Instruction Committee Year End Report 2015-2016

The Chair of the Instruction Committee

Chiaki Yanagisawa

Members of the Committee

The members of the committee were Profs. Robin E. Brown (Library), Bogdan Danila (Science), Meghan M. Fitzgerald (Teacher Education), Catarina Mata (Science), Chamutal Noimann (English), Rifat A. Salam (Continuing Education and Work Force Development), Yan Chen (Computer Information Systems), and Chiaki Yanagisawa (Science).

Officers

At the first meeting on September 2, 2015, we elected the following officers: Prof. Yanagisawa as the chair, Prof. Danila as the secretary and Prof. Fitzgerald as the representative for the Executive Committee of the Academic Senate.

Dates of the Meetings

We met on the following dates and a quorum was met at all the meetings: September 2, 2015; October 7, 2015; November 4, 2015; December 2, 2015; February 3, 2016; March 2, 2016; April 6, 2016; May 11, 2016.

Main Topics and Issued discussed

1) Distant learning
   This topic were discussed at more than one meetings. The discussed issues included the way a distant-learning course is approved, no compensation for developing a distant-learning course, outcome of distant-learning courses compared with regular in-class courses such as the student dropout rate etc. Prof. Yanagisawa obtained the data from Bettina Hansel of the Office of Institutional Effectiveness and Analytics (see Appendix A). As we found out that extensive surveys on distant-learning courses have been conducted with the instructors of these courses, especially by Profs. Hachey (Teacher Education), Conway (Business) and Wladis (Mathematics) (see Appendix B), we decided to conduct an anonymous survey on distant-learning courses with full- and part-time-instructors who have never done these courses. The survey questions proposed by Profs. Mata and Salam went through several iterations after discussion among the members of the committee, and they were translated into a format for SurveyMonkey.com by the Office of Institutional Effectiveness and Analytics. An email urging participation in the survey by Prof. Yanagisawa, addressed to the full- and part-time instructors of BMCC, was sent out via the Office of Academic Affair on May 17, 2016. The statistics of the survey will be monitored and a summary of the survey will be reported to the next Instruction Committee and the Academic Senate. **If it is found from the survey result that there is some issue, we recommend the next committee to take a look at it.**

2) Student evaluation
   At several meetings we discussed about the student evaluation conducted toward the end of every semester. Some issues we discussed about were: how much of weight should be placed on student evaluation results for tenure and promotion review, and some biases such as gender and racial biases as evidenced by some studies collected by Prof. Yanagisawa (see Appendices C, D, E, and F). We did not discuss these issues enough to make any recommendation. However, **we recommend that the next committee should look at these issues.**

3) Advisement
   There have been cases in which students took courses for which they were not well prepared to take. These non-major students took courses offered for students in specific majors. These students were ill-advised. Prof. Mata took the lead to discuss with the Administration about this issue and the advisors will be made aware of this problem.

4) An issue with remedial courses
   Among the members of this committee there were some concerns with water-down remedial courses that did not give enough skills for students to be prepared for credit-bearing courses. Several suggestions were discussed: check GPA records of the students who took certain remedial courses and the history of pass-failure rate of remedial courses; and identify a few remedial courses of concern and talk to the coordinators of these courses. We have not taken any action on this issue but **we recommend that this issue should be discussed by the next committee.**

5) Cross-listing problem
   It was pointed out that there has been some resistance from the Registrar's Office to cross-list some courses. **If this issue is not resolved, it should be discussed in the next committee.**

<u>Recommendation to the Instruction Committee of 2016-2017</u>

We recommend the next committee to discuss the issues highlighted in block letters in the list of the major issues described above.

<u>List of Appendices</u>

The following documents mentioned in this report are attached as the appendices:

A. A Comparison of attrition rates between online and in-class courses
B. Survey summaries of online courses by BMCC Professors
C. Article: Student evaluation of teaching by Boring and Stark
D. Article: Study finds gender perception affects evaluations by Mulhere
E. Article: An evaluation of course evaluations by Stark
F. Article: What's in a name: Exposing gender biases in student ratings of teaching evaluation by MacNell et al.

Appendix A

FALL 2013	IN PERSON										ONLINE													
DISCIPLINE	Total Enrollments	Total W,	Attrition % (W,	WU	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %	F	Fail %	Total Enrollments	Total W,	Attrition % (W,	WU	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %	F	Fail %	Total Enrollments	Total W,	Attrition % (W,	WU
ANT	724					8	1%	77%		7%	24					1	4%	75%		17%				
AST	522	39	7%	9	2%	9	2%	83%		3%	0										48	7	15%	0
BIO	2691	348	13%	73	3%	21	1%	78%		5%	0										84	10	12%	2
BUS	2363	167	7%	102	4%	43	2%	83%		5%	37	10	27%	1	3%	3	8%	49%		5%	0			
CHE	1413	163	12%	32	2%	17	1%	78%		6%	0										71	13	18%	1
CRT	728	32	4%	42	6%	3	0%	83%		4%	0										0			
ECE	777	35	5%	14	2%	8	1%	85%		5%	24	2	8%	0	0%		0%	88%		4%	0			
ECO	1373	128	9%	47	3%	12	1%	80%		4%	118	16	14%	8	7%	2	2%	74%		4%	0			
EDU	91	6	7%	1	1%	0	0%	88%		3%	28	10	36%	5	18%		0%	43%		4%	0			
ENG	11093	824	7%	620	6%	104	1%	76%		6%	264	39	15%	9	3%	7	3%	70%		8%	49	8	16%	2
HED	3276	165	5%	125	4%	74	2%	80%		5%	0										22		0%	2
HUM	716	44	6%	31	4%	8	1%	83%		3%	69	2	3%	0	0%		0%	90%		3%	44	1	2%	1
LIN	151	17	11%	13	9%	2	1%	66%		7%	14	1	7%	0	0%		0%	93%		0%	0			
MAR	748	37	5%	23	3%	12	2%	84%		9%	40	8	20%	10	25%		0%	45%		0%	0			
MAT	10290	945	9%	644	6%	161	2%	64%		4%	238	43	18%	16	7%	10	4%	60%		9%	204	33	16%	14
OFF	90	6	7%	0	0%	2	2%	80%		9%	37	6	16%	0	0%		0%	81%		0%	0			
PHI	938	77	8%	61	7%	14	1%	72%		8%	46	5	11%	0	0%		0%	89%		0%	0			
POL	1443	146	10%	67	5%	28	2%	75%		6%	24	3	13%	0	0%		0%	88%		0%	0			
PSY	3731	291	8%	161	4%	39	1%	79%		6%	136	17	13%	7	5%	2	1%	70%		10%	110	6	5%	7
SBE	66	8	12%	1	2%	1	2%	85%		3%	16	6	38%	0	0%	2	13%	25%		19%	0			
SOC	2564	221	9%	110	4%	36	1%	77%		5%	46	7	15%	5	11%	2	4%	65%		4%	0			
SPE	4723	325	7%	333	7%	96	2%	77%		6%	26	1	4%	1	4%		0%	65%		12%	45	7	16%	1
SPN	2937	301	10%	163	6%	20	1%	76%		6%	124	35	28%	4	3%		0%	54%		10%				
AVERAGES			8%		4%		1%	79%		6%	1311		17%		5%		2%	68%		6%	677		11%	

SPRING 2014	IN PERSON	ONLINE

DISCIPLINE	Total Enrollments	Total W,	Attrition % (W,	WU)	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %	Total Enrollments	Total W,	Attrition % (W,	WU)	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %	Total Enrollments	Total W,	Attrition % (W,	WU)
ANT	676	66	10%	30	4%	7	1%	88%		9%	22	2	9%	0		1	5%	82%		5%	0			
AST	520	33	6%	8	2%	4	1%	94%		4%	0										47	3	6%	0
BIO	2195	338	15%	45	2%	16	1%	91%		6%	0										65	12	18%	3
BUS	2030	193	10%	103	5%	20	1%	87%		6%	28	7	25%	0		2	7%	54%		14%	0			
CED	296	18	6%	4	1%	1	0%	89%		3%	0										10	2	20%	0
CHE	1472	246	17%	40	3%	12	1%	93%		6%	0										66	7	11%	1
CRT	565	45	8%	31	5%	6	1%	90%		5%	25	1	4%	0		1	4%	92%		0%	0			
ECE	642	28	4%	19	3%	3	0%	94%		5%	23	6	26%	0			0%	57%		17%	23	0	0%	4
ECO	1148	110	10%	30	3%	5	0%	93%		7%	114	12	11%	2	2%	5	4%	77%		6%	0			
EDU	92	6	7%	10	11%	1	1%	86%		3%	0										35	2	6%	0
ENG	9735	999	10%	737	8%	109	1%	81%		7%	269	39	14%	9	3%	10	4%	68%		9%	86	27	31%	4
HED	2758	216	8%	177	6%	46	2%	83%		6%	0										33	3	9%	9
HUM	643	38	6%	14	2%	0	0%	94%		7%	73	5	7%	3	4%	1	1%	84%		0%	28	1	4%	0
LIN	99	16	16%	4	4%	0	0%	96%		14%	21	6	29%	2	10%	1	5%	52%		0%	23	4	17%	2
MAR	670	25	4%	8	1%	5	1%	86%		11%	41	8	20%	6	15%		0%	46%		2%	0			
MAT	9328	944	10%	710	8%	143	2%	68%		4%	176	35	20%	7	4%	2	1%	65%		8%	142	14	10%	10
OFF	72	8	11%	8	11%	4	6%	82%		2%	17	4	24%	1	6%		0%	47%		0%	0			
PHI	876	93	11%	73	8%	10	1%	71%		5%	46	7	15%	3	7%		0%	76%		2%	0			
POL	1296	144	11%	64	5%	13	1%	71%		11%	45	5	11%	3	7%	2	4%	73%		4%	0			
PSY	3493	298	9%	148	4%	27	1%	78%		7%	134	19	14%	1	1%	2	1%	76%		6%	142	14	10%	6
SBE	104	7	7%	7	7%	0	0%	89%		4%	21	4	19%	0	0%	1	5%	52%		24%	0			
SOC	2466	216	9%	133	5%	24	1%	76%		7%	63	10	16%	1	2%	6	10%	70%		3%	0			
SPE	4671	388	8%	409	9%	68	1%	74%		6%	39	5	13%	2	5%	1	3%	64%		10%	65	6	9%	13
SPN	2649	279	11%	151	6%	13	0%	76%		6%	175	32	18%	8	5%		0%	69%		6%	26		0%	
THE	526	23	4%	23	4%	5	1%	86%		4%											10	1	10%	1
AVERAGES			9%		5%		1%	85%		6%	1332		17%		5%		3%	67%		7%	801		11%	
SUMMER 2014	IN PERSON										ONLINE													

DISCIPLINE	Total Enrollments	Total W,	Attrition % (W,	WU)	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %	Total Enrollments	Total W,	Attrition % (W,	WU)	Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %	Total Enrollments	Total W,	Attrition % (W,	WU)
ANT	47	1	2%	1		0	0%	96%		0%	21						0%	100%		0%	0			
BIO	617	47	8%	9		6	1%	87%		3%	0										68	4	6%	0
BUS	100	1	1%	3		1	1%	93%		0%	12		0%		0%		0%	100%		0%	0			
CHE	350	19	5%	1		1	0%	89%		5%	0										17	0	0%	0
ECO	140	2	1%	3		0	0%	94%		1%	69	3	4%	0	0%	2	3%	86%		7%	0			
ENG	968	43	4%	20		8	1%	86%		2%	194	12	6%	5	3%	4	2%	80%		7%	0			
HED	111	3	3%	3		0	0%	91%		3%	46	2	4%	1	2%		0%	85%		9%	0			
MAR	33	3	9%	0		0	0%	76%		6%	9	1	11%	0	0%		0%	67%		0%	0			
MAT	2498	144	6%	35		5	0%	72%		2%	130	6	5%	0	0%	2	2%	81%		8%	0			
PSY	228	6	3%	3		1	0%	92%		1%	129	7	5%	1	1%	1	1%	88%		4%	0			
SOC	114	5	4%	2		0	0%	92%		2%	22	2	9%				0%	86%		5%	0			
SPE	220	9	4%	1		6	3%	90%		2%	38	3	8%				0%	87%		5%	0			
SPN	203	6	3%	6		1	0%	93%		0%	53	1	2%			2	4%	91%		2%	0			
AVERAGES			4%		0%		1%	88%		2%	723		5%		1%		1%	86%		4%	85		3%	

FALL 2014	IN PERSON										ONLINE													
DISCIPLINE	Total Enrollments	Total W,	Attrition % (W,	WU	Attrition % WU	Total WN, WD)	Attrition % (WN, WD)	Pass (A-D) %	F	Fail %	Total Enrollments	Total W,	Attrition % (W,	WU	Attrition % WU	Total WN, WD)	Attrition % (WN, WD)	Pass (A-D) %	F	Fail %	Total Enrollments	Total W,	Attrition % (W,	WU
ANT	1025	79	8%	52	5%	12	1%	77%	82	8%	32	0	0%	0	0%	1	3%	81%	5	16%				
AST	667	28	4%	13	2%	5	1%	86%	24	4%											50	1	2%	3
BIO	2598	282	11%	56	2%	14	1%	78%	162	6%											97	13	13%	4
BUS	2734	128	5%	99	4%	43	2%	78%	223	8%	21	2	10%	0	0%	1	5%	76%	2	10%				
CHE	2011	188	9%	61	3%	11	1%	76%	159	8%											69	6	9%	0
CRT	1098	68	6%	45	4%	14	1%	78%	67	6%	24	5	21%	2	8%	3	13%	58%	0	0%				
ECE	717	35	5%	19	3%	3	0%	81%	45	6%	12	2	17%	0	0%	0	0%	25%	7	58%	19	0	0%	3
ECO	1419	96	7%	69	5%	16	1%	81%	54	4%	114	12	11%	4	4%	5	4%	75%	7	6%				
EDU	113	4	4%	2	2%	1	1%	88%	0	0%	19	4	21%	2	11%	2	11%	53%	0	0%				

ENG	12568	758	6%	708	6%	153	1%	74%	766	6%	296	28	9%	26	9%	14	5%	65%	32	11%	46	2	4%	2
HED	2886	151	5%	212	7%	52	2%	76%	149	5%	22	2	9%	4	18%	2	9%	55%	1	5%				
HUM	795	17	2%	14	2%	1	0%	87%	36	5%	61	6	10%	0	0%	1	2%	84%	2	3%				
ITL											25	0	0%	0	0%	3	12%	84%	0	0%				
LIN	173	7	4%	2	1%	1	1%	77%	18	10%	11	1	9%	0	0%	1	9%	82%	0	0%	9	0	0%	1
MAR	857	46	5%	30	4%	10	1%	76%	100	12%	43	5	12%	5	12%	0	0%	42%	5	12%				
MAT	12059	712	6%	635	5%	164	1%	61%	533	4%	166	25	15%	12	7%	10	6%	63%	13	8%	63	4	6%	3
OFF	34	1	3%	3	9%	0	0%	79%	2	6%	29	1	3%	6	21%	0	0%	62%	1	3%				
PHI	1060	82	8%	94	9%	14	1%	69%	79	7%	47	6	13%	3	6%	0	0%	68%	6	13%				
POL	1497	113	8%	103	7%	26	2%	71%	138	9%	46	1	2%	1	2%	0	0%	91%	2	4%	142	9	6%	4
PSY	3906	230	6%	163	4%	35	1%	80%	238	6%	111	17	15%	1	1%	4	4%	67%	12	11%				
SBE	116	3	3%	0	0%	1	1%	83%	7	6%	21	2	10%	0	0%	0	0%	62%	6	29%				
SOC	2750	169	6%	141	5%	37	1%	76%	211	8%	89	6	7%	3	3%	6	7%	66%	15	17%				
SPE	5487	322	6%	384	7%	84	2%	78%	278	5%	85	10	12%	5	6%	8	9%	59%	12	14%	116	11	9%	12
SPN	2650	255	10%	164	6%	21	1%	71%	252	10%	153	19	12%	3	2%	5	3%	63%	27	18%	95	9	9%	3
Totals and Averages			6%		4%		1%	77%		6%	1427		10%		5%		5%	66%		11%	559		6%	

HYBRID

Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %	F	Fail %
0%			71%		8%
2%	1	1%	79%		6%
1%		0%	75%		1%
4%	1	2%	65%		12%
9%		0%	91%		0%
2%		0%	91%		5%
7%	5	2%	56%		5%
6%		0%	80%		7%
2%		0%	60%		16%
4%		1%	74%		7%

HYBRID

Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %
0%		0%	83%		9%
5%		0%	75%		2%
0%		0%	80%		0%
2%		0%	79%		5%
17%	2	9%	70%		4%
0%		0%	80%		11%
5%	1	1%	56%		7%
27%		0%	58%		6%
0%		0%	93%		4%
9%		0%	74%		0%
7%		0%	66%		1%
4%	1	1%	74%		11%
20%	3	5%	54%		5%
0%		0%	100%		0%
10%		0%	60%		20%
7%		1%	73%		6%

HYBRID

Attrition % WU	Total WN	Attrition % (WN)	Pass (A-D) %		Fail %
	0	0%	99%		1%
	0	0%	100%		0%
		0%	100%		1%

HYBRID

Attrition % WU	Total WN, WD	Attrition % (WN, WD)	Pass (A-D) %	F	Fail %
6%	0	0%	80%	0	
4%	1	1%	64%	9	9%
0%	1	1%	78%	4	6%
16%	0	0%	63%	4	21%

4%	2	4%	76%	4	9%
11%	0	0%	89%	0	0%
5%	0	0%	73%	9	14%
3%	3	2%	73%	20	14%
10%	2	2%	60%	15	13%
3%	0	0%	72%	14	15%
6%		1%	73%		11%

Appendix B

Appendix B
WHAT THE DATA REVEALS

Alyse C. Hachey- Teacher Education/BMCC
Katherine Conway- Business/BMCC
Claire Wladis- Mathematics/BMCC

ONLINE LEARNING (COMMUNITY COLLEGES; FACTORS EFFECTING)

2015a

Wladis, C.W., Conway, K.M. & Hachey, A.C. (2015). Using course-level factors as predictors of online course outcomes: A multilevel analysis at an urban community college. ***Studies in Higher Education.***
http://www.tandfonline.com/doi/abs/10.1080/03075079.2015.1045478#.VaH90_l2OO0

Abstract: Research has documented lower retention rates in online versus face-to-face courses. However, little research has focused on the impact of course-level characteristics (e.g. elective versus distributional versus major requirements; difficulty level; STEM status) on online course outcomes. Yet, focusing interventions at the course level versus the student level may be a more economical approach to reducing online attrition. This study used multi-level modeling, and controlled for the effects of both instructor-level and student characteristics, to measure the relationship of course-level characteristics with successful completion of online and face-to-face courses. Elective courses, and to a lesser extent distributional course requirements, were significantly more likely to have a larger gap in successful course completion rates online versus face-to-face, when compared with major course requirements. Upper level courses had better course completion rates overall, but a larger gap in online versus face-to-face course outcomes than lower level courses.

