 0.35 ± 0.523 -1.48 ± 0.499**
Friend: (*reference level: No friend in group)* Friend in group	0.31	 -0.15 ± 0.500	0.39	 0.24 ± 0.308
Race/Ethnicity/Nationality x Friend: *(reference level: White American and No friend in group)* Asian American x Friend	0.01	 -0.34 ± 0.928	0.03	 0.58 ± 0.580

Underserved American x Friend International x Friend		-1.80 ± 1.40 -14.5 ± 902.8		-0.12 ± 1.169 0.719 ± 0.947
Gender x Friend: *(reference level: Male and No friend in group)* Female x Friend	0.05	 0.46 ± 0.890	0.04	 -0.10 ± 0.536
Gender x Race/Ethnicity/Nationality *(reference level: Male and White American)* Female x Asian American Female x Underserved American Female x International	NA	NA	0.03	 0.314 ± 0.530 -0.48 ± 1.081 0.97± 0.947

D. Study 3 Supplemental Tables and Methods: How comfortable are students with participating in group work and is this comfort any greater than their comfort answering instructor posed questions in front of the whole class?

1. Survey Items and Reliability

The survey used to address this question is given below. The responses to the questions were a likert scale with 4 possible responses: *Strongly Disagree, Disagree, Agree, or Strongly Agree.*

Suppl. Table 9. Survey Comparing Experience with Small and Large Group Work

Prompt 1: *Today you worked in a small group with other students on in-class problems (clicker questions or discussion questions). Please refer to this group when you answer the following questions.*	
Peer Discussion Questions:	**Factor from Original Paper**
I feel like I belong in my group.	Comfort being oneself in group
I feel like it's okay to make mistakes in front of others in my group.	Comfort being oneself in group
I feel like it's okay to ask 'dumb' questions in front of my group.	Comfort being oneself in group
I feel comfortable offering my own ideas in my group.	Comfort being oneself in group
I feel different from other students in my group.	Social Comparison Concern
I often feel intimidated to participate in my group.	Social Comparison Concern
I worry about being wrong when working in my group.	Social Comparison Concern
I have generally understood the material as well as the other people in my group.	Social Comparison Concern
I often leave class feeling like I'm not as smart as the other students in my group	Social Comparison Concern
I often leave the class feeling like I'm the only one in my group who doesn't understand the material.	Social Comparison Concern
Prompt 2: *At times you were also asked/encourage by the instructor to offer ideas, ask questions or give answers in front of the whole class. Please refer to this experience when answering the following questions.*	
Large Group Questions:	
I feel like I belong in this class	Comfort being oneself in group
I feel like it's okay to make mistakes in front of the whole class	Comfort being oneself in group

I feel like it's okay to ask 'dumb' questions in front of the whole class.	Comfort being oneself in group
I feel comfortable offering my own ideas in front of the whole class.	Comfort being oneself in group
I feel different from the other students in this class.	Social Comparison Concern
I often feel intimidated to talk in front of the whole class.	Social Comparison Concern
I worry about being wrong in front of the whole class.	Social Comparison Concern
I have generally understood the materials we well as other people in this class	Social Comparison Concern
I often leave the class feeling like I am not as smart as other students in the class.	Social Comparison Concern
I often leave class feeling like I am the only one who doesn't understand the material.	Social Comparison Concern

Reliability Analyses for Study 3:

Differential Item Functioning (DIF) Analyses for Study 3:

Researchers, specifically measurement specialists, are concerned with the fairness of the scales, and in particular the possibility that developed scales may be biased against male or female groups. For example, equal response levels of an item are expected for examinees who are matched in their observed total scores which reflect the trait being measured by the scale. In other words, the probability of answering an item correctly or of attaining a particular response level is modeled as a function of an individual's ability or latent trait. Unequal item correct response rates or item response levels between score-matched males and females indicates the potentially biased items that may favor males or females. Differential items functioning (DIF) is developed to detect scale item which function differently between groups of examinees matched in their scores measured by scales. An item showing differential item correct response rates or levels between score-matched males and females is identified as DIF items, especially referred to as gender DIF items as males and females are matched.

An item identified with DIF may be a *biased* item, but it is not necessarily the case. As Zumbo (1999) pointed out, biased items refer to the condition where examinees respond to scale items differently because they are unrelated to traits or constructs the scale is developed to measure. Thus, DIF is required, but not sufficient to claim that an item is biased, until the DIF for the item is proven to be unrelated to what the scale is developed to measure (Zumbo, 1999). If DIF items are proven to be biased after a substantial investigation following the identification of DIF items, scale scores should be adjusted to correct for the resulting DIF effect in the scale scores.