2014a

Wladis, C., Hachey, A. C. and Conway, K. (2014). The role of enrollment choice in online education: Course selection rationale and course difficulty as factors affecting retention, ***Journal of Asynchronous Learning Networks, 18*****(3).**
http://olj.onlinelearningconsortium.org/index.php/olj/article/view/391/109

Abstract: Previous research supports that retention is significantly lower in online courses in comparison to face-to face courses; however, much of the past research on student retention in the online environment focuses on student characteristics, with little existing on the impact of course type. This study identifies and analyzes two key factors that may be impacting online retention: the student's reason for taking the course (whether as an elective or a requirement) and course difficulty level. The results of this study indicate that a student's reason for taking a lower level course drastically impacts the likelihood of withdrawal in the online environment, while having no effect in face-to-face classes. In particular, for lower level courses which students took as an elective or distributional requirement, the online environment seemed to make them much more likely to drop out. The findings suggest that in the online environment, the student's reason for course enrollment (an elective versus a requirement) may be considered a risk indicator and that focused learner support targeted at particular course types may be needed to increase online persistence and retention.

2014b

Hachey, A. C., Wladis, C. and Conway, K. (2014). Do prior online course outcomes provide more information than G.P.A. alone in predicting subsequent online course grades and retention? An observational study at an urban community college, ***Computers & Education. 72*****, 59-67. doi:http://dx.doi.org/10.1016/j.compedu.2013.10.012**
http://www.sciencedirect.com/science/article/pii/S0360131513002972

Abstract: In this study, prior online course outcomes and pre-course enrollment G.P.A. were used as predictors of subsequent online course outcomes, and the interaction between these two factors was assessed in order to determine the extent to which students with similar G.P.A.'s but with different prior online course outcomes may differ in their likelihood of successfully completing a subsequent online course. This study used a sample of 962 students who took an online course

at a large urban community college from 2004 to 2010. Results indicate that prior online course experience is a very significant predictor of successful completion of subsequent online courses, even more so than G.P.A. For students with no prior online course experience, G.P.A. was a good predictor of future online course outcomes; but for students with previous online course experience prior online course outcomes was a more significant predictor of future online course grades and retention than G.P.A.

2013a

Hachey, A.C., Wladis, C. & Conway, K.M. (2013) Balancing retention and access in online courses: restricting enrollment… Is it worth the cost? ***Journal of College Student Retention: Research, Theory & Practice,*** **15(1), 9-36.**
http://csr.sagepub.com/content/15/1.toc

Abstract: Open access is central to the Community College mission. For this reason, any restriction in online enrollments should not be undertaken lightly. This study uses institutional data gathered from a large, urban community college to examine a policy aimed at increasing student retention in online courses by restricting those eligible to enroll based on G.P.A. The data, counter to expectations, show that the policy did not significantly impact attrition rates. Further analysis reveals that a high G.P.A. cut-off (3.0) is needed to significantly affect attrition rates; however, this would severely restrict those eligible to enroll. The data indicate that students in the middle G.P.A. range (2.0-3.5) have the highest proportional difference in attrition between online and face-to-face courses. The results suggest that rather than focusing on G.P.A. restrictions, community colleges may be better served by addressing research and interventions targeted toward other factors to increase student retention in online learning.

2013b

Hachey, A.C., Conway, K.M. and Wladis, C. (2013). Community colleges and underappreciated assets: Using institutional data to promote success in online learning. ***Online Journal of Distance Learning Administration, 16*****(1), Spring.**
http://www.westga.edu/~distance/ojdla/spring161/hachey_wladis.html

Abstract: Adapting to the 21st century, community colleges are not adding brick and mortar to meet enrollment demands. Instead, they are expanding services through online learning, with at least 61% of all community college students taking online courses today. As online learning is affording alternate pathways to education for students, it is facing difficulty in meeting outcome standards; attrition rates for the past decade have been found to be significantly higher for online courses than face-to-face courses. Yet, there is a lack of empirical investigation on community college online attrition, despite the fact that course and institutional management systems today are automatically collecting a wealth of data which are not being utilized but are readily available for study. This article presents a meta-review of one community college's realization of their underappreciated asset… the use of institutional data to address the dearth of evidence on factors effecting attrition in online learning.

2012a

Hachey, A. C., Wladis, C. and Conway, K. (2012) Is the second time the charm? Investigating trends in online re-enrollment, retention and success. ***The Journal of Educators Online, 9*****(1), 1-25.**
http://www.thejeo.com/Archives/Volume9Number1/HacheyetalPaper.pdf

Abstract: This study found that prior online course experience is strongly correlated with future online course success. In fact, knowing a student's prior online course success explains 13.2% of the variation in retention and 24.8% of the variation in online success in our sample, a large effect size. Students who have not successfully completed any previous online courses have very low success and retention rates, and students who have successfully completed all prior online courses have fairly high success and retention rates. Therefore, this study suggests that additional support services need to be provided to previously unsuccessful online learners, while students who succeed online should be encouraged to enroll in additional online courses in order to increase retention and success rates in online learning.

2011a

Conway, K., Hachey, A. C. and Wladis, C. (2011). Growth of online education in a community college, ***Academic Exchange Quarterly, 15*****(3), 96-101.**
http://rapidintellect.com/AEQweb/cho4929.htm

Abstract: This case study examines the evolution of online education at a large urban community college. It outlines issues related to course development, administration, student and faculty support. Online course enrollment, student and faculty perceptions and organizational issues were evaluated a decade after online education was introduced at the college. At both the inception of online education and in order to expand successfully, external funding was crucial for program success.

ONLINE STEM LEARNING

2015b

Wladis, C.W., Hachey, A.C. & Conway, K.M. (In Press). Which STEM majors enroll in online courses and why should we care? The impact of ethnicity, gender, and non-traditional student characteristics. ***Computers & Education, 87 (285-308).***

http://www.sciencedirect.com/science/article/pii/S036013151530004X

Abstract: Using data from roughly 27,800 undergraduate STEM (science, technology, engineering and mathematics) majors in the National Postsecondary Student Aid Study (NPSAS), this research examines the relationship between race/ethnicity, gender and non-traditional student characteristics and online course enrollment. Hispanic and Black STEM majors were significantly less likely, and female STEM majors significantly more likely, to take online courses even when academic preparation, socioeconomic status (SES), citizenship and English-as-second-language (ESL) status were controlled. Furthermore, non-traditional student characteristics strongly increased the likelihood of enrolling in an online course, more so than any other characteristic, with online enrollment probability increasing steeply as the number of non-traditional factors increased. The impact of non-traditional factors on online enrollment was significantly stronger for STEM than non-STEM majors.

2015c

Wladis, C. W., Hachey, A. C. & Conway, K. M. (2015). The online STEM classroom – Who succeeds? An exploration of the impact of ethnicity, gender and non-traditional student characteristics in the community college context. ***Community College Review*****, 43(2), 142-164.**

http://lib2.bmcc.cuny.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=101620296&site=ehost-live&scope=site

Abstract: This study used a sample of about 3,600 students in online and face-to-face courses matched by course, instructor, and semester from a large urban community college in the Northeast to analyze how ethnicity, gender and non-traditional student characteristics related to STEM [Science, Technology, Engineering, Mathematics] course outcomes online versus face-to-face. Multilevel logistic regression (with course/instructor as grouping factor) and propensity score matching were utilized. Results indicated that older students did significantly better in online STEM courses, and that women did significantly worse (although still no worse than men) online, than would be expected based on their outcomes in comparable face-to-face STEM courses. There was no significant interaction between the online medium and ethnicity, suggesting that while Black and Hispanic students may do worse than their White and Asian peers in both online and face-to-face STEM courses, this gap was not increased by the online environment.

2015d

Wladis, C., Hachey, A.C. & Conway, K.M. (2015). The representation of minority, female, and non-traditional STEM majors in the online environment at community colleges: A nationally representative study. ***Community College Review*****, 43(1), 89-114.**

http://lib2.bmcc.cuny.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=99752730&site=ehost-live&scope=site

Abstract: Using data from the more than 2,000 community college STEM majors in the National Postsecondary Student Aid Study, this research examines which groups may be underrepresented online and identifies characteristics which differ significantly between online and face-to-face students. It provides essential information on self-selection into online courses that is necessary for future observational studies of online versus face-to-face outcomes. The results show

that Hispanic students were significantly less likely to enroll online, with Black and Hispanic male students particularly underrepresented. Women were significantly more likely to enroll online, as were students with non-traditional student characteristics (delayed enrollment; no high school diploma; part-time enrollment; financially independent; have dependents; single parent status; working full-time). At community colleges, ethnicity was a stronger predictor than non-traditional characteristics, whereas at 4-year colleges the reverse was true: each additional non-traditional risk factor increased the likelihood of online enrollment by two and five percentage points at 2-year and 4-year colleges respectively.

2014c

Hachey, A. C., Wladis, C. and Conway, K. (2014). Prior online course experience and G.P.A. as predictors of subsequent online STEM course outcomes, *Internet and Higher Education, 25*, 11-17.

http://www.sciencedirect.com/science/article/pii/S1096751614000827

Abstract: This study found that G.P.A. and prior online experience both predicted online STEM course outcomes. While students with higher G.P.A.'s were also more likely to have successfully completed prior online courses, prior online course experience added significant information about likely future STEM online outcomes, even when controlling for G.P.A. Students who had successfully completed all prior online courses had significantly higher rates of successful online STEM course completion at all G.P.A. levels than students who had failed to complete even one prior online course successfully. Students who had dropped or earned a D or F grade in even one prior online course had significantly lower rates of successful online STEM course completion than students with no prior online experience, even when controlling for G.P.A. This suggests that prior online course outcomes should be combined with G.P.A. when attempting to identify community college students at highest risk in online STEM courses.

2014d

Wladis, C., Hachey, A.C. & Conway, K.M. (2014). An investigation of course-level factors as predictors of online STEM course outcomes. *Computers & Education, 77*, 145-150.

http://www.sciencedirect.com/science/article/pii/S0360131514001006

Abstract: This study analyzed students who took STEM courses online or face-to-face at a large urban community college in the Northeastern U.S. to determine which course-level characteristics most strongly predicted higher rates of dropout or D/F grades in online STEM courses than would be expected in comparable face-to-face courses. While career and elective STEM courses had significantly higher success rates *face-to-face* than liberal arts and major requirement STEM courses respectively, career STEM courses had significantly higher success rates online than would be expected, while elective STEM courses had significantly lower success rates online than would be expected given the face-to-face results. Once propensity score matching was used to generate a matched subsample which was balanced on a number of student characteristics, differences in course outcomes by course characteristics were no longer significant. This suggests that while certain types of STEM courses can be identified as higher or lower risk in the online environment, this appears not to be because of the courses themselves, but rather because of the particular characteristics of the students who choose to take these courses online. Findings suggests that one potential intervention for improving online STEM course outcomes could be to target students in specific courses which are at higher risk in the online environment; this may allow institutions to leverage interventions by focusing them on the STEM courses at greatest risk of lower online success rates, where the students who are at highest risk of online dropout seem to be concentrated.

2013c

Wladis, C., Hachey, A.C., Conway, K.M. (2013). Are online students in STEM (science, technology, engineering and mathematics) courses at greater risk of non-success? *American Journal of Educational Studies.* 6(1), 65-84.

http://lib2.bmcc.cuny.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=99178909&site=ehost-live&scope=site

Abstract: Both online and STEM courses have been shown to have lower student retention; however, there is little research indicating what effect the online environment may have on retention in STEM courses specifically. This study compares retention rates for online and face-to-face STEM and non-STEM courses to determine if the online environment affects STEM courses differently than non-STEM courses. In addition, different subcategories of STEM courses are compared to see if the effects of the online environment are different for different course subtypes. Each online course is matched with the same course taught face-to-face by the same instructor in the same semester to control for possible confounding effects. This study

found that retention rates in STEM courses were more strongly decreased by the online environment than in non-STEM courses. In particular, the course types which had significantly lower retention online were lower level STEM courses taken as electives or distributional requirements.

2012b

Wladis, C., Hachey, A. C. and Conway, K. (2012) An analysis of the effect of the online environment on STEM student success, In S. Brown, S. Larsen, K. Marrongelle, and M. Oehrtman (Eds.), *Proceedings of the 15th Annual Conference on Research in Undergraduate Mathematics Education*, (Vol.2). Portland, Oregon, 291-300.

(Not available from CUNY Libraries. May be available on the open web.)

Abstract: Both online and STEM courses have been shown to have lower student retention; however, there is little research indicating what effect the online environment may have on retention in STEM courses specifically. This study compares retention rates for online and face-to-face STEM and non-STEM courses to determine if the online environment affects STEM courses differently than non-STEM courses. In addition, different subcategories of STEM courses are compared to see if the effects of the online environment are different for different course subtypes. Each online course is matched with the same course taught face-to-face by the same instructor in the same semester to control for possible confounding effects. This study found that retention rates in STEM courses were more negatively impacted by the online environment than in non-STEM courses. In particular, the course types which had significantly lower retention online were lower level STEM courses taken as electives or distributional requirements.

ONLINE LEARNING AND HISPANIC STUDENTS

2014e

Conway, K.L., Hachey, A.C. and Wladis, C.W. (2014). A new disapora: Latino(a)s in the online environment. In Y. Medina and A. D. Macaya (Eds.). *Latinos on the East Coast: A critical reader*. NY, NY: Peter Lang.

(On order for the BMCC Library.)

Abstract: The Latino/a diaspora from the Caribbean, Central and South America to the U.S. is well documented. Many of these immigrants have settled in communities where they now constitute a majority. As noted herein, the Latino/a population varies by region, in its ethnicity, immigration status and longevity in the U.S. In the Northeast, the Latino/a population grew at a rate ten times as fast as the rest of the population in the decade ending 2010. Overall, and specifically in higher education, Latino/a students are the largest minority group and the fastest growing. Many of these students begin at community colleges. But as Latino/a students succeed in college in greater numbers, a new migration is occurring in higher education: to the online environment. This chapter examines Northeast Latino/a student enrollments and persistence in online courses in comparison to the traditional face-to-face classroom and in comparison to other ethnicities. Latino/a students, while enrolling in college in large numbers, continue to lag other student groups in graduation rates, and it is critical to understand if an increase in online course offerings will help or hinder Latino/a student success.

2011b

Conway, K., Wladis, C. and Hachey, A. C. (2011) Minority student access in the online environment, *Hispanic Educational Technologies Services (HETs) Journal*, *II*.

http://hets.org/ejournal/2014/07/30/minority-student-access-in-the-online-environment/

Abstract: Using registration and transcript data, the authors explored differences in online course enrollment across different student groups. This study revealed that minority students do not enroll in online courses to the same extent as their White student peers. An even greater issue is that Black and Hispanic students, regardless of the course delivery medium, continue to have lower G.P.A. s than their White and Asian/Pacific Islander (PI) student peers. This finding reinforces prior research that suggests Black and Hispanic student groups need additional support in order to be successful in college, and that greater recruitment efforts for online courses are needed for all minority groups.

Prior research has also shown that students who enroll in online courses at the college have higher G.P.A.'s than students who enroll in face-to-face courses; however, this study reveals a notable exception to this pattern. In contrast to other ethnic groups, there is no significant difference between Asian/PI students who select face to face versus online

courses, suggesting that there are differences in the factors that determine online enrollment in this group compared to others.

Appendix C

Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness

Anne Boring*

OFCE, SciencesPo, Paris

PSL, Université Paris-Dauphine, LEDa, UMR DIAL

Kellie Ottoboni and Philip B. Stark

Department of Statistics

University of California, Berkeley

January 5, 2016

*This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 612413, for the EGERA (Effective Gender Equality in Research and the Academia) European project.

The truth will set you free, but first it will piss you off.

Gloria Steinem

Abstract

Student evaluations of teaching (SET) are widely used in academic personnel decisions as a measure of teaching effectiveness. We show:

- SET are biased against female instructors by an amount that is large and statistically significant
- the bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded
- the bias varies by discipline and by student gender, among other things
- it is not possible to adjust for the bias, because it depends on so many factors
- SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness
- gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors

These findings are based on nonparametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five-year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university.

1 Background

Student evaluations of teaching (SET) are used widely in decisions about hiring, promoting, and firing instructors. Measuring teaching effectiveness is difficult—for students, faculty, and administrators alike. Universities generally treat SET as if they primarily measure teaching effectiveness or teaching quality. While it may seem natural to think that students' answers to questions like "how effective was the instructor?" measure teaching effectiveness, it is not a foregone conclusion that they do. Indeed, the best evidence so far shows that they do not: they have *biases*[1] that are stronger than any connection they might have with effectiveness. Worse, in some circumstances the association between SET and an objective measure of teaching effectiveness is *negative*, as our results below reinforce.

Randomized experiments [Carrell and West, 2010, Braga et al., 2014] have shown that students confuse grades and grade expectations with the long-term value of a course and that SET are not associated with student performance in follow-on courses, a proxy for teaching effectiveness. On the whole, high SET seem to be a reward students give instructors who make them anticipate getting a good grade, for whatever reason; for extensive discussion, see Johnson [2003, Chapters 3–5].

Gender matters too. Boring [2015a] finds that SET are affected by gender biases and stereotypes. Male first-year undergraduate students give more *excellent* scores to male instructors, even though there is no difference between the academic performance of male students of male and of female instructors. Experimental work by MacNell et al. [2014] finds that when students think an instructor is female, students rate the instructor lower on every aspect of teaching, including putatively objective

[1] Centra and Gaubatz [2000, p.17] define bias to occur when "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching, such as increased student learning."

measures such as the timeliness with which instructors return assignments.

Here, we apply nonparametric permutation tests to data from Boring [2015a] and MacNell et al. [2014] to investigate whether SET primarily measure teaching effectiveness or biases using a higher level of statistical rigor. The two main sources of bias we study are students' grade expectations and the gender of the instructor. We also investigate variations in bias by discipline and by student gender.

Permutation tests allow us to avoid contrived, counterfactual assumptions about parametric generative models for the data, which regression-based methods (including ordinary linear regression, mixed effects models, logistic regression, etc.) and methods such as t-tests and ANOVA generally require. The null hypotheses for our tests are that some characteristic—e.g., instructor gender—amounts to an arbitrary label and might as well have been assigned at random.

We work with course-level summaries to match how institutions use SET: typically, SET are averaged for each offering of a course, and those averages are compared across instances of the course, across courses in a department, across instructors, and across departments. Stark and Freishtat [2014] discuss statistical problems with this reduction to and reliance upon averages.

We find that the association between SET and an objective measure of teaching effectiveness, performance on the anonymously graded final, is weak and—for these data—generally not statistically significant. In contrast, the association between SET and (perceived) instructor gender is large and statistically significant: instructors whom (students believe) are male receive significantly higher average SET.

In the French data, *male* students tend to rate male instructors higher than they rate female instructors, with little difference in ratings by female students. In the US data, *female* students tend to rate (perceived) male instructors higher than they rate (perceived) female instructors, with little difference in ratings by male students.

The French data also show that gender biases vary by course topic, and that SET have a strong positive association with students' grade expectations.

We therefore conclude that SET primarily do not measure teaching effectiveness; that they are strongly and non-uniformly biased by factors including the genders of the instructor and student; that they disadvantage female instructors; and that it is impossible to adjust for these biases. SET should not be relied upon as a measure of teaching effectiveness. Relying on SET for personnel decisions has disparate impact by gender, in general.

2 Data

2.1 French Natural Experiment

These data, collected between 2008 and 2013, are a census of 23,001 SET from 4,423 first-year students at a French university students (57% women) in 1,177 sections, taught by 379 instructors (34% women). The data are not public, owing to French restrictions on human subjects data. Boring [2015a] describes the data in detail. Key features include:

- All first-year students take the same six mandatory courses: History, Macroeconomics, Microeconomics, Political Institutions, Political Science, and Sociology. Each course has one (male) professor who delivers the lectures to groups of approximately 900 students. Courses have sections of 10–24 students. Those sections are taught by a variety of instructors, male and female. The instructors have considerable pedagogical freedom.

- Students enroll in "triads" of sections of these courses (three courses per semester). The enrollment process does not allow students to select individual instructors.

The assignment of instructors to sets of students is as if at random, forming a *natural experiment*. It is reasonable to treat the assignment as if it is independent across courses.

- Section instructors assign interim grades during the semester. Interim grades are known to the students before the students submit SET. Interim grades are thus a proxy for students' grade expectations.

- Final exams are written by the course professor, not the section instructors. Students in all sections of a course in a given year take the same final. Final exams are graded anonymously, except in Political Institutions, which we therefore omit from analyses involving final exam scores. To the extent that the final exam measures appropriate learning outcomes, performance on the final is a measure of the effectiveness of an instructor: in a given course in a given year, students of more effective instructors should do better on the final exam, on average, than students of less effective instructors.

- SET are mandatory: response rates are nearly 100%.

SET include closed-ended and open-ended questions. The item that attracts the most attention, especially from the administration, is the *overall score*, which is treated as a summary of the other items. The SET data include students' individual evaluations of section instructors in microeconomics, history, political institutions, and macroeconomics for the five academic years 2008–2013, and for political science and sociology for the three academic years 2010–2013 (these two subjects were introduced in 2010). The SET are anonymous to the instructors, who have access to SET only after all grades have been officially recorded.

Table 1: Summary statistics of sections

course	# sections	# instructors	% Female instructors
Overall	**1,194**	**379**	**33.8%**
History	230	72	30.6%
Political Institutions	229	65	20.0%
Microeconomics	230	96	38.5%
Macroeconomics	230	93	34.4%
Political Science	137	49	32.7%
Sociology	138	56	46.4%

Data for a section of Political Institutions that had an experimental online format are omitted. Political Science and Sociology originally were not in the triad system; students were randomly assigned by the administration to different sections.

2.2 US Randomized Experiment

These data, described in detail by MacNell et al. [2014], are available at `http://n2t.net/ark:/b6078/d1mw2k`. Students in an online course were randomized into six sections of about a dozen students each, two taught by the primary professor, two taught by a female graduate teaching assistant (TA), and two taught by a male TA. In one of the two sections taught by each TA, the TA used her or his true name; in the other, she or he used the other TA's identity. Thus, in two sections, the students were led to believe they were being taught by a woman and in two they were led to believe they were being taught by a man. Students had no direct contact with TAs: the primary interactions were through online discussion boards. The TA credentials presented to the students were comparable; the TAs covered the same material; and assignments were returned at the same time in all sections (hence, objectively, the TAs returned assignments equally promptly in all four sections).

SET included an overall score and questions relating to professionalism, respectfulness, care, enthusiasm, communication, helpfulness, feedback, promptness, consistency, fairness, responsiveness, praise, knowledge, and clarity. Fourty-seven students in the four sections taught by TAs finished the class, of whom 43 submitted SET.

The SET data include the genders and birth years of the students;[2] the grade data do not. The SET data are not linked to the grade data.

3 Methods

Previous analyses of these data relied on parametric tests based on null hypotheses that do not match the experimental design. For example, the tests assumed that SET of male and female instructors are independent random samples from normally distributed populations with equal variances and possibly different means. As a result, the p-values reported in those studies are for unrealistic null hypotheses and might be misleading.

In contrast, we use permutation tests based on the as-if-random (French natural experiment) or truly random (US experiment) assignment of students to class sections, with no counterfactual assumption that the students, SET scores, grades, or any other variables comprise random samples from any populations, much less populations with normal distributions.

In most cases, our tests are *stratified*. For the US data, for instance, the randomization is stratified on the actual TA: students are randomized within the two sections taught by each TA, but students assigned to different TAs comprise different strata. The randomization is independent across strata. For the French data, the randomization is stratified on course and year: students in different courses or in different years comprise different strata, and the randomization is independent across strata. The null distributions of the test statistics[3] are induced by this random assignment, with no assumption about the distribution of SET or other variables, no parameter

[2] One birth year is obviously incorrect, but our analyses do not rely on the birth years.

[3] The test statistics are correlations of a response variable with experimental variables, or differences in the means of a response variable across experimental conditions, aggregated across strata.

estimates, and no model.

3.1 Illustration: French natural Experiment

The selection of course sections by students at the French university—and the implicit assignment of instructors to sets of students—is as if at random within sections of each course each year, independent across courses and across years. The university's triad system groups students in their classes across disciplines, building small cohorts for each semester. Hence, the randomization for our test keeps these groups of students intact. Stratifying on course topic and year keeps students who took the same final exam grouped in the randomization and honors the design of the natural experiment.

Teaching effectiveness is multidimensional [Marsh and Roche, 1997] and difficult to define, much less measure. But whatever it is, effective teaching should promote student learning: *ceteris paribus*, students of an effective instructor should have better learning outcomes than students of an ineffective instructor have. In the French university, in all courses other than Political Institutions,[4] students in every section of a course in a given year take the same anonymously graded final exam. To the extent that final exams are designed well, scores on these exams reflect relevant learning outcomes for the course. Hence, in each course each semester, students of more effective instructors should do better on the final, on average, than students of less effective instructors.

Consider testing the hypothesis that SET are unrelated to performance on the final exam against the alternative that, all else equal, students of instructors who get higher average SET get higher final exam scores, indicating that they learned more.

[4]The final exam in Political Institutions is oral and hence not graded anonymously.

For this hypothesis test, we omit Political Institutions because the final exam was not anonymous.

The test statistic is the average over courses and years of the Pearson correlation between mean SET and mean final exam score among sections of each course each year. If SET do measure instructors' contributions to learning, we would expect this average correlation to be positive: sections with above-average mean SET in each discipline each year would tend to be sections with above-average mean final exam scores. How surprising is the observed average correlation, if there is no overall connection between mean SET and mean final exam for sections of a course?

There are 950 "individuals," course sections of subjects other than Political Institutions. Each of the 950 course sections has an average SET and an average final exam score. These fall in $3 \times 5 + 2 \times 3 = 21$ year-by-course strata. Under the randomization, within each stratum, instructors are assigned sections independently across years and courses, with the number of sections of each course that each instructor teaches each year held fixed. For instance, if in 2008 there were N sections of History taught by K instructors in all, with instructor k teaching N_k sections, then in the randomization, all

$$\binom{N}{N_1 \cdots N_K} \tag{1}$$

ways of assigning N_k of the N 2008 History sections to instructor k, for $k = 1, \ldots, K$, would be equally likely. The same would hold for sections of other courses and other years. Each combination of assignments across courses and years is equally likely: the assignments are independent across strata.

Under the null hypothesis that SET have no relationship to final exam scores, average final exam scores for sections in each course each year are *exchangeable* given the average SET for the sections. Imagine "shuffling" (i.e., permuting) the average

final exam scores across sections of each course each year, independently for different courses and different years. For each permutation, compute the Pearson correlation between average SET for each section and average final exam score for each section, for each course, for each year. Average the resulting 21 Pearson correlations. The probability distribution of that average is the null distribution of the test statistic. The p-value is the upper tail probability beyond the observed value of the test statistic, for that null distribution.

The hypothetical randomization holds triads fixed, to allow for cohort effects and to match the natural experiment. Hence, the test is conditional on which students happen to sign up for which triad. However, if we test at level no greater than α conditionally on the grouping of students into triads, the unconditional level of the resulting test across all possible groupings is no greater than α:

$$\begin{aligned}
\Pr\{\text{ Type I error }\} &= \sum_{\text{all possible sets of triads}} \Pr\{\text{ Type I error } | \text{ triads }\} \Pr\{\text{ triads }\} \\
&\leq \sum_{\text{all possible sets of triads}} \alpha \Pr\{\text{triads }\} \\
&= \alpha \sum_{\text{all possible sets of triads}} \Pr\{\text{triads }\} \\
&= \alpha. \qquad (2)
\end{aligned}$$

It is not practical to enumerate all possible permutations of sections within courses and years, so we estimate the p-value by performing 10^5 random permutations within each stratum, finding the value of the test statistic for each overall assignment, and comparing the observed value of the test statistic to the empirical distribution of those 10^5 random values. The probability distribution of the number of random permutations assignments for which the test statistic is greater than or

equal to its observed value is Binomial, with n equal to the number of overall random permutations and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at https://github.com/kellieotto/SET-and-Gender-Bias. Results for the French data are below in section 4.

3.2 Illustration: US Experiment

To test whether perceived instructor gender affects SET in the US experiment, we use the Neyman "potential outcomes" framework [Neyman et al., 1990]. A fixed number N of individuals—e.g., students or classes—are assigned randomly (or as if at random by Nature) into $k \geq 2$ groups of sizes $N_1, \ldots, N_k$. Each group receives a different treatment. "Treatment" is notional. For instance, the treatment might be the gender of the class instructor.

For each individual i, we observe a numerical response R_i. If individual i is assigned to treatment j, then $R_i = r_{ij}$. The numbers $\{r_{ij}\}$ are considered to have been fixed before the experiment. (They are not assumed to be a random sample from any population; they are not assumed to be realizations of any underlying random variables.) Implicit in this notation is the *non-interference* assumption that each individual's response depends only on the treatment that individual receives, and not on which treatments other individuals receive.

We observe only one potential outcome for individual i, depending on which treatment she or he receives. In this model, the responses $\{R_i\}_{i=1}^N$ are random, but only because individuals are assigned to treatments at random, and the assignment determines which of the fixed values $\{r_{ij}\}$ are observed.

In the experiment conducted by MacNell et al. [2014], N students were assigned

at random to six sections of an online course, of which four were taught by TAs. Our analysis focuses on the four sections taught by TAs. We condition on the assignment of students to the two sections taught by the professor. Each remaining student i could be assigned to any of $k = 4$ treatment conditions: either of two TAs, each identified as either male or female. The assignment of students to sections was random: each of the

$$\binom{N}{N_1 N_2 N_3 N_4} = \frac{N!}{N_1! N_2! N_3! N_4!} \tag{3}$$

possible assignments of N_1 students to TA 1 identified as male, N_2 student to TA 1 identified as female, etc., was equally likely.

Let r_{i1} and r_{i2} be the ratings student i would give TA 1 when TA 1 is identified as male and as female, respectively; and let r_{i3} and r_{i4} the ratings student i would give TA 2 when that TA is identified as male and as female, respectively. Typically, the null hypotheses we test assert that for each i, some subset of $\{r_{ij}\}$ are equal. For assessing whether the identified gender of the TA affects SET, the null hypothesis is that for each i, $r_{i1} = r_{i2}$ (the rating the ith student would give TA 1 is the same, whether TA 1 is identified as male or female), and $r_{i3} = r_{i4}$ (the rating the ith student would give TA 2 is the same, whether TA 2 is identified as male or female). Different students might give different ratings under the same treatment condition (the null does not assert that $r_{ij} = r_{\ell j}$ for $i \neq \ell$), and the ith student might give different ratings to TA 1 and TA 2 (the null does not assert that $r_{i1} = r_{i3}$). The null hypothesis makes no assertion about the population distributions of $\{r_{i1}\}$ and $\{r_{i3}\}$, nor does it assert that $\{r_{ij}\}$ are a sample from some super-population.

For student i, we observe exactly one of $\{r_{i1}, r_{i2}, r_{i3}, r_{i4}\}$. If we observe r_{i1}, then—if the null hypothesis is true—we also know what r_{i2} is, and vice versa, but we do not know anything about r_{i3} or r_{i4}. Similarly, if we observe either r_{i3} or r_{i4} and the

null hypothesis is true, we know the value of both, but we do not know anything about r_{i1} or r_{i2}.

Consider the average SET (for any particular item) given by the N_2+N_4 students assigned to sections taught by an apparently female TA, minus the average SET given by the N_1+N_3 students assigned to sections taught by an apparently male TA. This is what MacNell et al. [2014] tabulate as their key result. If the perceived gender of the TA made no difference in how students rated the TA, we would expect the difference of averages to be close to zero.[5] How "surprising" is the observed difference in averages?

Consider the

$$\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3} \tag{4}$$

assignments that keep the same N_1+N_2 students in TA 1's sections (but might change which of those sections a student is in) and the same N_3+N_4 students in TA 2's sections. For each of those assignments, we know what $\{R_i\}_{i=1}^{N}$ would have been if the null hypothesis is true: each would be exactly the same as its observed value, since those assignments keep students in sections taught by the same TA. Hence, we can calculate the value that the test statistic would have had for each of those assignments.

Because all $\binom{N}{N_1 N_2 N_3 N_4}$ possible assignments of students to sections are equally likely, these $\binom{N_1+N_2}{N_1} \times \binom{N_3+N_4}{N_3}$ assignments in particular are also equally likely. The fraction of those assignments for which the value of the test statistic is at least as large (in absolute value) as the observed value of the test statistic is the p-value of the null hypothesis that students give the same rating (or none) to an TA, regardless

[5] We would expect it to be a least a little different from zero both because of the luck of the draw in assigning students to sections and because students might rate the two TAs differently, regardless of the TA's perceived gender, and the groups are not all the same size.

of the gender that TA appears to have.

This test is conditional on which of the students are assigned to each of the two TAs, but if we test at level no greater than α conditionally on the assignment, the unconditional level of the resulting test across all assignments is no greater than α, as shown above.

In principle, one could enumerate all the equally likely assignments and compute the value of the test statistic for each, to determine the (conditional) null distribution of the test statistic. In practice, there are prohibitively many assignments (for instance, there are $\binom{23}{11}\binom{24}{11} > 3.3 \times 10^{12}$ possible assignments of 47 students to the 4 TA-led sections that keep constant which students are assigned to each TA). Hence, we estimate p-values by simulation, drawing 10^5 equally likely assignments at random, with one exception, noted below. The distribution of the number of simulated assignments for which the test statistic is greater than or equal to its observed value is Binomial with n equal to the number of simulated assignments and p equal to the true p-value. Hence, the standard error of the estimated p-values is hence no larger than $(1/2)/\sqrt{10^5} \approx 0.0016$. Code for all our analyses is at https://github.com/kellieotto/SET-and-Gender-Bias. Results for the US data are in section 5.

4 The French Natural Experiment

In this section, we test hypotheses about relationships among SET, teaching effectiveness, grade expectations, and student and instructor gender. Our tests aggregate data within course sections, to match how SET are typically used in personnel deci-

sions. We use the average of Pearson correlations across strata as the test statistic,[6] which allows us to test both for differences in means (which can be written as correlations with a dummy variable) and for association with ordinal or quantitative variables.

In these analyses, individual i is a section of a course; the "treatment" is the instructor's gender, the average interim grade, or the average final exam score; and the "response" is the average SET or the average final exam score. Strata consist of all sections of a single course in a single year.

Our tests for overall effects stratify on the course subject, to account for systematic differences across departments: the hypothetical randomization shuffles characteristics among courses in a given department, but not across departments. We also perform tests separately in different departments, and in some cases separately by student gender.

4.1 SET and final exam scores

We test whether average SET scores and average final exam scores for course sections are associated (Table 2). The null hypothesis is that the pairing of average final grade and average SET for sections of a course each year is as if at random, independent across courses and across years. We test this hypothesis overall and separately by discipline, using the average Pearson correlation across strata, as described in section 3.1. If the null hypothesis were true, we would expect the test statistic to be close to zero. On the other hand, if SET do measure teaching effectiveness, we would expect average SET and average final exam score to be positively correlated

[6] As discussed above, we find p-values from the (nonparametric) permutation distribution, not from the theoretical distribution of the Pearson correlation under the parametric assumption of bivariate normality.

within courses within years, making the test statistic positive.

The numbers show that SET scores do not measure teaching effectiveness well, overall: the one-sided p-value for the hypothesis that the correlation is zero is 0.09. Separate tests by discipline find that for History, the association is positive and statistically significant (p-value of 0.01), while the other disciplines (Macroeconomics, Microeconomics, Political science and Sociology), the association is either negative or positive but not statistically significant (p-values 0.19, 0.55, 0.62, and 0.61 respectively).