Concern about DIF items has led measurement professionals to develop various DIF detecting methods for investigating such occurrences. DIF analyses are typically conducted at the individual item level. It is assumed that the absence of DIF items will lead to an unbiased scale.

Methods:

The simultaneous Item Bias Test (SIBTest), developed by (Shealy and Stout, 1993a, 1993b), is a statistical method to detects items with DIF implemented by a computer software program called DIFPACK. In this study, SIBTest was selected based on the following reasons. The SIBTest DIF method was selected in my DIF study is based on the following reasons. First, the SIBTest has been found to be more effective in detecting DIF than the Mantel-Haenszel and logistic regression DIF methods (Bolt and Stout, 1996; Jiang and Stout, 1998). Second, the SIBTest uses non-parametric approach to design its DIF detection model, which is different from those DIF methods developed by the parametric approach. The parameter-oriented DIF method requires stronger assumptions that a dataset may fail to meet its assumption. Third, after items identified as DIF, they can be grouped as a bundle to examine the potential sources that contribute to the DIF effect. Fourth, instead of using observed test score, the SIBTest runs a regression to estimate the true score used to match students on ability, which results in an improved estimation of ability level where the examinees should stay. Because the collected data includes both dichotomously and polytomously scored items, the Poly-SIBTest option in the DIFPACK v1.7 software application was selected to perform gender DIF item analysis. To identify items as DIF status, a nominal α equal to .01, which is commonly used, was selected in the hypothesis testing to control Type I error rate. In such a circumstance, 1% of the items would be falsely identified as DIF when they did not truly show DIF. When items identified as DIF, the guidelines developed by Roussos and Stout (1996) can be used to evaluate magnitude of the DIF as negligible, moderate or large effect. Student records with missing data were eliminated from the data to meet the requirement of Poly-SIBTest DIF method. The number of students involved in the gender DIF analysis was smaller than the collected ones.

Results and Conclusions:

For other 10 items administered to the participants in peer discussions, there were 279 males and 388 females involved in the DIF analysis. Three items including, Q1, Q2, and SG4, were flagged with DIF (B = .116, -.239, and .117, $p < .01$). Items Q1 (*I feel comfortable offering my own ideas in my group*) and SG4 (*I often leave class feeling like I'm not as smart as the other students in my group*) favored males whereas item Q2 (*I feel different from other students in my group*) favored females. For the identical 10 items administered to the participants in large class discussion, there were 275 males and 391 females involved in the DIF analysis. The DIF analysis showed there is only one item Q4 (*I feel different from the other students in this class.*) identified as DIF favoring females (B = -.225, $p < .01$). All five identified gender DIF items were classified as large magnitude of DIF effects as the estimated betas were larger than .08.

Four out of 31 items were identified as DIF status which indicate unequal response levels between males and females matched in their observed survey scores. Consistent with the previous studies, the finding supports that the gender DIF items do occur in a survey. However, as pointed out, they do not count as biased items unless the source of DIF contributing to unequal possibilities can be proved irrelevant to the ability or trait a scale developed to measure. The identified gender DIF items may place a threat to the validity of the scale. Thus, further comprehensive investigation aiming at flagged DIF items is suggested to ensure the validity of the items before those items are allowed to use to reflect what the scale is designed to measure.

2. Model Selection Results for Social Comparison Concern and Comfort Being Yourself factors

Suppl. Table 10. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student characteristics on Factor 1: Comfort Being Oneself. Models also have a random effect (1|Stu.ID)

Rank	Model	AICc	Δi	ωi
1	Gender + RBC + Context + Gender x Context	5428.8	0.0	0.36
2	Gender + RBC + Context + Ethn + Gender x Ethn + Gender x Context + Ethn x Context	5429.1	0.31	0.31
3	Gender + RBC + Context + Ethn + Gender x Context + Ethn x Context	5429.9	1.04	0.21
4	Gender + RBC + Context + Ethn + Gender x Ethn + Gender x Context	5432.2	3.36	0.07

Suppl. Table 11. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student characteristics on Factor 2: Social Comparison Concern. Models also have a random effect (1|Stu.ID)