Table 2: Average correlation between SET and final exam score, by subject

	strata	$\bar{\rho}$	p-value
Overall	26 (21)	0.04	0.09
History	5	0.16	0.01
Political Institutions	5	N/A	N/A
Macroeconomics	5	0.06	0.19
Microeconomics	5	-0.01	0.55
Political science	3	-0.03	0.62
Sociology	3	-0.02	0.61

Note: p-values are one-sided, since, if SET measured teaching effectiveness, mean SET should be positively associated with mean final exam scores. Correlations are computed for course-level averages of SET and final exam score within strata, then averaged across strata. Political Institutions is not reported, because the final exam was not graded anonymously. The five strata of Political Institutions are not included in the overall average, which is computed from the remaining 21 strata-level correlation coefficients.

4.2 SET and Instructor Gender

The second null hypothesis we test is that the pairing (by section) of instructor gender and SET is as if at random within courses each year, independently across years and courses. If gender does not affect SET, we would expect the correlation between average SET and instructor gender to be small in each course in each year. On the other hand, if students tend to rate instructors of one gender higher, we would

expect the average correlation to be large in absolute value. We find that average SET are significantly associated with instructor gender, with male instructors getting higher ratings (overall p-value 0.00). Male instructors get higher SET on average in every discipline (Table 3) with two-sided p-values ranging from 0.08 for History to 0.63 for Political Science.

Table 3: Average correlation between SET and instructor gender

	$\bar{\rho}$	p-value
Overall	0.09	0.00
History	0.11	0.08
Political institutions	0.11	0.10
Macroeconomics	0.10	0.16
Microeconomics	0.09	0.16
Political science	0.04	0.63
Sociology	0.08	0.34

Note: p-values are two-sided.

4.3 Instructor Gender and Learning Outcomes

Do men receive higher SET scores overall because they are better instructors? The third null hypothesis we test is that the pairing (by course) of instructor gender and average final exam score is as if at random within courses each year, independent across courses and across years. If this hypothesis is true, we would expect the average correlations to be small. If the effectiveness of instructors differs systematically by gender, we would expect average correlation to be large in absolute value. Table 4 shows that on the whole, students of male instructors perform worse on the final than students of female instructors, by an amount that is statistically significant (p-value 0.07 overall). In all disciplines, students of male instructors perform worse, but by amounts that are not statistically significant (p-values ranging from 0.22 for History to 0.70 for Political Science). This suggests that male instructors are not

noticeably more effective than female instructors, and perhaps are less effective: The statistically significant difference in SET scores for male and female instructors does not seem to reflect a difference in their teaching effectiveness.

Table 4: Average correlation between final exam scores and instructor gender

	$\bar{\rho}$	p-value
Overall	-0.06	0.07
History	-0.08	0.22
Macroeconomics	-0.06	0.37
Microeconomics	-0.06	0.37
Political science	-0.03	0.70
Sociology	-0.05	0.55

Note: p-values are two-sided. Negative values of $\bar{\rho}$ indicate that students of female instructors did better on average than students of male instructors.

4.4 Gender Interactions

Why do male instructors receive higher SET scores? Separate analyses by student gender shows that male students tend to give higher SET scores to male instructors (Table 5). These permutation tests confirm the results found by Boring [2015a]. Gender concordance is a good predictor of SET scores for men (p-value 0.00 overall). Male students give significantly higher SET scores to male instructors in History (p-value 0.01), Microeconomics (p-value 0.01), Macroeconomics (p-value 0.04), Political Science (p-value 0.06), and Political Institutions (p-value 0.08). Male students give higher SET scores to male instructors in Sociology as well, but the effect is not statistically significant (p-value 0.16).

The correlation between gender concordance and overall satisfaction scores for female students is also positive overall and weakly significant (p-value 0.09). The correlation is negative in some fields (History, Political Institutions, Macroeconomics,

Microeconomics and Sociology) and positive in only one field (Political Science), but in no case statistically significant (p-values range from 0.12 to 0.97).

Table 5: Average correlation between SET and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	0.15	0.00	0.05	0.09
History	0.17	0.01	-0.03	0.60
Political institutions	0.12	0.08	-0.11	0.12
Macroeconomics	0.14	0.04	-0.05	0.49
Microeconomics	0.18	0.01	-0.00	0.97
Political science	0.17	0.06	0.04	0.64
Sociology	0.12	0.16	-0.03	0.76

Note: p-values are two-sided.

Do male instructors receive higher SET scores from male students because their teaching styles match male students' learning styles? If so, we would expect male students of male instructors to perform better on the final exam. However, they do not (Table 6). If anything, male students of male instructors perform worse overall on the final exam (the correlation is negative but not statistically significant, with a p-value 0.75). In History, the amount by which male students of male instructors do worse on the final is significant (p-value 0.03): male History students give significantly higher SET scores to male instructors, despite the fact that they seem to learn more from female instructors. SET do not appear to measure teaching effectiveness, at least not primarily.

4.5 SET and grade expectations

The next null hypothesis we test is that the pairing by course of average SET scores with average interim grades is as if at random. Because interim grades may set student grade expectations, if students give higher SET in courses where they expect

Table 6: Average correlation between student performance and gender concordance

	Male student		Female student	
	$\bar{\rho}$	p-value	$\bar{\rho}$	p-value
Overall	-0.01	0.75	0.06	0.07
History	-0.15	0.03	-0.02	0.74
Macroeconomics	0.04	0.60	0.11	0.10
Microeconomics	0.02	0.80	0.07	0.29
Political science	0.08	0.37	0.11	0.23
Sociology	0.01	0.94	0.06	0.47

Note: p-values are two-sided.

higher grades, the association should be positive. Indeed, the association is positive and generally highly statistically significant (Table 7). Political institutions is the only discipline for which the average correlation between interim grades and SET scores is negative, but the correlation is not significant (p-value 0.61). The estimated p-values for all other courses are between 0.0 and 0.03. The average correlations are especially high in History (0.32) and Sociology (0.24).

Table 7: Average correlation between SET and interim grades

	$\bar{\rho}$	p-value
Overall	0.16	0.00
History	0.32	0.00
Political institutions	-0.02	0.61
Macroeconomics	0.15	0.01
Microeconomics	0.13	0.03
Political science	0.17	0.02
Sociology	0.24	0.00

Note: p-values are one-sided.

In summary, the average correlation between SET and final exam grades (at the level of class sections) is positive, but only weakly significant overall and not significant for most disciplines. However, the average correlation between SET and grade expectations (at the level of class sections) is positive and significant overall

and across most disciplines. The average correlation between instructor gender and SET is statistically significant—male instructors get higher SET—but if anything, students of male instructors do worse on final exams than students of female instructors. Male students tend to give male instructors higher SET, even though they might be learning less than they do from female instructors. We conclude that SET are influenced more by instructor gender and student grade expectations than by teaching effectiveness.

5 The US Randomized Experiment

The previous section suggests that SET have little connection to teaching effectiveness, but the natural experiment does not allow us to control for differences in teaching styles across instructors. MacNell et al. [2014] does. As discussed above, MacNell et al. [2014] collected SET from an online course in which 43 students were randomly assigned to four[7] discussion groups, each taught by one of two TAs, one male and one female. The TAs gave similar feedback to students, returned assignments at exactly the same time, etc.

Biases in student ratings are revealed by differences in ratings each TA received when that TA is identified to the students as male versus as female. MacNell et al. [2014] find that "the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the student ratings index ... Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female, regardless of the actual

[7]As discussed above, there were six sections in all, of which two were taught by the professor and four were taught by TAs.

gender of the assistant instructor." MacNell et al. [2014] used parametric tests whose assumptions did not match their experimental design; part of our contribution is to show that their data admit a more rigorous analysis using permutation tests that honor the underlying randomization and that avoid parametric assumptions about SET. The new analysis supports their overall conclusions, in some cases substantially more strongly than the original analysis (for instance, p-values of 0.01 versus 0.19 for promptness and fairness). In other cases, the original parametric tests overstated the evidence (for instance, a p-value of 0.29 versus 0.04 for knowledgeability).

We use permutation tests as described above in section 3. Individual i is a student; the treatment is the combination of the TA's identity and the TA's apparent gender (there are $K = 4$ treatments). The null hypothesis is that each student would give a TA the same SET score, whether that TA is apparently male or apparently female. A student might give the two TAs different scores, and different students might give different scores to the same TA.

Because of how the experimental randomization was performed, all allocations of students to TA sections that preserve the number of students in each section are equally likely, including allocations that keep the same students assigned to each actual TA constant.

To test whether there is a systematic difference in how students rate apparently male and apparently female TAs, we use the difference in pooled means as our test statistic: We pool the SET for both instructors when they are identified as female and take the mean, pool the SET for both instructors when they are identified as male and take the mean, then subtract the second mean from the first mean (Table 8). This is what MacNell et al. [2014] report as their main result.

As described above, the randomization is stratified and conditions on the set of students allocated to each TA, because, under the null hypothesis, we then know what

SET students would have given for each possible allocation, completely specifying the null distribution of the test statistic. The randomization includes the nonresponders, who are omitted from the averages of the group they are assigned to.

We also perform tests involving the association of concordance of student and apparent TA gender, (Table 9) and SET and concordance of student and actual TA gender (Table 10) using the pooled difference in means as the test statistic. We test the association between grades and actual TA gender (Table 11) using the average Pearson correlation across strata as the test statistic. We find the p-values from the stratified permutation distribution of the test statistic, avoiding parametric assumptions.

5.1 SET and Perceived Instructor Gender

The first hypothesis we test is that students would rate a given TA the same, whether the student thinks the TA is female or male. A positive value of the test statistic means that students give higher SET on average to apparently male instructors. There is weak evidence that the overall SET score depends on the perceived gender (p-value 0.12). The evidence is stronger for several other items students rated: fairness (p-value 0.01), promptness (p-value 0.01), giving praise (p-value 0.01), enthusiasm (p-value 0.06), communication (p-value 0.07), professionalism (p-value 0.07), respect (p-value 0.06), and caring (p-value 0.10). For seven items, the nonparametric permutation p-values are smaller than the parametric p-values reported by MacNell et al. [2014]. Items for which the permutation p-values were greater than 0.10 include clarity, consistency, feedback, helpfulness, responsiveness, and knowledgeability. SET were on a 5-point scale, so a difference in means of 0.80, observed in student ratings of the promptness with which assignments were returned, is 16%

of the full scale—an enormous difference. Since assignments were returned at exactly the same time in all four sections of the class, this seriously impugns the ability of SET to measure even putatively objective characteristics of teaching.

Table 8: Mean ratings and reported instructor gender (male minus female)

	difference in means	nonparametric p-value	MacNell et al. p-value
Overall	0.47	0.12	0.128
Professional	0.61	0.07	0.124
Respectful	0.61	0.06	0.124
Caring	0.52	0.10	0.071
Enthusiastic	0.57	0.06	0.112
Communicate	0.57	0.07	NA
Helpful	0.46	0.17	0.049
Feedback	0.47	0.16	0.054
Prompt	0.80	0.01	0.191
Consistent	0.46	0.21	0.045
Fair	0.76	0.01	0.188
Responsive	0.22	0.48	0.013
Praise	0.67	0.01	0.153
Knowledge	0.35	0.29	0.038
Clear	0.41	0.29	NA

Note: p-values are two-sided.

We also conducted separate tests by student gender. In contrast to our findings for the French data, where male students rated male instructors higher, in the MacNell et al. [2014] experiment, perceived male instructors received significantly higher evaluation scores because female students rated the perceived male instructors higher (Table 9). Male students rated the perceived male instructor significantly (though weakly) higher on only one criterion: fairness (p-value 0.09). Female students, however, rated the perceived male instructor higher on overall satisfaction (p-value 0.11) and most teaching dimensions: praise (p-value 0.01), enthusiasm (p-value 0.05), caring (p-value 0.05), fairness (p-value 0.04), respectfulness (p-value 0.12), communication (p-value 0.10), professionalism (p-value 0.12), and feedback (p-value 0.10).

Female students rate (perceived) female instructors lower on helpfulness, promptness, consistency, responsiveness, knowledge, and clarity, although the differences are not statistically significant.

Table 9: SET and reported instructor gender (male minus female)

	Male students		Female students	
	difference in means	p-value	difference in means	p-value
Overall	0.17	0.82	0.79	0.11
Professional	0.42	0.55	0.82	0.12
Respectful	0.42	0.55	0.82	0.12
Caring	0.04	1.00	0.96	0.05
Enthusiastic	0.17	0.83	0.96	0.05
Communicate	0.25	0.68	0.87	0.10
Helpful	0.46	0.43	0.51	0.35
Feedback	0.08	1.00	0.88	0.10
Prompt	0.71	0.15	0.86	0.13
Consistent	0.17	0.85	0.77	0.17
Fair	0.75	0.09	0.88	0.04
Responsive	0.38	0.54	0.06	1.00
Praise	0.58	0.29	0.81	0.01
Knowledge	0.17	0.84	0.54	0.21
Clear	0.13	0.85	0.67	0.29

Note: p-values are two-sided.

Students of both genders rated the apparently male instructor higher on all dimensions, by an amount that often was statistically significant for female students (Table 9). However, students rated the actual male instructor higher on some dimensions and lower on others, by amounts that generally were not statistically significant (Table 10). The exceptions were praise (p-value 0.02) and responsiveness (p-value 0.05), where female students tended to rate the actual female instructor significantly higher.

Students of the actual male instructor performed worse in the course on average, by an amount that was statistically significant (Table 11). The difference in student performance by perceived gender of the instructor is not statistically significant.

Table 10: SET and actual instructor gender (male minus female)

	Male students		Female students	
	difference in means	p-value	difference in means	p-value
Overall	-0.13	0.61	-0.29	0.48
Professional	0.15	0.96	-0.09	0.73
Respectful	0.15	0.96	-0.09	0.73
Caring	-0.22	0.52	-0.07	0.75
Enthusiastic	-0.13	0.62	-0.44	0.29
Communicate	-0.02	0.80	-0.18	0.61
Helpful	0.03	0.89	0.26	0.71
Feedback	-0.24	0.48	-0.41	0.36
Prompt	-0.09	0.69	-0.33	0.44
Consistent	0.12	0.97	-0.40	0.35
Fair	-0.06	0.71	-0.59	0.12
Responsive	-0.13	0.64	-0.68	0.05
Praise	0.02	0.86	-0.60	0.02
Knowledge	0.22	0.83	-0.44	0.17
Clear	-0.26	0.49	-0.98	0.07

Note: p-values are two-sided.

Table 11: Mean grade and instructor gender (male minus female)

	difference in means	p-value
Perceived	1.76	0.54
Actual	-6.81	0.02

Note: p-values are two-sided.

These results suggest that students rate instructors more on the basis of the instructor's perceived gender than on the basis of the instructor's effectiveness. Students of the TA who is actually female did substantially better in the course, but students rated apparently male TAs higher.

6 Multiplicity

We did not adjust the p-values reported above for multiplicity. We performed a total of approximately 50 tests on the French data, of which we consider four to be our primary results:

1FR lack of association between SET and final exam scores (a negative result, so multiplicity is not an issue)

2FR lack of association between instructor gender and final exam scores (a negative result, so multiplicity is not an issue)

3FR association between SET and instructor gender

4FR association between SET and interim grades

Bonferroni's adjustment for these four tests would leave the last two associations highly significant, with adjusted p-values less than 0.01.

We performed a total of 77 tests on the US data. We consider the three primary null hypotheses to be

1US perceived instructor gender plays no role in SET

2US male students rate perceived male and female instructors the same

3US female students rate perceived male and female instructors the same

To account for multiplicity, we tested these three "omnibus" hypotheses using the nonparametric combination of tests (NPC) method with Fisher's combining function [Pesarin and Salmaso, 2010, Chapter 4] to summarize the 15 dimensions of teaching into a single test statistic that measures how "surprising" the 15 observed differences would be for each of the three null hypotheses. In 10^5 replications, the empirical p-values for these three omnibus hypotheses were 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$), 0.464 (99% confidence interval $[0.460, 0.468]$), and 0 (99% confidence interval $[0.0, 5.3 \times 10^{-5}]$), respectively. (The confidence bounds were obtained by inverting Binomial hypothesis tests.) Thus, we reject hypotheses 1US and 3US.

We made no attempt to optimize the tests to have power against the alternatives considered. For instance, with the US data, the test statistic grouped the two identified-as-female sections and the two identified-as-male conditions, in keeping with how MacNell et al. [2014] tabulated their results, rather than using each TA as his or her own control (although the randomization keeps the two strata intact). Given the relatively small number of students in the US experiment, it is remarkable that *any* of the p-values is small, much less that the p-values for the omnibus tests are effectively zero.

7 Code and Data

Jupyter (`http://jupyter.org/`) notebooks containing our analyses are at `https://github.com/kellieotto/SET-and-Gender-Bias`; they rely on the `permute` Python library (`https://pypi.python.org/pypi/permute/`). The US data are available at `http://n2t.net/ark:/b6078/d1mw2k`. French privacy law prohibits publishing the French data.

8 Discussion

8.1 Other studies

To our knowledge, only two experiments have controlled for teaching style in their designs: Arbuckle and Williams [2003] and MacNell et al. [2014]. In both experiments, students generally gave higher SET when they *thought* the instructor was male, regardless of the actual gender of the instructor. Both experiments found that systematic differences in SET by instructor gender reflect gender bias rather than a match of teaching style and student learning style or a difference in actual teaching effectiveness.

Arbuckle and Williams [2003] showed a group of 352 students "slides of an age- and gender-neutral stick figure and listened to a neutral voice presenting a lecture and then evaluated it on teacher evaluation forms that indicated 1 of 4 different age and gender conditions (male, female, 'old,' and 'young')" [Arbuckle and Williams, 2003, p.507]. All students saw the same stick figure and heard the same voice, so differences in SET could be attributed to the age and gender the students were *told* the instructor had. When students were told the instructor was young and male, students rated the instructor higher than for the other three combinations, especially on "enthusiasm," "showed interest in subject," and "using a meaningful voice tone."

Instructor race is also associated with SET. In the US, SET of instructors of color appear to be biased downwards: minority instructors tend to receive significantly lower SET scores compared to white (male) instructors [Merritt, 2008].[8] Age, [Arbuckle and Williams, 2003], charisma [Shevlin et al., 2000], and physical attractiveness [Riniolo et al., 2006, Hamermesh and Parker, 2005] are also associated

[8]French law does not allow the use of race-related variables in data sets. We were thus unable to test for racial biases in SET using the French data.

with SET. Other factors generally not in the instructor's control that may affect SET scores include class time, class size, mathematical or technical content, and the physical classroom environment [Hill and Epps, 2010].