Rank	Model	AICc	Δi	ωi
1	Gender + Context + RBC + Ethn + Gender x Context + Ethn x Context + Gender x Ethn	5925.2	0.0	0.50
2	Gender + Context + RBC + Gender x Context	5926.6	1.44	0.24
3	Gender + Context + RBC + Ethn + Gender x Context + Ethn x Context	5927.7	2.52	0.14
4	Gender + Context + RBC + Ethn + Gender x Context + Gender x Ethn	5930.1	4.93	0.04
5	Gender + Context + RBC + Ethn + Ethn x Context + Gender x Ethn	5930.9	5.69	0.03

Suppl.Table 12. Students report more positively on both factors in peer discussions relative whole class discussions, especially women. Model averaged coefficients for Comfort and Social Comparison Concern. Models also have a random effect (1|Stu.ID). ŧ = p-value < 0.1, * <0.05, **<0.01, ***<0.001.

	Factor 1: Comfort Being Oneself		Factor 2: Social Comparison Concern	
Parameter	Relative Variable Importance	Model Averaged Coefficients	Relative Variable Importance	Model Averaged Coefficients
Intercept	NA	12.8 ± 0.194***	NA	17.1 ± 0.256***
Relative Biology Competency:	1	0.51 ± 0.068***	1	1.27 ± 0.099***
Gender:	1		0.98	
(reference level: Male)				
Female		-0.003 ± 0.254		0.018 ± 0.298
Ethn:	0.64		0.74	
(reference level: White American)				
Asian American		-0.13 ± 0.313		-0.08 ± 0.347
Underserved American		0.11 ± 0.379		-0.22 ± 0.574
International		-0.08 ± 0.542		-1.84 ± 0.751*
Gender x Ethn:	0.38		0.57	
(reference level: White and Male)				
Female x Asian American		-0.70 ± 0.293*		-0.44 ± 0.424
Female x Underserved American		0.03 ± 0.490		0.54 ± 0.710
Female x International		-0.44 ± 0.609		1.47 ± 0.886 ŧ
Participation Context	1		1	

(reference level: Peer Discussions) Whole Class		 -2.30 ± 0.245***		 -2.44 ± 0.270***
Gender x Participation Context: *(reference level: Male and Peer Discussions)* Female*Whole Class	1	 -1.11 ± 0.218***	0.93	 -0.68 ± 0.225**
Ethn x Participation Context: *(reference level: White and Peer Discussions)* Asian American x Whole Class Underserved American x Whole Class International x Whole Class	0.52	 0.43 ± 0.231ŧ 0.16 ± 0.485 1.44 ± 0.485**	0.68	 0.44 ± 0.239 ŧ 0.05 ± 0.502 1.62 ± 0.502**

E. Supplemental Analyses for Study 4: Value of Peer discussion

DIF Analyses Results:

Methods were the same as those applied to the survey in Study 3. It's important to note that for this data set, the number of students dramatically decreased because of the missing data elimination procedure.

The gender DIF analysis is summarized as follows. For 11 items only administered to the participants in peer discussion, there were 112 males and 159 females involved in the DIF analysis. As a result, only one item, Q22 (*After discussing a clicker question in my group, I am more likely to answer it correctly than if I had worked by myself*), was flagged with DIF (B = -.233, p < .01). The item favored females indicating higher response level than that of males.

Factor Analysis:

Suppl. Table 13. Factor Loadings for Groupwork survey. An asterisk (*) after a loading indicates that the question belongs to that factor.

Questions:	Factor 1 Loadings: Group Function	Factor 2 Loadings: Comfort and Confidence with Contribution
My group worked well together.	0.980*	0.021
My group members made significant contributions of knowledge and/or ideas to the group.	0.977*	-.001
I made significant contributions of knowledge and/or ideas to my group.	0.977*	0.024
There was one (or more) person in my group who dominated most of the discussion.	0.976*	0.021
There was one (or more) person in my group who dominated most of the discussion.	0.971*	0.015
I feel like I can be myself in my group.	-0.065	0.841*
I feel like I belong in my group.	-0.065	0.821*
I understand the material as well as other students in my group.	-0.095	.790*
I feel comfortable offering my own ideas in my group.	0.225	0.778*
I often feel intimidated to participate in my group.	0.042	0.775*
I worry about being wrong when working in my group.	-0.020	.741*
I feel different from other students in my group.	0.097	0.684*

In addition, we asked students 5 questions related to the value they perceived in peer discussions. These questions were used as a third factor that was a response variable in our analysis:

- *Explaining the material to my group improved my understanding of it. If you did not get the opportunity to explain material, please leave blank.*
- *A group member explaining the material to me improved my understanding of it. If you did not get the opportunity to have material explained to you, please leave blank.*
- *After discussing a clicker question in my group, I am more likely to answer it correctly than if I had worked by myself.*
- *Listening to a lecture helps me understand a topic better than discussing it with other students in a group.*
- *Discussing a topic with other students in a group helps me understand a topic better than listening to a lecture.*

Finally we asked 2 questions intended to control for how familiar a student was with their groupmates. We ultimately chose to use the friend outside of class question as our control and we collapsed the Likert scale question to a binary with the No Friend condition including Strongly Disagree and Disagree and the Friend condition including Strongly Agree and Agree. The original questions were:

- *I am friends outside of class with at least one of the students in my group.*
- *I did not know the other students in my group before taking this class.*

Model Selection for Factor 1: Group Function and Factor 2: Comfort and Confidence with Group work

Seven potential variables were initially considered tocontribute to responses on the survey questions: 1) a student's overall performance in the course (BI.GPA); 2) student gender identity (a factor with twolevels; Stu.Gender); 3) student race/ethnicity/nationality (afactor with 4 levels; Ethn); 4) an interaction betweenstudent gender identity and race/ethnicity/nationality (Stu.Gender*Ethn); 5) whether not a student had a friend in the group (a factor with two levels; Friend); 6) an interactions between gender identity and friend (Stu.Gender*Friend); and 7) an interaction between race/ethnicity/nationality and friend (Ethn*Friend).Only students with a complete setof these variables were included in this analysis.

Combinations of these 7 variables produced a total of 72 potential models to describe our data. The total number of models tested was substantially lower than our number of observations (n=360 students), which justified fully exploring this set of models. Thus, we systematically explored the possible models for our data and ultimately chose the model that best fits the data according to the model-selection statistics. We also calculated the model averaged regression coefficients for the fixed effects in our model. Our initial full model was as follows:

Factor =BI.GPA + Stu.Gender + Ethn + Stu.Gender∗Ethn + Friend + Stu.Gender∗Friend + Ethn*Friend

Factor 1: Group Function. Model selection identified 9 models that had the strongest support ($\Delta i < 4$) for predicting student responses on this factor. The top three models had the majority of support (summed $\omega = 0.53$; Supp. Table 11). The top two models had almost equal support and differed only in whether

or not a the dichotomous variable indicating whether a student had a friend in the group was present. The best model included Friend as well as Exam Performance, Term, Race/Ethnicity/Nationality and explained 14% of the variation in student responses. This low R^2 along with the instability in terms present in the top model indicates our explanatory variables capture some of the variation in this factor, but are not sufficient for capturing the main causes of variation.

Specifically, exam performance, the iteration of the course, and Race/Ethnicity/Nationality all significantly predicted student responses on Factor 1 (Supp. Table 12). As student exam performance increased so did their sense that their group functioned well ($\beta = 0.135 \pm 0.190$). Relative to White American students Asian and International students felt their groups functioned less well ($\beta = -0.371 \pm 0.161$ and $\beta = -0.561 \pm 0.278$ respectively).

Suppl. Table 14. 95% confidence set of best ranked models (summed $\omega > 0.95$) examining the effect of student characteristics on Factor 1: Group Function.

Rank	Model	AICc	Δi	ωi
1	Term + BI.GPA + Friend + Ethn	695.44	0	0.21
2	Term + BI.GPA + Ethn	695.67	0.22	0.19
3	Term + BI.GPA + Gender + Ethn	696.39	0.94	0.13
4	Term + BI.GPA + Friend + Gender + Ethn	696.47	1.03	0.12
5	Term + BI.GPA + Friend + Gender + Ethn+ Friend*Gender	696.90	1.46	0.10
6	Term + BI.GPA	698.57	3.13	0.04
7	Term + BI.GPA + Friend	699.14	3.70	0.03
8	Term + BI.GPA + Ethn + Friend + Friend*Ethn	699.34	3.90	0.03
9	Term + BI.GPA + Gender	699.42	3.98	0.03
10	Term + BI.GPA + Friend + Gender	700.10	4.65	0.02
11	Term + BI.GPA + Friend + Gender +Ethn + Friend*Ethn	700.43	4.99	0.02
12	Term + BI.GPA + Friend +Gender + Friend*Gender	700.82	5.38	0.01
13	Term + BI.GPA + Friend +Gender + Ethn + Gender*Ethn	701.02	5.58	0.01
14	Term + BI.GPA + Friend + Gender + Ethn + Gender*Ethn	701.23	5.82	0.01
15	Term + BI.GPA + Friend + Gender + Ethn +Friend*Gender + Friend*Ethn	701.26	5.82	0.01
16	Term + BI.GPA + Friend + Gender + Ethn + Friend*Gender + Gender*Ethn	701.89	6.45	0.01
17	BI.GPA +Friend + Ethn	702.59	7.14	0.01
18	BI.GPA + Friend +Gender + Ethn	703.76	8.32	0.00
19	BI.GPA + Ethn	704.18	8.74	0.00
20	Term + BI.GPA + Gender + Ethn + Friend*Gender	704.30	8.86	0.00