Many studies cast doubt on the validity of SET as a measure of teaching effectiveness (see Johnson [2003, Chapters 3–5] for a review and analysis, Pounder [2007] for a review, and Galbraith et al. [2012], Carrell and West [2010] for exemplars). Some studies find that gender and SET are not significantly associated [Bennett, 1982, Centra and Gaubatz, 2000, Elmore and LaPointe, 1974] and that SET are valid and reliable measures of teaching effectiveness [Benton and Cashin, 2012, Centra, 1977].[9] The contradictions among conclusions suggests that if SET are ever valid, they are not valid in general: universities should not assume that SET are broadly valid at their institution, valid in any particular department, or valid for any particular course. Given the many sources of bias in SET and the variability in magnitude of the bias by topic, item, student gender, etc., as a practical matter it is impossible to adjust for biases to make SET a valid, useful measure of teaching effectiveness.

8.2 Summary

We used permutation tests to examine data collected by Boring [2015a] and MacNell et al. [2014], both of which find that gender biases prevent SET from measuring teaching effectiveness accurately and fairly. SET are more strongly related to instructor's perceived gender and to students' grade expectations than they are to learning, as measured by performance on anonymously graded, uniform final exams. The extent and direction of gender biases depend on context, so it is impossible to adjust for such biases to level the playing field. While the French university data show a

[9]Some authors who claim that SET are valid have a financial interest in developing SET instruments and conducting SET.

positive male student bias for male instructors, the experimental US setting suggests a positive female student bias for male instructors. The biases in the French university data vary by course topic; the biases in the US data vary by item. We would also expect the bias to depend on class size, format, level, physical characteristics of the classroom, instructor ethnicity and a host of other variables.

We do not claim that there is *no* connection between SET and student performance. However, the observed association is sometimes positive and sometimes negative, and in general is not statistically significant—in contrast to the statistically significant strong associations between SET and grade expectations and between SET and instructor gender. SET appear to measure student satisfaction and grade expectations more than they measure teaching effectiveness [Stark and Freishtat, 2014, Johnson, 2003]. While student satisfaction may *contribute* to teaching effectiveness, it is not itself teaching effectiveness. Students may be satisfied or dissatisfied with courses for reasons unrelated to learning outcomes—and not in the instructor's control (e.g., the instructor's gender).

In the US, SET have two primary uses: instructional improvement and personnel decisions, including hiring, firing, and promoting instructors. We recommend caution in the first use, and discontinuing the second use, given the strong student biases that influence SET, even on "objective" items such as how promptly instructors return assignments.[10]

[10]In 2009, the French Ministry of Higher Education and Research upheld a 1997 decision of the French State Council that public universities can use SET only to help tenured instructors improve their pedagogy, and that the administration may not use SET in decisions that might affect tenured instructors' careers (c.f. Boring [2015b]).

8.3 Conclusion

In two very different universities and in a broad range of course topics, SET measure students' gender biases better than they measure the instructor's teaching effectiveness. Overall, SET disadvantage female instructors. There is no evidence that this is the exception rather than the rule. Hence, the onus should be on universities that rely on SET for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, under-represented minorities, or other protected groups. Because the bias varies by course and institution, affirmative evidence needs to be specific to a given course in a given department in a given university. Absent such specific evidence, SET should not be used for personnel decisions.

References

J. Arbuckle and B. D. Williams. Students' Perceptions of Expressiveness : Age and Gender Effects on Teacher Evaluations. *Sex Roles*, 49(November):507–516, 2003.

S. K. Bennett. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74(2):170–179, 1982.

S. L. Benton and W. E. Cashin. Student ratings of teaching: A summary of research and literature. IDEA Paper 50, The IDEA Center, 2012.

A. Boring. Gender biases in student evaluations of teachers. Document de travail OFCE 13, OFCE, April 2015a.

A. Boring. Can students evaluate teaching quality objectively? Le blog de l'ofce, OFCE, 2015b. URL http://www.ofce.sciences-po.fr/blog/can-students-evaluate-teaching-quality-objectively/.

M. Braga, M. Paccagnella, and M. Pellizzari. Evaluating students evaluations of professors. *Economics of Education Review*, 41:71–88, 2014.

S. E. Carrell and J. E. West. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3):409–432, June 2010. ISSN 0022-3808. doi: 10.1086/653808. URL http://www.jstor.org/stable/10.1086/653808.

J. A. Centra. Student ratings of instruction and their relationship to student learning. *American educational research journal*, 14(1):17–24, 1977.

J. A. Centra and N. B. Gaubatz. Is There Gender Bias in Student Evaluations of Teaching? *Journal of Higher Education*, 71(1):17–33, 2000. URL http://www.jstor.org/stable/10.2307/2649280.

P. B. Elmore and K. A. LaPointe. Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology*, 66(3):386–389, 1974.

C. S. Galbraith, G. B. Merrill, and D. M. Kline. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? a neural network and bayesian analyses. *Research in Higher Education*, 53(3):353–374, 2012.

D. S. Hamermesh and A. Parker. Beauty in the classroom: Instructors pulchritude

and putative pedagogical productivity. *Economics of Education Review*, 24(4): 369–376, 2005.

M. C. Hill and K. K. Epps. The impact of physical classroom environment on student satisfaction and student evaluation of teaching in the university environment. *Academy of Educational Leadership Journal*, 14(4):65–79, 2010.

V. E. Johnson. *Grade Inflation: A Crisis in College Education*. Springer-Verlag, New York, 2003.

L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.

H. W. Marsh and L. A. Roche. Making Students' Evaluations of Teaching Effectiveness Effective. *American Psychologist*, 52(11):1187–1197, 1997.

D. J. Merritt. Bias, the brain, and student evaluations of teaching. *St. John's Law Review*, 81(1):235–288, 2008.

J. Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley, New York, 2010.

J. S. Pounder. Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. *Quality Assurance in Education*, 15(2):178–191, 2007. ISSN 0968-4883. doi: 10.1108/09684880710748938. URL http://www.emeraldinsight.com/10.1108/09684880710748938.

T. C. Riniolo, K. C. Johnson, T. R. Sherman, and J. A. Misso. Hot or not: do professors perceived as physically attractive receive higher student evaluations? *The Journal of general psychology*, 133(1):19–35, Jan. 2006. ISSN 0022-1309. doi: 10.3200/GENP.133.1.19-35. URL http://www.ncbi.nlm.nih.gov/pubmed/16475667.

M. Shevlin, P. Banyard, M. Davies, and M. Griffiths. The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4):397–405, 2000.

P. B. Stark and R. Freishtat. An evaluation of course evaluations. *Science Open Research*, 2014. doi: 10.14293/S2199-1006.1.-.AOFRQA.v1. URL https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4.

Appendix D

(https://www.insidehighered.com)

Study finds gender perception affects evaluations

Submitted by Kaitlin Mulhere on December 10, 2014 - 3:00am

College students' assessments of their instructors' teaching ability is linked to whether they think those instructors are male or female, according to new research from North Carolina State University.

In the study [1], students in an online course gave better evaluations to the instructors they thought were male, even though the two instructors – one male and one female – had switched their identities. The research is based on a small pilot study of one class.

Student evaluations can carry a lot of weight in decisions about promotions, tenure and pay raises. But the findings demonstrate that gender bias can have a big impact on student ratings of teachers, according to the study.

To conduct the study, researchers compared instructor evaluations of four discussion groups in a technology and society class within the sociology and anthropology department at North Carolina State. Two groups were taught by a female instructor and two were taught by a male instructor. Students in one of the female instructor's groups were told their instructor was male, and vice versa.

Neither the actual male nor actual female instructor received significantly higher ratings. But the same instructors received different ratings when they "switched" genders. The male instructor had lower ratings when students were told their instructor was female. The female instructor had higher ratings when students were told their instructor was male.

The authors say that their findings suggest a female instructor would have to work harder than a male to receive comparable ratings. The study was published this month in the journal *Innovative Higher Education.*

The lead author, Lillian MacNell, said personal experience encouraged her to conduct the study. (The co-authors are Adam Driscoll, an assistant professor at the University of Wisconsin-La Crosse and Andrea Hunt, an assistant professor at the University of North Alabama. Driscoll and Hunt earned their Ph.D.s from N.C. State.)

MacNell, a doctoral student at North Carolina State, was grading for an online course and often received emails from students challenging her decisions. They complained about her grading,

and in some cases, went over her head and emailed the professor directly.

She vented to a male colleague who was also grading for the course, saying that it was frustrating how much students were protesting her decisions.

“He said, ‘What are you talking about?’ He hadn’t received anything.”

Both MacNell and the male colleague were using the same language and rubric to grade students, she said, so there was no reason why students should be accepting his decisions but not hers.

The study cites previous research that has found gender bias in students' evaluations of articles -- identical articles were ranked higher if they had male names -- and in students' judgment of faculty qualifications -- male candidates were judged as more qualified despite identical credentials.

In teaching evaluations, previous studies have focused on how female instructors are expected to be nurturing and supportive; when they’re not, it may count against them in evaluations. At the same time, if they are nurturing and supportive, female instructors risk being perceived as less authoritative and knowledgeable than their male counterparts, according to the study.

But it's difficult when evaluating teaching to isolate the instructor's gender from other factors that influence class instruction, such as teaching style. So while previous research has revealed differences between instructor evaluations for each gender, it hasn't determined whether those differences were the result of gender bias, according to the authors.

That’s where online education comes in. Unlike those doing research on face-to-face instruction, researchers in this study were able to hide the gender of the instructors, and to keep equal all the teaching components, such as grading standards and the speed of responses.

When comparing the evaluations of the perceived gender identities, the male identity received higher scores across all 12 variables students evaluated. In six variables -- professionalism, promptness, fairness, respectfulness, enthusiasm and giving praise -- the differences were statistically significant.

In promptness, for example, the instructors matched their grading schedules so that students in all groups received feedback at about the same rate. The instructor whom students thought was male was graded a 4.35 out of 5 for promptness, while the instructor perceived to be female received a 3.55.

With just 43 subjects, this study was a pilot; the authors plan to expand their research with more classes and different types of courses. Still, higher education administrators should be aware of the findings when using evaluations to make faculty decisions, since evaluations could reflect a gender bias rather than an actual difference in teaching abilities, MacNell said.

Faculty [2]

Links:
[1] http://link.springer.com/article/10.1007/s10755-014-9313-4
[2] https://www.insidehighered.com/news/news-sections/faculty

undefined
undefined

Appendix E

An Evaluation of Course Evaluations

Published in *ScienceOpen*:
https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4?0

DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Philip B. Stark*

Department of Statistics, University of California, Berkeley
Berkeley, CA 94720, United States
stark@stat.berkeley.edu

Richard Freishtat

Center for Teaching and Learning, University of California, Berkeley
Berkeley, CA 94720, United States
rfreishtat@berkeley.edu

26 September 2014

* Corresponding author. E-mail: stark@stat.berkeley.edu

26 September 2014

Student ratings of teaching have been used, studied, and debated for almost a century. This article examines student ratings of teaching from a statistical perspective. The common practice of relying on averages of student teaching evaluation scores as the primary measure of teaching effectiveness for promotion and tenure decisions should be abandoned for substantive and statistical reasons: There is strong evidence that student responses to questions of "effectiveness" do not measure teaching effectiveness. Response rates and response variability matter. And comparing averages of categorical responses, even if the categories are represented by numbers, makes little sense. Student ratings of teaching are valuable when they ask the right questions, report response rates and score distributions, and are balanced by a variety of other sources and methods to evaluate teaching.

Since 1975, course evaluations at *University of California, Berkeley* have asked:

> Considering both the limitations and possibilities of the subject matter and course, how would you rate the overall teaching effectiveness of this instructor?
>
> 1 (not at all effective), 2, 3, 4 (moderately effective), 5, 6, 7 (extremely effective)

Among faculty, student evaluations of teaching (SET) are a source of pride and satisfaction—and frustration and anxiety. High-stakes decisions including tenure and promotions rely on SET. Yet it is widely believed that they are primarily a popularity contest; that it's easy to "game" ratings; that good teachers get bad ratings and *vice versa*; and that rating anxiety stifles pedagogical innovation and encourages faculty to water down course content. What's the truth?

We review statistical issues in analyzing and comparing SET scores, problems defining and measuring teaching effectiveness, and pernicious distortions that result from using SET scores as a proxy for teaching quality and effectiveness. We argue here--and the literature shows--that students are in a good position to evaluate *some* aspects of teaching, but SET are at best tenuously connected to teaching effectiveness (Defining and measuring teaching effectiveness are knotty problems in themselves; we discuss this below). Other ways of evaluating teaching can be combined with student comments to produce a more reliable and meaningful composite. We make recommendations regarding the use of SET and discuss new policies implemented at *University of California, Berkeley*, in 2013.

Background

SET scores are the most common method to evaluate teaching (Cashin, 1999; Clayson, 2009; Davis, 2009; Seldin, 1999). They define "effective teaching" for many purposes. They are popular partly because the measurement is easy and takes little class or faculty time. Averages of SET ratings have an air of objectivity simply by virtue of being numerical. And comparing an instructor's average rating to departmental averages is simple. However, questions about using SET as the sole source of evidence about teaching for merit and promotion, and the efficacy of evaluation questions and methods of interpretation persist (Pounder, 2007).

Statistics and SET

Who responds?

Some students do not fill out SET surveys. The *response rate* will be less than 100%. The lower the response rate, the less representative the responses might be: there's no reason nonresponders should be like responders--and good reasons they might not be. For instance, anger motivates people to action more than satisfaction does. Have you ever seen a public demonstration where people screamed "we're content!"? (See, e.g., http://xkcd.com/470/)

Nonresponse produces uncertainty: Suppose half the class responds, and that they rate the instructor's handwriting legibility as 2. The average for the entire class might be as low as 1.5, if all the "nonresponders" would also have rated it 1. Or it might be as high as 4.5, if the nonresponders would have rated it 7.

Some schools require faculty to explain low response rates. This seems to presume that it is the instructor's fault if the response rate is low, and that a low response rate is in itself a sign of bad teaching. Consider these scenarios:

(1) The instructor has invested an enormous amount of effort in providing the material in several forms, including online materials, online self-test exercises, and webcast lectures; the course is at 8am. We might expect attendance and response rates to in-class evaluations to be low.

(2) The instructor is not following any text and has not provided notes or

supplementary materials. Attending lecture is the only way to know what is covered. We might expect attendance and response rates to in-class evaluations to be high.

(3) The instructor is exceptionally entertaining, gives "hints" in lecture about exams; the course is at 11am. We might expect high attendance and high response rates for in-class evaluations.

The point: Response rates themselves say little about teaching effectiveness. In reality, if the response rate is low, the data should not be considered representative of the class as a whole. An explanation solves nothing.

Averages of small samples are more susceptible to "the luck of the draw" than averages of larger samples. This can make SET in small classes more extreme than evaluations in larger classes, even if the response rate is 100%. And students in small classes might imagine their anonymity to be more tenuous, perhaps reducing their willingness to respond truthfully or to respond at all.

Averages

Personnel reviews routinely compare instructors' average scores to departmental averages. Such comparisons make no sense, as a matter of Statistics. They presume that the difference between 3 and 4 means the same thing as the difference between 6 and 7. They presume that the difference between 3 and 4 means the same thing to different students. They presume that 5 means the same thing to different students and to students in different courses.

They presume that a 3 “balances” a 7 to make two 5s. For teaching evaluations, there’s no reason any of those things should be true (See, e.g., McCullough & Radson, 2011).

SET scores are *ordinal categorical* variables: The ratings fall in categories that have a natural order, from worst (1) to best (7). But the numbers are *labels*, not *values*. We could replace the numbers with descriptions and no information would be lost: The ratings might as well be “not at all effective,” … , “extremely effective.” It doesn’t make sense to average labels. Relying on averages equates two ratings of 5 with ratings of 3 and 7, since both sets average to 5.

They are not equivalent, as this joke shows: Three statisticians go hunting. They spot a deer. The first statistician shoots; the shot passes a yard to the left of the deer. The second shoots; the shot passes a yard to the right of the deer. The third one yells, “we got it!”

Scatter matters

Comparing an individual instructor’s average with the average for a course or a department is meaningless: Suppose that the departmental average for a particular course is 4.5, and the average for a particular instructor in a particular semester is 4.2. The instructor’s rating is below average. How bad is that?

If other instructors get an average of exactly 4.5 when they teach the course, 4.2 might be atypically low. On the other hand, if other instructors get 6s half the time and 3s half the time, 4.2 is well within the spread of scores. Even if

averaging made sense, the mere fact that one instructor's average rating is above or below the departmental average says little. We should report the *distribution* of scores for instructors and for courses: the percentage of ratings in each category (1–7). The distribution is easy to convey using a bar chart.

All the children are above average

At least half the faculty in any department will have average scores at or below median for that department. Deans and Chairs sometimes argue that a faculty member with below-average teaching evaluations is an excellent teacher—just not as good as the other, superlative teachers in that department. With apologies to Garrison Keillor, all faculty members in all departments cannot be above average.

Comparing incommensurables

Students' interest in courses varies by course type (e.g., prerequisite versus major elective). The nature of the interaction between students and faculty varies with the type and size of courses. Freshmen have less experience than seniors. These variations are large and may be confounded with SET (Cranton & Smith, 1986; Feldman, 1984, 1978). It is not clear how to make fair comparisons of SET across seminars, studios, labs, prerequisites, large lower-division courses, required major courses, etc (See, e.g., McKeachie, 1997).

Student Comments

Students are ideally situated to comment *about their experience* of the

course, including factors that influence teaching effectiveness, such as the instructor's audibility, legibility, and perhaps the instructor's availability outside class. They can comment on whether they feel more excited about the subject after taking the class, and—for electives—whether the course inspired them to take a follow-up course. They might be able to judge clarity, but clarity may be confounded with the difficulty of the material. While some student comments are informative, one must be quite careful interpreting the comments: faculty and students use the same vocabulary quite differently, ascribing quite different meanings to words such as "fair," "professional," "organized," "challenging," and "respectful" (Lauer, 2012). Moreover, it is not easy to compare comments across disciplines (Cashin, 1990; Cashin & Clegg, 1987; Cranton & Smith, 1986; Feldman, 1978), because the depth and quality of students' comments vary widely by discipline. In context, these comments are all glowing:

Physical Sciences class.

> "Lectures are well organized and clear"
>
> "Very clear, organized and easy to work with"

Humanities class.

> "Before this course I had only read two plays because they were required in High School. My only expectation was to become more familiar with

the works. I did not expect to enjoy the selected texts as much as I did, once they were explained and analyzed in class. It was fascinating to see texts that the author's were influenced by; I had no idea that such a web of influence in Literature existed. I wish I could be more 'helpful' in this evaluation, but I cannot. I would not change a single thing about this course. I looked forward to coming to class everyday. I looked forward to doing the reading for this class. I only wish that it was a year long course so that I could be around the material, graduate instructor's and professor for another semester."

What SET Measure

If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.

-D. Huff (1954)

This is what we do with SET. We don't measure teaching effectiveness. We measure what students say, and pretend it's the same thing. We calculate statistics, report numbers, and call it a day.

What is effective teaching? One definition is that an effective teacher is skillful at creating conditions conducive to learning. Some learning happens no matter what the instructor does. Some students do not learn much no matter what the instructor does. How can we tell how much the instructor helped or hindered?

Measuring learning is hard: Grades are poor proxies, because courses and exams can be easy or hard (Beleche, Fairris and Marks, 2012). If exams were set by someone other than the instructor—as they are in some universities—we might be able to use exam scores to measure learning (See, e.g., http://xkcd.com/135/). But that's not how most universities work, and teaching to the test could be confounded with learning.

Performance in follow-on courses and career success may be better measures, but those measurements are hard to make. And how much of someone's career success can be attributed to a given course, years later?

There is a large research literature on SET, most of which addresses *reliability*: Do different students give the same instructor similar marks (See, e.g., Abrami, et al., 2001; Braskamp and Ory, 1994; Centra, 2003; Ory, 2001; Wachtel, 1998; Marsh and Roche, 1997)? Would a student rate the same instructor consistently later (See, e.g., Braskamp and Ory, 1994; Centra, 1993; Marsh, 2007; Marsh and Dunkin, 1992; Overall and Marsh, 1980)? That has nothing to do with

whether SET measure effectiveness. A hundred bathroom scales might all report your weight to be the same. That doesn't mean the readings are accurate measures of your *height*—or even your weight, for that matter.

Moreover, inter-rater reliability is an odd thing to worry about, in part because it's easy to report the full distribution of student ratings, as advocated above. Scatter matters, and it can be measured *in situ* in every course.

Observation versus Randomization

Most of the research on SET is based on *observational studies*, not *experiments*. In the entire history of Science, there are few observational studies that justify inferences about causes (A notable exception is John Snow's research on the cause of cholera; his study amounts to a "natural experiment." See http://www.stat.berkeley.edu/~stark/SticiGui/Text/experiments.htm#cholera for a discussion). In general, to infer causes, such as whether good teaching results in good evaluation scores, requires a *controlled, randomized experiment*: individuals are assigned to groups at random; the groups get different *treatments*; the outcomes are compared statistically across groups to test whether the treatments have different effects and to estimate the sizes of those differences.

Randomized experiments use a blind, non-discretionary chance mechanism to assign treatments to individuals. Randomization tends to mix individuals across groups in a balanced way. Absent randomization, other things can *confound* the effect of the treatment (See, e.g., http://xkcd.com/552/).

For instance, suppose some students choose classes by finding the professor reputed to be the most lenient grader. Such students might then rate that professor highly for an "easy A." If those students choose sequel courses the same way, they may get good grades in those easy classes too, "proving" that the first ratings were justified.

The best way to reduce confounding is to assign students randomly to classes. That tends to mix students with different abilities and from easy and hard sections of the prequel across sections of sequels. This experiment has been done at the U.S. Air Force Academy (Carrell and West, 2008) and Bocconi University in Milan, Italy (Braga, Paccagnella, and Pellizzari, 2011).

These experiments found that teaching effectiveness, as measured by subsequent performance and career success, is *negatively* associated with SET scores. While these two student populations might not be representative of all students, the studies are the best we have seen. And their findings are concordant.

What do student teaching evaluations measure?

SET may be *reliable*, in the sense that students often agree (Braskamp and Ory, 1994; Centra, 1993; Marsh, 2007; Marsh and Dunkin, 1992; Overall and Marsh, 1980). But that's an odd focus. We don't expect instructors to be equally effective with students with different background, preparation, skill, disposition, maturity, and "learning style." Hence, if ratings are extremely consistent, they probably don't measure teaching effectiveness: If a laboratory instrument always

gives the same reading when its inputs vary substantially, it's probably broken. There is no consensus on what SET do measure:

- SET scores are highly correlated with students' grade expectations (Marsh and Cooper, 1980; Short et al., 2012; Worthington, 2002)
- SET scores and enjoyment scores are related (In the UC Berkeley Department of Statistics in fall 2012, for the 1486 students who rated the instructor's overall effectiveness and their enjoyment of the course, the correlation between instructor effectiveness and course enjoyment was 0.75, and the correlation between course effectiveness and course enjoyment was 0.8.)
- SET can be predicted from the students' reaction to 30 seconds of silent video of the instructor; physical attractiveness matters (Ambady and Rosenthal, 1993).
- gender, ethnicity, and the instructor's age matter (Anderson and Miller, 1997; Basow, 1995; Cramer and Alexitch, 2000; Marsh and Dunkin, 1992; Wachtel, 1998; Weinberg et al., 2007; Worthington, 2002).
- omnibus questions about curriculum design, effectiveness, etc. appear most influenced by factors unrelated to learning (Worthington, 2002)

<u>What good are SET?</u>

Students are in a good position to observe some aspects of teaching, such as clarity, pace, legibility, audibility, and their own excitement (or boredom).

SET can measure these things; the statistical issues raised above still matter, as do differences between how students and faculty use the same words (Lauer, 2012).

But students cannot rate effectiveness--regardless of their intentions. Calling SET a measure of effectiveness does not make it one, any more than you can make a bathroom scale measure height by relabeling its dial "height." Averaging "height" measurements made with 100 different scales would not help.

What's better?

Let's drop the pretense. We will never be able to measure teaching effectiveness reliably and routinely. In some disciplines, measurement is possible but would require structural changes, randomization, and years of follow-up.

If we want to assess and improve teaching, we have to pay attention to the teaching, not the average of a list of student-reported numbers with a troubled and tenuous relationship to teaching. Instead, we can watch each other teach and talk to each other about teaching. We can look at student comments. We can look at materials created to design, redesign, and teach courses, such as syllabi, lecture notes, websites, textbooks, software, videos, assignments, and exams. We can look at faculty teaching statements. We can look at samples of student work. We can survey former students, advisees, and graduate instructors. We can look at the job placement success of former graduate students. Etc.

We can ask: Is the teacher putting in appropriate effort? Is she following

practices found to work in the discipline? Is she available to students? Is she creating new materials, new courses, or new pedagogical approaches? Is she revising, refreshing, and reworking existing courses? Is she helping keep the curriculum in the department up to date? Is she trying to improve? Is she supervising undergraduates for research, internships, and honors theses? Is she advising graduate students? Is she serving on qualifying exams and thesis committees? Do her students do well when they graduate?

Or, is she "checked out"? Does she use lecture notes she inherited two decades ago the first time she taught the course? Does she mumble, facing the board, scribbling illegibly? Do her actions and demeanor discourage students from asking questions? Is she unavailable to students outside of class? Does she cancel class frequently? Does she return student work with helpful comments? Does she refuse to serve on qualifying exams or dissertation committees?

In 2013, the University of California, Berkeley Department of Statistics adopted as standard practice a more holistic assessment of teaching. Every candidate is asked to produce a teaching portfolio for personnel reviews, consisting of a teaching statement, syllabi, notes, websites, assignments, exams, videos, statements on mentoring, or any other materials the candidate feels are relevant. The chair and promotion committee read and comment on the portfolio in the review. At least before every "milestone" review (mid-career, tenure, full, step VI), a faculty member attends at least one of the candidate's lectures and

comments on it, in writing. These observations complement the portfolio and student comments. Distributions of SET scores are reported, along with response rates. Averages of scores are not reported.

Classroom observation took the reviewer about four hours, including the observation time itself. The process included conversations between the candidate and the observer, the opportunity for the candidate to respond to the written comments, and a provision for a "no-fault do-over" at the candidate's sole discretion. The candidates and the reviewer reported that the process was valuable and interesting. Based on this experience, the Dean of the Division now recommends peer observation prior to milestone reviews.

Observing more than one class session and more than one course would be better. Adding informal classroom observation and discussion between reviews would be better. Periodic surveys of former students, advisees, and teaching assistants would bring another, complementary source of information about teaching. But we feel that using teaching portfolios and even a little classroom observation improves on SET alone.

The following sample letter is a redacted amalgam of chair's letters submitted with merit and promotion cases since the Department of Statistics adopted a policy of more comprehensive assessment of teaching, including peer observation:

Smith is, by all accounts, an excellent teacher, as confirmed by the

classroom observations of Professor Jones, who calls out Smith's ability to explain key concepts in a broad variety of ways, to hold the attention of the class throughout a 90-minute session, to use both the board and slides effectively, and to engage a large class in discussion. Prof. Jones's peer observation report is included in the case materials; conversations with Jones confirm that the report is Jones's candid opinion: Jones was impressed, and commented in particular on Smith's rapport with the class, Smith's sensitivity to the mood in the room and whether students were following the presentation, Smith's facility in blending derivations on the board with projected computer simulations to illustrate the mathematics, and Smith's ability to construct alternative explanations and illustrations of difficult concepts when students did not follow the first exposition.

While interpreting "effectiveness" scores is problematic, Smith's teaching evaluation scores are consistently high: in courses with a response rate of 80% or above, less than 1% of students rate Smith below a 6.

Smith's classroom skills are evidenced by student comments in teaching evaluations and by the teaching materials in her portfolio.

Examples of comments on Smith's teaching include:

> *I was dreading taking a statistics course, but after this class, I decided to major in statistics.*
>
> *the best I've ever met...hands down best teacher I've had in 10 years of university education*
>
> *overall amazing…she is the best teacher I have ever had*
>
> *absolutely love it*
>
> *loves to teach, humble, always helpful*
>
> *extremely clear … amazing professor*
>
> *awesome, clear*
>
> *highly recommended*
>
> *just an amazing lecturer*
>
> *great teacher … best instructor to date*
>
> *inspiring and an excellent role model*
>
> *the professor is GREAT*

Critical student comments primarily concerned the difficulty of the material or the homework. None of the critical comments reflected on the pedagogy or teaching effectiveness, only the workload.

I reviewed Smith's syllabus, assignments, exams, lecture notes, and other materials for Statistics X (a prerequisite for many majors), Y (a seminar course she developed), Z (a graduate course she developed for the revised MA program, which she has spearheaded), and Q (a topics course in her research area). They are very high quality and clearly the result of considerable thought and effort.

In particular, Smith devoted an enormous amount of time to developing online materials for X over the last five years. The materials required designing and creating a substantial amount of supporting technology, representing at least 500 hours per year of effort to build and maintain. The undertaking is highly creative and advanced the state of the art. Not only are those online materials superb, they are having an impact on pedagogy elsewhere: a Google search shows over 1,200 links to those materials, of which more than half are from other countries. I am quite impressed with the pedagogy, novelty, and functionality. I have a few minor suggestions about the content, which I will discuss with Smith, but those are a matter of taste, not of correctness.

The materials for X and Y are extremely polished. Notably, Smith assigned a term project in an introductory course, harnessing the power of inquiry-based learning. I reviewed a handful of the term projects, which were ambitious and impressive. The materials for Z and Q are also well organized and interesting, and demand an impressively high level of performance from the students. The materials for Q include a great selection of data sets and computational examples that are documented well. Overall, the materials are exemplary; I would estimate that they represent well over 1,500 hours of development during the review period.

Smith's lectures in X were webcast in fall, 2013. I watched portions of a dozen of Smith's recorded lectures for X—a course I have taught many times. Smith's lectures are excellent: clear, correct, engaging, interactive, well paced, and with well organized and legible boardwork. Smith does an excellent job keeping the students involved in discussion, even in large (300+ student) lectures. Smith is particularly good at keeping the students thinking during the lecture and of inviting questions and comments. Smith responds generously and sensitively to questions, and is tuned in well to the mood of the class.

Notably, some of Smith's lecture videos have been viewed nearly 300,000 times! This is a testament to the quality of Smith's pedagogy and reach. Moreover, these recorded lectures increase the visibility of the Department and the University, and have garnered unsolicited effusive thanks and praise from across the world.

Conversations with teaching assistants indicate that Smith spent a considerable amount of time mentoring them, including weekly meetings and observing their classes several times each semester. She also played a leading role in revising the PhD curriculum in the department.

Smith has been quite active as an advisor to graduate students. In addition to serving as a member of sixteen exam committees and more than a dozen MA and PhD committees, she advised three PhD recipients (all of whom got jobs in top-ten departments), co-advised two others, and is currently advising three more. Smith advised two MA recipients who went to jobs in industry, co-advised another who went to a job in government, advised one who changed advisors. Smith is currently advising a fifth. Smith supervised three undergraduate honors theses and two undergraduate internships during the review period.

This is an exceptionally strong record of teaching and mentoring for an assistant professor. Prof. Smith's teaching greatly exceeds expectations.

We feel that a review along these lines would better reflect whether faculty are dedicated teachers, the effort they devote, and the effectiveness their teaching; would comprise a much fairer assessment; and would put more appropriate attention on teaching.

Recap

- SET does not measure teaching effectiveness.
- Controlled, randomized experiments find that SET ratings are negatively associated with direct measures of effectiveness. SET seem to be influenced by the gender, ethnicity, and attractiveness of the instructor.
- Summary items such as "overall effectiveness" seem most influenced by

irrelevant factors.

- Student comments contain valuable information about students' *experiences*.
- Survey response rates matter. Low response rates make it impossible to generalize reliably from the respondents to the whole class.
- It is practical and valuable to have faculty observe each other's classes.
- It is practical and valuable to create and review teaching portfolios.
- Teaching is unlikely to improve without serious, regular attention.

Recommendations

1. Drop omnibus items about "overall teaching effectiveness" and "value of the course" from teaching evaluations: They are misleading.
2. Do not average or compare averages of SET scores: Such averages do not make sense statistically. Instead, report the distribution of scores, the number of responders, and the response rate.
3. When response rates are low, extrapolating from responders to the whole class is unreliable.
4. Pay attention to student comments—but understand their limitations. Students typically are not well situated to evaluate pedagogy.
5. Avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines.
6. Use teaching portfolios as part of the review process.

7. Use classroom observation as part of milestone reviews.
8. To improve teaching and evaluate teaching fairly and honestly, spend more time observing the teaching and looking at teaching materials.

References

Abrami, P.C., Marilyn, H.M. & Raiszadeh, F. (2001). Business students' perceptions of faculty evaluations. *The International Journal of Educational Management, 15*(1), 12–22.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*(3), 431.

Anderson, K., & Miller, E.D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics, 30*(2), 216-219.

Basow, S.A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656-665.

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review*, *31*(5), 709-719.

Braga, M., Paccagnella, M., & Pellizzari, M. (2011). Evaluating students' evaluations of professors. *Bank of Italy Temi di Discussione (Working Paper) No*, *825*.

Braskamp, L.A., & Ory, J.C. (1994). *Assessing faculty work: Enhancing individual and institutional performance.* San Francisco: Jossey-Bass.

Carrell, S.E., & West, J.E. (2008). *Does professor quality matter? Evidence from random assignment of students to professors* (No. w14081). National

Bureau of Economic Research.

Cashin, W.E. (1990). Students do rate different academic fields differently. In M. Theall and J. Franklin (eds.), *Student Ratings of Instruction: Issues for improving practice*. San Francisco: Jossey-Bass Inc.

Cashin, W.E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions.* Bolton, MA: Anker.

Cashin, W.E. and Clegg, V.L. (1987). *Are student ratings of different academic fields different?* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

Centra, J.A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less coursework? *Research in Higher Education*, *44*(5), 495-518.

Clayson, D.E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16-30.

Cramer, K.M. & Alexitch, L.R. (2000). Student evaluations of college professors: identifying sources of bias. *Canadian Journal of Higher Education, 30*(2),

143-64.

Cranton, P.A. and Smith, R.A. (1986). A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal, 23*(1), 117–128.

Davis, B.G. (2009). *Tools for Teaching, 2nd edition*. San Francisco, CA: John Wiley & Sons.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't know. *Research in Higher Education, 9*, 199–242.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*(11), 45–116.

Huff, D. (1954). *How To Lie With Statistics*, New York: W.W. Norton.

Lauer, C. (2012). A Comparison of Faculty and Student Perspectives on Course Evaluation Terminology. In *To Improve the Academy: Resources for Faculty, Instructional, and Organizational Development*, edited by J. Groccia & L. Cruz, 195-212. San Francisco, CA: Wiley & Sons, Inc.

Marsh. H.W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective,* 319–383. Dordrecht, The Netherlands: Springer.

Marsh, H.W., & Cooper, T. (1980) *Prior subject interest, students evaluations, and instructional effectiveness* Paper presented at the annual meeting of the American Educational Research Association.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*, Vol. 8. New York: Agathon Press.

Marsh, H.W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187–1197.

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation & Research in Education*, *24*(3), 183–202.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225.

Ory, J.C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning, 87*, 3–15.

Overall, J.U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*, 321–325.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile?: An

analytical framework for answering the question. *Quality Assurance in Education*, *15*(2), 178-191.

Seldin, P. (1999). Building successful teaching evaluation programs. In P. Seldin (ed.), *Current practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Short, H., Boyle, R., Braithwaite, R., Brookes, M., Mustard, J., & Saundage, D. (2008). A comparison of student evaluation of teaching with student performance. In *OZCOTS 2008: Proceedings of the 6th Australian Conference on Teaching Statistics* (pp. 1–10).

Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191–211.

Weinberg, B.A., Fleisher, B.M., & Hashimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education (NBER Working Paper No. 12844)*. Retrieved 5 August 2013 from http://www.nber.org/papers/w12844http://www.nber.org/papers/w12844

Worthington, A.C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education, *Assessment and Evaluation in Higher Education*, *27*(1), 49–64.