Suppl. Table 15. Asian American and International students report more negatively on both factors whereas performance in the course and having a friend in the group leads to more positive responses. Relative Variable Importance and Model averaged coefficients for Group Function and Comfort and Confidence Factors for peer discussions. ɬ = p-value < 0.1, * <0.05, **<0.01, ***<0.001.

Parameter	Factor 1: Group Function		Factor 2: Comfort and Confidence	
	Relative Variable Importance	Model Averaged Coefficients	Relative Variable Importance	Model Averaged Coefficients
Intercept	NA	0.24 ± 0.180	NA	0.03 ± 0.172
BI.GPA	1	0.24 ± 0.064***	1	0.23 ± 0.063***
Friend:	0.60		0.99	

(reference level: no friend in group) Friend in Group		0.14 ± 0.190		0.44 ± 0.156**
Gender: (reference level: Male) Female	0.49	-0.21 ± 0.199	0.80	-0.24 ± 0.159
Term: (reference level: larger class) Smaller Class	0.98	-0.43 ± 0.136**	0.38	-0.14 ± 0.134
Ethn: (reference level: White American) Asian American Underserved American International	0.86	-0.37 ± 0.161* -0.18 ± 0.301 -0.56 ± 0.278*	0.99	-0.41 ± 0.141** -0.13 ± 0.274 -0.80 ± 0.271**
Gender x Ethn: (reference level: White American and Male) Female*Asian American Female*Underserved American Female*International	0.03	0.28 ± 0.261 -0.13 ± 0.564 0.38 ± 0.527	0.04	0.13 ± 0.256 0.12 ± 0.553 0.37 ± 0.519
Gender x Friend: (reference level: Male and no Friend) Female*Friend	0.14	0.37 ± 0.271	0.20	-0.04 ± 0.266
Ethn x Friend: (reference level: White and no Friend) Asian American*Friend Underserved American*Friend International*Friend	0.06	0.33 ± 0.292 0.59 ± 0.565 -0.17 ± 0.513	0.05	-0.06 ± 0.288 0.35 ± 0.557 -0.23 ± 0.508

Factor 2 – Comfort and Confidence with Contributions to Group. Model selection identified 6 models that predicted student responses on this factor. The top 2 models had the majority of the support (summed $\omega = 0.53$; Supp. Table 13). The top model is 1.65 times more likely to be the best model than the second best model. The best model includes Friend, Race/Ethnicity/Nationality, Exam Performance and Gender. The second best model includes the additional variable Term. The best model has a low R^2 (17%) indicating that our explanatory variables were not able to capture the majority of the variation in this factor.

Across all the potential models, Exam Performance, Friend, and Race/Ethnicity/Nationality were the only variables that significantly explained student responses on this factor (Supp. Table 12). Both having a friend in the group and performing better on exams caused students to answer more positively on this factor. Asian American and international students were less likely to respond as positively on this factor as White American students.

Suppl. Table 16. 95% confidence set of best ranked models (summed $\omega > 0.95$) examining the effect of student characteristics on Factor 2: *Comfort and Confidence with Contributions to Group*.

Rank	Model	AICc	Δi	ωi
1	Friend + Ethn + BI.GPA + Gender	686.78	0	0.33
2	Friend + Ethn + BI.GPA + Gender + Term	687.76	0.98	0.20
3	Friend + Ethn + BI.GPA	688.89	2.10	0.12
4	Friend + Ethn + BI.GPA + Gender + Friend*Gender	688.91	2.12	0.12
5	Friend + Ethn + BI.GPA + Gender + Term + Friend*Gender	689.90	3.12	0.07
6	Friend + Ethn + BI.GPA + Term	690.01	3.23	0.07

Suppl. Table 17.Value of Group work. 95% confidence set of best ranked models (summed ω> 0.95) examining the effect of student characteristics on Factor 3: Value of Group Work. Function = Response on Group Function factor, Comfort = Response on the Comfort and Confidence with Participation factor.