Appendix F

Innov High Educ (2015) 40:291–303
DOI 10.1007/s10755-014-9313-4

What's in a Name: Exposing Gender Bias in Student Ratings of Teaching

Lillian MacNell · Adam Driscoll · Andrea N. Hunt

Published online: 5 December 2014

Abstract Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention.

Keywords gender inequality · gender bias · student ratings of teaching · student evaluations of instruction

Lillian MacNell is a doctoral candidate in Sociology at North Carolina State University. She received her Master's degree in Sociology at the University of Central Florida. Her research and teaching interests include food access, food justice, and the environment.

Adam Driscoll is Assistant Professor of Sociology at the University of Wisconsin-La Crosse. He received his Master's degree in Sociology at East Carolina University and his Ph.D. in Sociology at North Carolina State University. His research and teaching focus upon the environmental impacts of industrial agriculture and effective online pedagogy.

Andrea N. Hunt has a Ph.D. in Sociology from North Carolina State University and is currently Assistant Professor in Sociology and Family Studies at the University of North Alabama. Her research interests include gender, race and ethnicity, mentoring in undergraduate research, engaging teaching practices, and the role of academic advising in student retention.

L. MacNell (✉)
Department of Sociology and Anthropology, 334 1911 Building, Campus Box 8107,
Raleigh, North Carolina 27695, USA
e-mail: loconne@ncsu.edu

A. Driscoll
University of Wisconsin-La Crosse, La Crosse, WI, USA
e-mail: adriscoll@uw.lax.edu

A. N. Hunt
University of North Alabama, Florence, AL, USA
e-mail: ahunt3@una.edu

Student ratings of teaching are often used as an indicator of the quality of an instructor's teaching and play an important role in tenure and promotion decisions (Abrami, d'Apollonia, & Rosenfield, 2007; Benton & Cashin, 2014). Gender bias in these ratings constitutes an important form of inequality facing women in academia that is often unaccounted for in such decisions. Students perceive, evaluate, and treat female instructors quite differently than they do male instructors (Basow, 1995; Centra & Gaubatz, 2000; Feldman, 1992; Young, Rush, & Shaw, 2009). While a general consensus exists that gender plays a vital role in how students perceive and interact with their instructors, there is conflicting evidence as to whether or not this translates into a bias in student ratings due to variations in several mediating factors such as teaching styles and subject material.

Prior studies of student ratings of instruction have been limited in their ability to test for the existence of gender bias because it is difficult to separate the gender of an instructor from their teaching practices in a face-to-face classroom. In online courses, however, students usually base the categorization of their instructor's gender on the instructor's name and, if provided, photograph. It is possible for students to believe that their instructor is actually a man, based solely on a name or photograph, when in reality she is a woman, or vice versa. Therefore, the online environment affords researchers a unique opportunity to assign one instructor two different gender identities in order to understand whether or not differences in student ratings are a result of differences in teaching or simply based on unequal student expectations for male and female instructors. Such experimentation allows researchers to control for potentially confounding factors and therefore attribute observed differences solely to the variable of interest—in this case, the perceived gender of the instructor (Morgan & Winship, 2007).

This study analyzed differences in student ratings of their instructors[1] from an online course, independent of actual gender. The course professor randomly assigned students to one of six discussion groups, two of which the professor taught directly. The other four were taught by one of two assistant instructors—one male and one female. Each instructor was responsible for grading the work of students in their group and interacting with those students on course discussion boards. Each assistant instructor taught one of their groups under their own identity and the second group under the other assistant instructor's identity. Thus, of the two groups who believed they had the female assistant instructor, one actually had the male. Similarly, of the two groups who believed they had the male assistant instructor, one actually had the female (see Table 1). At the end of the course, the professor asked students to rate their instructor through the use of an online survey. This design created a controlled experiment that allowed us to isolate the effects of the gender identity of the assistant instructors, independent of their actual gender. If gender bias was present, than the students from the two groups who believed they had a female assistant instructor should have given their instructor significantly lower evaluations than the two groups who believed they had a male assistant instructor.

Student Ratings of Teaching

Though far from perfect, student ratings of teaching provide valuable feedback about an instructor's teaching effectiveness (Svinicki & McKeachie, 2010). They may be reliably interpreted as both a direct measure of student satisfaction with instruction and as an indirect

[1] To clarify the language we use throughout the paper, we refer to all three persons responsible for grading and directly interacting with students as "instructors." The course "professor" was the person responsible for course design and content preparation, while the two "assistant instructors" worked under the professor's direction to manage and teach their respective discussion groups.

Table 1 Experimental Design.

Discussion Group	Instructor's Perceived Gender	Instructor's Actual Gender
Group A (n =8)	Female	Female
Group B (n =12)	Female	Male
Group C (n =12)	Male	Female
Group D (n =11)	Male	Male

measure of student learning (Marsh, 2007; Murray, 2007). They also play an important role in the selection of teaching award winners, institutional reviews of programs, and student course selection (Benton & Cashin, 2014). More importantly to the careers of educators, these ratings are "used by faculty committees and administrators to make decisions about merit increases, promotion, and tenure" (Davis, 2009, p. 534). In particular, quantitative evaluations of instructors' overall teaching effectiveness are frequently emphasized in personnel decisions (Centra & Gaubatz, 2000). Given the widespread reliance on student ratings of teaching and their effect on career advancement, any potential bias in those ratings is a matter of great consequence.

Gender Bias in Academia

Sociological studies of gender and gender inequality are careful to distinguish between sex (a biological identity) and gender (a socially constructed category built around cultural expectations of male- and female-appropriate behavior). Gender is part of an ongoing performance based on producing a configuration of behaviors that are seen by others as normative. West and Zimmerman (1987) suggested that people engage in gendered behaviors not only to live up to normative standards, but also to minimize the risk of accountability or gender assessment from others. Thus, gender is a process that is accomplished at the interactional level and reinforced through the organization of social institutions such as academia (Lorber, 1994). Gender then contributes to a hierarchal system of power relations that is embedded within the interactional and institutional levels of society and shapes gendered expectations and experiences in the workplace (Risman, 2004).

An examination of gender bias in student ratings of teaching must be framed within the broader context of the pervasive devaluation of women, relative to men, that occurs in professional settings in the United States (Monroe, Ozyurt, Wrigley, & Alexander, 2008). In general, Western culture accords men an automatic credibility or competence that it does not extend to women (Johnson, 2006). Stereotypes that women are less logical, less confident, and occupy lower positions still pervade our organizational structures (Acker, 1990). Conversely, men are automatically assumed to have legitimate authority, while women must prove their expertise to earn the same level of respect. This disparity has been well documented in the field of academia, where men tend to be regarded as "professors" and women as "teachers" (Miller & Chamberlin, 2000) and women face a disparate amount of gender-based obstacles, relative to men (Morris, 2011).

In experiments where researchers gave students identical articles to evaluate—half of which bore a man's name and half of which bore a woman's—the students rated the research they thought had been done by men more highly (Goldberg, 1968; Paludi & Strayer, 1985). In a similar study, college students evaluated two hypothetical applicants for a faculty position and tended to judge the male candidate as more qualified despite the fact that both applicants had identical credentials (Burns-Glover & Veith, 1995). Additionally, a study of student

evaluations of instructors' educational attainment revealed that students misattribute male instructors' education upward and female instructors' education downward (Miller & Chamberlin, 2000). Overall, women in academia tend to be regarded as less capable and less accomplished than men, regardless of their actual achievements and abilities.

Gender Role Expectations

Students often expect their male and female professors to behave in different ways or to respectively exhibit certain "masculine" and "feminine" traits. Commonly held masculine, or "effectiveness," traits include professionalism and objectivity; feminine, or "interpersonal," traits include warmth and accessibility. Students hold their instructors accountable to these gendered behaviors and are critical of instructors who violate these expectations (Bachen, McLoughlin, & Garcia, 1999; Chamberlin & Hickey, 2001; Dalmia, Giedeman, Klein, & Levenburg, 2005; Sprague & Massoni, 2005). Consequently, instructors who adhere to gendered expectations are viewed more favorably by their students (Andersen & Miller, 1997; Bennet, 1982). When female instructors exhibit strong interpersonal traits, they are viewed comparably to their male counterparts. When female instructors fail to meet these gendered expectations, however, they are sanctioned, while male instructors who do not exhibit strong interpersonal traits are not (Basow & Montgomery, 2005; Basow, Phelan, & Capotosto, 2006). At the same time, students are less tolerant of female instructors whom they perceive as lacking professionalism and objectivity than they are of male instructors who lack the same qualities (Bennet, 1982). In general, "students' perceptions and evaluations of female faculty are tied more closely to their gender expectations than for male faculty" (Bachen et al., 1999, p. 196).

These different standards can place female instructors in a difficult "double-bind," where gendered expectations (that women be nurturing and supportive) conflict with the professional expectations of a higher-education instructor (that they be authoritative and knowledgeable) (Sandler, 1991; Statham, Richardson, & Cook, 1991). On the one hand, students expect female instructors to embody gendered interpersonal traits by being more accessible and personable. However, these same traits can cause students to view female instructors as less competent or effective. On the other hand, female instructors who are authoritative and knowledgeable are violating students' gendered expectations, which can also result in student disapproval. Therefore, female instructors are expected to be more open and accessible to students *as well as* to maintain a high degree of professionalism and objectivity. Female instructors who fail to meet these higher expectations are viewed as less effective teachers than men (Basow, 1995).

Male instructors, however, are rated more highly when they exhibit interpersonal characteristics in addition to the expected effectiveness characteristics (Andersen & Miller, 1997). In other words, female instructors who fail to exhibit an ideal mix of traits are rated lower for not meeting expectations, while male instructors are not held to such a standard. Consequently, gendered expectations represent a greater burden for female than male instructors (Sandler, 1991; Sprague & Massoni, 2005). An important manifestation of that disparity is bias in student ratings of instructors, where female instructors may receive lower ratings than males, not because of differences in teaching but for failing to meet gendered expectations.

Methodological Concerns with Previous Studies of Gender Bias

Studies of gender bias in student ratings of instruction have presented complicated and sometimes contradictory results. Sometimes men received significantly higher ratings (Basow & Silberg, 1987; Sidanius & Crane, 1989), sometimes women (Bachen et al., 1999; Rowden & Carlson, 1996), and sometimes neither (Centra & Gaubatz, 2000; Feldman, 1993). The

variety of results in these studies suggests that gender does play a role in students' ratings of their instructors, but that it is a complex and multifaceted one (Basow et al., 2006).

One reason why prior research on gender bias in student ratings of teaching has provided such inconclusive results may lie in the research design of these previous studies. A large portion of research on student ratings of teaching directly utilized those ratings for their data (e.g. Basow, 1995; Bennett, 1982; Centra, 2007; Centra & Gaubatz, 2000; Marsh, 2001). This strategy allows for the analysis of a large amount of data, but it does not control for differences in actual teaching and therefore may fail to capture gender bias in student ratings. Studies that compare student ratings of instructors explore whether or not there are differences—not whether or not those differences are the result of gender bias (Feldman, 1993). For example, a study of ratings may find that a female instructor received significantly lower scores than a male peer, but it could not assess whether that indicates a true difference in teaching quality. Perhaps she was not perceived as warm and engaging; failing to meet the gendered expectations of the students, she may have been rated more poorly than her male peer despite being an equally effective instructor. Similarly, the lack of a gender disparity in student ratings of instruction could actually obscure a gender bias if at a particular institution the female faculty members were, on average, stronger instructors than the males, yet were being penalized by the students due to bias (Feldman, 1993).

Additionally, a number of situational elements may serve to sway student ratings of male verses female instructors as male and female professors tend to occupy somewhat different teaching situations. Men are overrepresented in the higher ranks of academic positions as well as in STEM fields. They are also more likely to teach upper-level courses whereas women are more likely to teach introductory courses (Simeone, 1987; Statham et al., 1991). Women are also more likely than men to be employed in full-time non-tenure track positions as well as in part-time positions (Curtis, 2011). These factors are highly relevant because instructor rank, academic area, and class level of the course have all been found to directly impact student ratings of instruction (Feldman, 1993; Liu, 2012). All of these factors serve to complicate the relationship between instructor gender and student ratings of instruction and obfuscate the conclusions that can be drawn from direct studies of such ratings. Studies of actual student ratings of instruction may tell us more about women's position in academia than about actual gender bias in student ratings. In contrast, experimental studies allow the researcher to control for both the quality and character of the teaching as well as the academic position of the instructor; ensuring that any differences registered in student ratings indicate, as much as possible, a bias rather than an actual difference in teaching (Feldman, 1993).

Research Question and Related Hypotheses

The fundamental question examined in this study is whether or not students rate their instructors differently on the basis of what they perceive those instructors' gender to be. We expected that there would be no difference between the ratings for the actual male and female instructors in the course as every attempt was made to minimize any differences in interaction and teaching. However, we expected that student ratings of instructors would reflect the different expectations for male and female instructors discussed above. Instructors whom students perceived to be male would be afforded an automatic credibility on their competence and professionalism. Furthermore, they would not be penalized for any perceived deficiency in their interpersonal skills. Therefore, we expected that students would rate the instructors they *believed* to be male more highly than ones they believed to be female, regardless of the instructors' actual gender.

The Study and Methodology

This study examined gender bias in student ratings of teaching by falsifying the gender of assistant instructors in an online course and asking students to evaluate them along a number of instructional criteria. By using a 2-by-2 experimental design (see Table 1), we were able to compare student evaluations of a perceived gender while holding the instructor's actual gender (and any associated differences in teaching style) constant. Any observed differences in how students rated one perceived gender versus the other must have therefore derived from bias on the students' part, given that the exact same two instructors (one of each gender) were being evaluated in both cases.

Subjects

Data were collected from an online introductory-level anthropology/sociology course offered during a five-week summer session at a large (20,000+), public, 4-year university in North Carolina. The University's institutional review board had approved this study (IRB# 2640). The course fulfilled one of the university's general education requirements, and the students represented a range of majors and grade levels. The majority of the participants were traditional college-aged students with a median age of 21 years. The instructors taught the course entirely through a learning management system and students' only contact with their instructors was either through e-mail or comments posted on the learning management system. The professor delivered course content through assigned readings and written PowerPoint slideshow lectures. The course was broken up into nine different content sections. For each section, students were required to read the assigned material and make a series of posts on a structured discussion board. The course had 72 students who were randomly divided into six discussion groups for the entirety of the course. All discussion board activity took place within the assigned discussion group. Each discussion group had one instructor responsible for moderating the discussion boards and grading all assignments for that group. The course professor took two groups and divided the remaining four between the two assistant instructors, each taking one group under their own identity and a second under their fellow assistant instructor's identity (see Table 1). All instructors were aware of the study being conducted and cooperated fully.

The section discussion boards were the primary source of interaction between students and the course instructors and, as such, represented 30% of the students' final grades. The discussion boards were also an important part of student learning because they were the main arena in which students could analyze and voice questions about course concepts and material. The instructor assigned to each discussion group maintained an active presence on each discussion board, offering comments and posing questions. The instructor also graded students' posts and provided detailed feedback on where students had lost points. The two assistant instructors for the four discussion groups employed a wide range of strategies so as to maintain consistency in teaching style and grading. The two assistant instructors composed personal introduction posts that indicated similar biographical information and background credentials. They posted on the discussion boards and graded assignments at the same time of day three days each week to ensure that no group received significantly faster or slower feedback than others. The professor provided detailed grading rubrics for the discussion boards, and the instructors coordinated their grading to ensure that these rubrics were applied to students' work equitably.[2]

[2] A one-way ANOVA test confirmed that there was no significant variation among all six groups' discussion board grades and overall grades for the course.

Toward the end of the course the professor sent students reminder e-mails requesting that they complete an online evaluation of their instructor. These evaluations were explained as serving the purpose of providing the professor with feedback about the instructors' performance. The survey asked students to rate their instructor on various factors such as accessibility, effectiveness, and overall quality. Over 90% of the class completed the evaluation. For the purpose of this study, we only analyzed data from the discussion groups assigned to the assistant instructors, leaving us with 43 subjects.

Instrument

The instructor evaluation consisted of 15 closed-ended questions that ask students to rate their instructors on a variety of measures using a five-point Likert scale (1 = Strongly disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly agree). The survey had six questions designed to measure effectiveness traits (e.g. professionalism, knowledge, and objectivity) and six questions designed to measure interpersonal traits (e.g. respect, enthusiasm, and warmth). In addition, there were two questions designed to measure communication skills and one question that asked students to evaluate the instructor's overall quality as a teacher. We also asked students to indicate which discussion group they were in and to provide basic demographic and academic background information including gender, age, year in school, and number of credit hours currently being taken. All students fully completed the evaluation, leaving us with no missing data.

We performed all analyses with the 13th version of the Stata statistical analysis program. We used exploratory factor analysis to test how well the separate questions reflected a common underlying dimension. Principal component factor analysis revealed that 12 of our items characterized a single factor for which the individual factor loadings ranged from .7370 to .9489; sufficiently high to justify merging them into a single index (Hair, Anderson, Tatham, & Black, 1998). This indicates that those 12 questions on our survey were all measuring the same latent variable, which we interpret to be a general evaluation of the instructor's teaching. A reliability test yielded a Cronbach's alpha above .950 for the 12 questions. In order to confirm the factor structure, we used structural equation modeling to test a single latent variable indicated by our 12 separate questions. Our model was a strong fit to the data (N =43, $\chi^2(47)$ =59.18 (not significant), RMSEA =0.078, CFI =0.980, SRMR =0.043) with all loadings significant at the p <0.001 level. Therefore, we extracted a factor score, *student ratings index*, which weighed each question by how strongly it loaded onto the single factor, providing us with a single representation of how well each student evaluated their instructor's teaching.

Analysis

To test for the existence of gender bias in student ratings of teaching, we made two types of comparisons. First we compared across the *actual* gender of the assistant instructor, combining the two groups that had the female assistant instructor (one of which thought they had a male) into one category and doing the same with the two groups that had the male assistant instructor. Second, we compared across the *perceived* gender of the assistant instructor, combining the two groups that thought they had a female assistant instructor (one of which was actually a male) into one category and doing the same with the two groups that thought they had a male assistant instructor. We made both comparisons for the 12 individual questions, as well as the *student ratings index*. We used Welch's *t*-tests (an adaptation of the Student's *t*-test that does not assume equal variance) to establish the statistical significance of each difference. We also ran two general linear multivariate analyses of variance (MANOVAs) on the set of 12 variables

to test the effects of instructor gender (perceived and actual) on all of the questions considered as a group. A MANOVA allows a researcher to test a set of correlated dependent variables and conduct a single, overall comparison between the groups formed by categorical independent variables (Garson, 2012). This *F*-test of all means addresses the potential for false positive findings as the result of multiple comparisons.[3]

Results

Student Ratings of Perceived and Actual Gender

By comparing differences across the *actual* gender of the assistant instructor with those observed across the *perceived* gender of the instructor it is possible to observe whether or not students rated their instructors differently depending on the gender of the instructor. The results of this comparison are found in Table 2.

Our MANOVAs indicate that there is a significant difference in how students rated the perceived male and female instructors ($p < 0.05$), but not the actual male and female instructors. When looking at the individual questions as well as the *student ratings index*, there are no significant differences between the ratings of the actual male and female instructor (the first and second columns in Table 2). Students in the two groups that had the female assistant instructor (one of which thought they had a male) did not rate their instructor any differently than did the students in the two groups that had the male assistant instructor. The left two columns of Fig. 1 provide a graphic representation of this comparison for the *student ratings index*. The overlapping error bars (+/- one standard error) indicate the lack of a significant difference between how students rated the actual male and female assistant instructors.

When comparing between the perceived gender identities of the instructors (the fourth and fifth columns in Table 2), we found that the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the *student ratings index*.[4] Looking at the *R*-squares, all seven of these comparisons yielded a medium sized effect. It is worth noting, particularly given the small sample size, that the male instructor identity also received higher scores on the other six questions, though not to a statistically significant degree. Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female, regardless of the actual gender of the assistant instructor. This comparison is represented graphically by the right two columns of Fig. 1, where a clear difference can be observed.

[3] We acknowledge that the application of parametric analytical techniques (ANOVA, MANOVA, and *t*-tests) to ordinal data (the Likert scale responses) remains controversial among social scientists and statisticians. (See Knapp (1990) for a relatively balanced review of the debate.) We side with the arguments of Gaito (1980) and Armstrong (1981) and argue that it is appropriate to do so in our case as the concept being measured is interval, even if the data labels are not. This practice is common within higher education research. (e.g. Centra & Gaubatz [2000] Young, Rush, & Shaw [2009]; Basow [1995]; and Knol et al. [2013])

[4] While we acknowledge that a significance level of .05 is conventional in social science and higher education research, we side with Skipper, Guenther, and Nass (1967), Labovitz (1968), and Lai (1973) in pointing out the arbitrary nature of conventional significance levels. Considering our study design, we have used a significance level of .10 for some tests where: 1) the results support the hypothesis and we are consequently more willing to reject the null hypothesis of no difference; 2) our hypothesis is strongly supported theoretically and by empirical results in other studies that use lower significance levels; 3) our small *n* may be obscuring large differences; and 4) the gravity of an increased risk of Type I error is diminished in light of the benefit of decreasing the risk of a Type II error (Labovitz, 1968; Lai, 1973).

Table 2 Comparison of means of student ratings of teaching across the actual gender of the assistant instructor and the perceived gender of the assistant instructor

Question	Actual Female	Actual Male	Difference	Perceived Female	Perceived Male	Difference
Caring	4.00 (1.257)	3.87 (0.868)	0.13 (*0.004*)	3.65 (1.226)	4.17 (0.834)	-0.52 (*0.071*)
Consistent	3.80 (1.322)	3.70 (1.020)	0.10 (*0.002*)	3.50 (1.357)	3.96 (0.928)	-0.47 (*0.045*)
Enthusiastic	4.05 (1.191)	3.78 (0.850)	0.27 (*0.019*)	3.60 (1.314)	4.17 (0.576)	-0.57† (*0.112*)
Fair	4.05 (1.050)	3.78 (0.951)	0.27 (*0.018*)	3.50 (1.192)	4.26 (0.619)	-0.76* (*0.188*)
Feedback	4.10 (1.252)	3.83 (1.029)	0.27 (*0.015*)	3.70 (1.380)	4.17 (0.834)	-0.47 (*0.054*)
Helpful	3.65 (1.309)	3.83 (0.834)	-0.18 (*0.008*)	3.50 (1.192)	3.96 (0.928)	-0.46 (*0.049*)
Knowledgeable	4.20 (1.056)	4.09 (0.949)	0.11 (*0.003*)	3.95 (1.191)	4.30 (0.765)	-0.35 (*0.038*)
Praise	4.35 (0.988)	4.09 (0.900)	0.26 (*0.020*)	3.85 (1.089)	4.52 (0.665)	-0.67* (*0.153*)
Professional	4.30 (1.218)	4.35 (0.935)	-0.05 (*0.000*)	4.00 (1.414)	4.61 (0.499)	-0.61† (*0.124*)
Prompt	4.10 (1.252)	3.87 (0.919)	0.23 (*0.013*)	3.55 (1.356)	4.35 (0.573)	-0.80* (*0.191*)
Respectful	4.30 (1.218)	4.35 (0.935)	-0.05 (*0.001*)	4.00 (1.414)	4.61 (0.499)	-0.61† (*0.124*)
Responsive	4.00 (1.124)	3.57 (0.843)	0.43 (*0.052*)	3.65 (1.137)	3.87 (0.869)	-0.22 (*0.013*)
Student Rating Index	0.09 (1.165)	-0.08 (0.850)	0.17 (*0.008*)	-0.33 (1.267)	0.284 (0.584)	-0.61† (*0.128*)
N	20	23		20	23	

Note: Each cell contains the mean student response for the question with the standard deviations in parentheses. The cells in the Difference columns contain the difference between the means with the *r*-squared in italics and parentheses. Welch's *t*-tests were used to establish the significance of the observed differences.

† p < =0.10.
* p < =0.05.

Discussion

With the design of this experiment, we are able to attribute any differences between how students rated the two perceived genders to gender bias as the students actually evaluated the same two instructors in each case. Our findings support the existence of gender bias in that

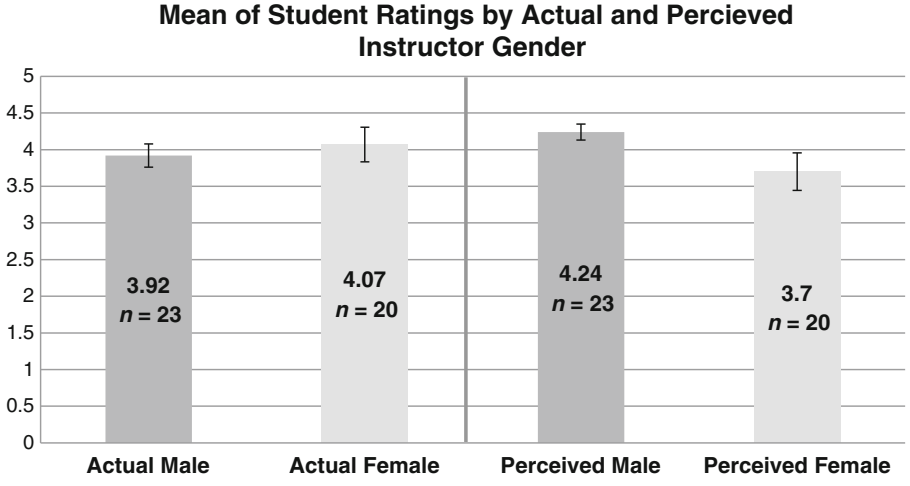

Figure 1 Comparison of the mean of student ratings across actual instructor gender (left two columns) and perceived instructor gender (right two columns). The difference between the right two columns is significant to the p<=0.10 level.

students rated the instructors they perceived to be female lower than those they perceived to be male, regardless of teaching quality or actual gender of the instructor. The perceived female instructor received significantly lower ratings on six of the 12 metrics on the survey, as well as on the *student ratings index*.

The difference between how students rated the two perceived genders stands in stark contrast to the fact that neither the actual male nor actual female instructor received significantly higher ratings than the other. Both instructors performed equally well from the students' perspective. However, in both cases the *same* instructor received different ratings depending solely on their perceived gender. In other words, when the actual male instructor was perceived to be female, he received significantly lower ratings than when he was perceived to be a male. For example, when the actual male and female instructors posted grades after two days *as a male*, this was considered by students to be a 4.35 out of 5 level of promptness, but when the same two instructors posted grades at the same time *as a female*, it was considered to be a 3.55 out of 5 level of promptness. In each case, the same instructor, grading under two different identities, was given lower ratings half the time with the only difference being the perceived gender of the instructor. Similarly, students rated the perceived female instructors an average of 0.75 points lower on the question regarding fairness, despite both instructors utilizing the same grading rubrics and there being no significant differences in the average grades of any of the groups. These findings support the argument that male instructors are often afforded an automatic credibility in terms of their professionalism, expertise, and effectiveness as instructors. Despite the fact that the students were equally satisfied with the promptness and fairness of the *actual* instructors, the instructor that students perceived to be male was considered to be more effective.

Similarly, both actual instructors demonstrated the same level of interpersonal interaction in their attempts to create a sense of immediacy in the online classroom. Yet the perceived male instructor received higher ratings on all six interpersonal measures, three of them significantly. We contend that female instructors are *expected* to exhibit such traits and therefore are not rewarded when they do so, while male instructors are perceived as going above and beyond expectations when they exhibit these traits. In other words, students have higher interpersonal standards for their female instructors (Sandler, 1991). Our findings support the existence of this bias. In the online environment, it is more difficult to create immediacy through verbal communication, and nonverbal communication and body language are eliminated entirely (O'Sullivan, Hunt, & Lippert, 2004). Students sanctioned the perceived female instructor for failing to demonstrate strong interpersonal traits, yet did not do the same for the perceived male instructor. Both instructors were working within the same confines of online, text-based communications, but students only penalized the instructor they perceived to be female for this shortcoming.

Although this experiment was conducted in the online environment, we believe that the findings apply more broadly to all student ratings of teaching. Rather than testing for gender bias in the online environment, we used this environment as a natural laboratory to test for the existence of gender bias in student ratings as a whole. We argue that the demonstrated bias exists in the general student population and will manifest itself in both online and face-to-face classrooms. The combination of higher expectations and lower automatic credibility translates into very real differences in student ratings of female versus male instructors. Though it is easier to affect interpersonal characteristics in a face-to-face environment, the fact remains that some

professors are *expected* to do so while others are given a ratings boost for those same behaviors.

Because student ratings of teaching are considered an important measure of teaching proficiency, the existence of gender bias in those scores needs to be better understood and acknowledged within the institutional framework of our higher-education system. These results provide strong evidence that gender bias exists in student ratings of their instructors, but more work is needed. First and foremost, these results need to be replicated in other similar online classes. A single case study cannot establish a broad pattern. However, it does suggest the existence of one and provides incentive for further exploration. Additional studies of this type could lend weight to these findings and better establish the existence of this bias throughout academia. Additionally, courses in other subject areas with a variety of both male and female instructors should follow a similar model to corroborate these findings.

Conclusions

Our findings show that the bias we saw here is *not* a result of gendered behavior on the part of the instructors, but of actual bias on the part of the students. Regardless of actual gender or performance, students rated the perceived female instructor significantly more harshly than the perceived male instructor, which suggests that a female instructor would have to work harder than a male to receive comparable ratings. If female professors and instructors are continually receiving lower evaluations from their students for no other reason than that they are women, then this particular form of inequality needs to be taken into consideration as women apply for academic jobs and come up for promotion and review.

These findings represent an important contribution to existing debates over the validity of student ratings of teaching. (See Benton & Cashin, 2014; Perry & Smart, 2007; and Theall, Abrami, & Mets, 2001 for reviews.) These debates have highlighted a number of weaknesses and shortcomings of student ratings of teaching as a reflection of the quality of instruction being rated (Greenwald, 1997; Johnson, 2003; Svanum & Aigner, 2011). They have also shown that there is substantial room for updating and improving how student ratings of teaching are collected, interpreted, and utilized (Hampton & Reiser, 2004; Subramanya, 2014). However, for better or worse, they remain one of the primary tools used to evaluate educators' teaching for the purposes of promotion and tenure decisions (Davis, 2009; Svinicki & McKeachie, 2010). This study demonstrates that gender bias is an important deficiency of student ratings of teaching. Therefore, the continued use of student ratings of teaching as a primary means of assessing the quality of an instructor's teaching systematically disadvantages women in academia. As this limitation is one of numerous problems associated with the emphasis on quantitative student ratings of teaching, this work adds to the growing call for re-evaluation and modification of the current system of evaluating the quality of instruction in higher education (Hampton & Reiser, 2004; Morrison & Johnson, 2013).

It is also worth noting that this experiment is only scratching the surface of what is possible with gender studies in the online environment. The online environment presents a unique opportunity to experiment directly with gender identity. Analyzing the difference in online behavior of individuals when they perceive that they are interacting with a male or female could provide a wealth of data on how gender is constructed and treated. We hope that this experiment serves as a model for future work that will enhance our ability to test for gender bias in order to further our understanding of its basis, means of perpetuation, and potential avenues of amelioration.

References

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–445). Dordrecht, The Netherlands: Springer.

Acker, J. (1990). Hierarchies, job, and bodies: A theory of gendered organizations. *Gender and Society, 4*, 81–95.

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *Ps-Political Science and Politics, 30*, 216–219.

Armstrong, G. D. (1981). Parametric statistics and ordinal data: A pervasive misconception. *Nursing Research, 30*, 60–62.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*, 193–210.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656–665.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-rating of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91–106.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*, 25–35.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308–314.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology, 74*, 170–179.

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 279–326). Dordrecht, The Netherlands: Springer.

Burns-Glover, A. L., & Veith, D. J. (1995). Revisiting gender and teaching evaluations: Sex still makes a difference. *Journal of Social Behavior and Personality, 10*, 69–80.

Centra, J. A. (2007). *Differences in responses to the student instructional report: Is it bias?* Princeton, NJ: Educational Testing Service.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*, 17–33.

Chamberlin, M. S., & Hickey, J. S. (2001). Student evaluations of faculty performance: The role of gender expectations in differential evaluations. *Educational Research Quarterly, 25*, 3–14.

Curtis, J. W. (2011). *Persistent inequity: Gender and academic employment.* Report from the American Association of University Professors. Retrieved from http://www.aaup.org/NR/rdonlyres/08E023AB-E6D8-4DBD-99A0-24E5EB73A760/0/persistent_inequity.pdf

Dalmia, S., Giedeman, D. C., Klein, H. A., & Levenburg, N. M. (2005). Women in academia: An analysis of their expectations, performance and pay. *Forum on Public Policy, 1*, 160–177.

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.

Feldman, K. A. (1992). College students' views of male and female college teachers: Evidence from the social laboratory and experiments – Part 1. *Research in Higher Education, 33*, 317–375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Evidence from the social laboratory and experiments – Part 2. *Research in Higher Education, 34*, 151–211.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564–567.

Garson, G. D. (2012). *General linear models: Multivariate GLM & MANOVA/MANCOVA*. Asheboro, NC: Statistical Associates.

Goldberg, P. (1968). Are women prejudiced against women? *Trans-action, 5*, 28–30.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182–1186.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis with readings* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*, 497–527.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer.

Johnson, A. (2006). *Power, privilege, and difference*. Boston, MA: McGraw-Hill.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*, 121–123.

Knol, M. H., Veld, R., Vorst, H. C. M., van Driel, J. H., & Mellenbergh, G. J. (2013). Experimental effects of student evaluations coupled with collaborative consultation on college professors' instructional skills. *Research in Higher Education, 54*, 825–850.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of.05. *The American Sociologist, 3*, 220–222.

Lai, M.K. (1973). *The case against tests of statistical significance*. Report from the Teacher Education Division Publication Series. Retrieved from http://files.eric.ed.gov/fulltext/ED093926.pdf

Liu, O. L. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education, 53*, 471–486.

Lorber, J. (1994). *Paradoxes of gender*. New Haven, CT: Yale University Press.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal, 38*, 183–212.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology, 28*, 283–298.

Monroe, K., Ozyurt, S., Wrigley, T., & Alexander, A. (2008). Gender equality in academia: Bad news from the trenches, and some possible solutions. *Perspectives on Politics, 6*, 215–233.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, MA: Cambridge University Press.

Morris, L. V. (2011). Women in higher education: Access, success, and the future. *Innovative Higher Education, 36*, 145–147.

Morrison, K., & Johnson, T. (2013). Editorial. *Educational Research and Evaluation, 19*, 579–584.

Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 145–183). Dordrecht, The Netherlands: Springer.

O'Sullivan, P. D., Hunt, S. K., & Lippert, L. R. (2004). Mediated immediacy: A language of affiliation in a technological age. *Journal of Language and Social Psychology, 23*, 464–490.

Paludi, M. A., & Strayer, L. A. (1985). What's in an author's name? Differential evaluations of performance as a function of author's name. *Sex Roles, 12*, 353–361.

Perry, R. P., & Smart, J. C. (Eds.). (2007). *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht, The Netherlands: Springer.

Risman, B. J. (2004). Gender as a social structure: Theory wrestling with activism. *Gender & Society, 18*, 429–450.

Rowden, G. V., & Carlson, R. E. (1996). Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychological Reports, 78*, 835–839.

Sandler, B. R. (1991). Women faculty at work in the classroom, or, why it still hurts to be a woman in labor. *Communication Education, 40*, 6–15.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology, 19*, 174–197.

Simeone, A. (1987). *Academic women: Working toward equality.* South Hadley, MA: Bergin and Garvey.

Skipper, J. K., Guenther, A. C., & Nass, G. (1967). The sacredness of.05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 1*, 16–18.

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles, 53*, 779–793.

Statham, A., Richardson, L., & Cook, J. A. (1991). *Gender and university teaching: A negotiated difference*. Albany, NY: State University of New York Press.

Subramanya, S. R. (2014). Toward a more effective and useful end-of-course evaluation scheme. *Journal of Research in Innovative Teaching, 7*, 143–157.

Svanum, S., & Aigner, C. (2011). The influences of course effort, mastery and performance goals, grade expectancies, and earned course grades on student ratings of course satisfaction. *British Journal of Educational Psychology, 81*, 667–679.

Svinicki, M., & McKeachie, W. J. (2010). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). *The student ratings debate: Are they valid? How can we best use them?* San Francisco, CA: Jossey-Bass.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1*, 125–151.

Young, S., Rush, L., & Shaw, D. (2009). Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal of Scholarship of Teaching and Learning, 3*, 1–14.