Rank	Model	AICc	Δi	ωi
1	Function + Comfort + Friend + Gender + Term + RBC + Friend x Gender	1074.72	0	0.28
2	Comfort + Friend + Gender + Term + RBC + Friend x Gender	1075.56	0.84	0.19
3	Function + Comfort + Friend + Term + RBC	1077.13	2.41	0.08
4	Function + Comfort + Friend + Gender + Term + Ethn + RBC + Friend x Gender	1077.38	2.66	0.07
5	Function + Friend + Term + RBC	1077.48	2.75	0.07
6	Function + Friend + Gender + Term +Ethn + RBC + Friend x Gender	1078.11	3.39	0.05
7	Function + Comfort + Friend + Gender + RBC+ Gender x Friend	1078.40	3.68	0.04
8	Function + Comfort + Friend + Gender + Term + RBC	1079.07	4.35	0.03
9	Function + Friend + Gender + Term + RBC	1079.23	4.51	0.03
10	Function + Friend + Gender + RBC + Gender x Friend	1079.63	4.91	0.02
11	Function + Comfort + Friend + Term + RBC + Ethn	1079.90	5.17	0.02
12	Function + Friend + Term + RBC + Ethn	1080.25	5.53	0.02
13	Function + Comfort + Friend + RBC	1081.00	6.28	0.01
14	Function + Comfort + Friend + Gender + Ethn + RBC + Friend x Gender	1081.28	6.56	0.01
15	Function + Friend + RBC	1081.74	7.02	0.01
16	Function + Comfort + Friend + Gender + Term + RBC + Ethn + Gender x Friend	1081.75	7.03	0.01
17	Function + Friend + Gender + Term + RBC + Ethn + Gender x Friend	1081.86	7.14	0.01

Suppl. Table 18. Value of Group Work model averaged coefficients. ɬ = p-value < 0.1, * <0.05, **<0.01, ***<0.001.

Parameter	Relative Variable Importance	Model Averaged Coefficients	p-value
Intercept	NA	5.15 ± 1.23	< 0.001***
Factor 1: Group Function	1	0.51 ± 0.089	< 0.0001***
Factor 2: Comfort and Confidence with Participation	0.59	0.09 ± 0.054	0.098 ɬ
Relative Biology Competency	0.99	-0.50 ± 0.140	0.0004***
Friend:	0.99		
(reference level: no friend in group)			
Friend in Group		0.32 ± 0.580	0.578
Gender:	0.77		
(reference level: Male)			
Female		-1.05 ± 0.579	0.0684 ɬ
Term:	0.88		
(reference level: Class B)			
Class A		0.71 ± 0.293	0.015*
Race/Ethn/Nationality:	0.21		
(reference level: White American)			
Asian American		-0.31 ± 0.283	0.282
Underserved American		0.31 ± 0.497	0.889
International		0.50 ± 0.552	0.365
Gender x Friend:	0.68		
(reference level: Male and no Friend)			
Female*Friend		1.42 ± 0.248	0.014*

Work Cited:

Bolt, D., and Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBtest procedure. Behaviormetrika *23*, 67–95.

Glaser, B.G., and Strauss, A.L. (2009). The Discovery of Grounded Theory: Strategies for Qualitative Research (Transaction Publishers).

Jiang, H., and Stout, W. (1998). Improved Type I Error control and reduced estimation bias for DIF detection using SIBTEST. J. Educ. Behav. Stat. *23*, 291–322.

Roussos, L., and Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. Appl. Psychol. Meas. *20*, 355–371.

Shealy, R., and Stout, W.F. (1993a). An item response theory model for test bias. In Differential Item Functioning, (Hillsdale, NJ: Erlbaum), pp. 197–239.

Shealy, R., and Stout, W.F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. Psychometrika *58*, 159–194.

Strauss, A., and Corbin, J.M. (1990). Basics of Qualitative Research: Grounded Theory Procedures and Techniques (Sage).

Zumbo, B.D. (1999). A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores (Ottowa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense).