Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career

Bob Uttl [Corresp., 1] , **Dylan Smibert** [2]

1 Department of Psychology, Mount Royal University, Calgary, Alberta, Canada

2 Department of Psychology, Saint Mary's University, Halifax, Nova Scotia, Canada

Corresponding Author: Bob Uttl
Email address: uttlbob@gmail.com

Anonymous student evaluations of teaching (SETs) are used by colleges and universities to measure teaching effectiveness and to make decisions about faculty hiring, firing, re-appointment, promotion, tenure, and merit pay. Although numerous studies have found that SETs correlate with various teaching effectiveness irrelevant factors (TEIFs) such as subject, class size, and grading standards, it has been argued that such correlations are small and do not undermine the validity of SETs as measures of professors' teaching effectiveness. However, previous research has generally used inappropriate parametric statistics and effect sizes to examine and to evaluate the significance of TEIFs on personnel decisions. Accordingly, we examined the influence of quantitative vs. non quantitative courses on SET ratings and SET based personnel decisions using 14, 872 publicly posted class evaluations where each evaluation represents a summary of SET ratings provided by individual students responding in each class. In total, 325,538 individual student evaluations from a US mid-size university contributed to theses class evaluations. The results demonstrate that class subject (math vs. English) is strongly associated with SET ratings, has a substantial impact on professors being labeled satisfactory vs. unsatisfactory and excellent vs. non-excellent, and the impact varies substantially depending on the criteria used to classify professors as satisfactory vs. unsatisfactory. Professors teaching quantitative courses are far more likely not to receive tenure, promotion, and/or merit pay when their performance is evaluated against common standards.

2
3
4
5
6 **Title:**
7 Student Evaluations of Teaching: Teaching Quantitative Courses Can Be Hazardous to One's Career
8
9 **Abstract:**

10 Anonymous student evaluations of teaching (SETs) are used by colleges and universities to measure
11 teaching effectiveness and to make decisions about faculty hiring, firing, re-appointment, promotion,
12 tenure, and merit pay. Although numerous studies have found that SETs correlate with various teaching
13 effectiveness irrelevant factors (TEIFs) such as subject, class size, and grading standards, it has been
14 argued that such correlations are small and do not undermine the validity of SETs as measures of
15 professors' teaching effectiveness. However, previous research has generally used inappropriate
16 parametric statistics and effect sizes to examine and to evaluate the significance of TEIFs on personnel
17 decisions. Accordingly, we examined the influence of quantitative vs. non quantitative courses on SET
18 ratings and SET based personnel decisions using 14, 872 publicly posted class evaluations where each
19 evaluation represents a summary of SET ratings provided by individual students responding in each
20 class. In total, 325,538 individual student evaluations from a US mid-size university contributed to
21 theses class evaluations. The results demonstrate that class subject (math vs. English) is strongly
22 associated with SET ratings, has a substantial impact on professors being labeled satisfactory vs.
23 unsatisfactory and excellent vs. non-excellent, and the impact varies substantially depending on the
24 criteria used to classify professors as satisfactory vs. unsatisfactory. Professors teaching quantitative
25 courses are far more likely not to receive tenure, promotion, and/or merit pay when their performance
26 is evaluated against common standards.

27
28
29 **Authors:**

30 Bob Uttl, Department of Psychology, Mount Royal University, Calgary, Alberta, Canada
31 Dylan Smibert, Department of Psychology, Saint Mary's University, Halifax, Nova Scotia, Canada
32
33 **Corresponding author:**
34 Bob Uttl, Address: Department of Psychology, Mount Royal University, 4825 Mount Royal Gate SW,
35 Calgary, AB, Canada; Phone: +1-403-440-8539; Email: uttlbob@gmail.com

36

Introduction

37

38   Anonymous student evaluations of teaching (SETs) are used by colleges and universities to
39   measure teaching effectiveness and to make decisions about faculty hiring, firing, re-appointment,
40   promotion, tenure, and merit pay. Although SETs are relatively reliable when average ratings across
41   five or more courses (depending on class size) are used, their validity has been questioned.
42   Specifically, numerous studies have found that SETs correlate with various teaching effectiveness
43   irrelevant factors (TEIFs) such as class size (Benton & Cashin, 2012), subject (Benton & Cashin,
44   2012), and professor hotness/sexiness (Felton, Koper, Mitchell, & Stinson, 2008; Felton, Mitchell, &
45   Stinson, 2004). However, it is often argued that correlations between TEIFs and SETs are small and
46   therefore do not undermine the validity of SETs (Beran & Violato, 2005; Centra, 2009). To illustrate,
47   Beran and Violato (2005) examined correlations between several TEIFs and SETs using over 370,000
48   individual student ratings. Although they reported $d = 0.61$ between ratings of courses in natural vs
49   social science, they further analyzed their data using regression analyses and concluded that course
50   characteristics, including the discipline, were not important. They wrote: "From examining numerous
51   student and course characteristics as possible correlates of student ratings, results from the present
52   study suggest they are not important factors." (p. 599). Similarly, using Educational Testing Service
53   data from 238,471 classes, Centra (2009) found that the natural sciences, mathematics, engineering,
54   and computer science courses were rated about 0.30 standard deviation lower than courses in the
55   humanities (English, history, languages) and concluded that "a third of a standard deviation does not
56   have much practical significance". If so, one may argue, SETs are both reliable and valid and TEIFs
57   can be ignored by administrators when making judgments about faculty's teaching effectiveness for
58   personnel decisions.
59   However, SET research has been plagued by several unrecognized methodological
60   shortcomings that render much of the previous research on reliability, validity and other aspects of SET
61   invalid and uninterpretable. First, SET rating distributions are typically strongly negatively skewed due
62   to severe ceiling effects, that is, due to a large proportion of students giving professors the highest
63   possible ratings. In turn, it is inappropriate and invalid to describe and analyze these ceiling-limited
64   ratings using parametric statistics that assume a normal distribution of data (i.e., means, SDs, ds, rs, r^2;
65   see Uttl (2005), for an extensive discussion of the problems associated with severe ceiling effects,
66   including detection of ceiling effects and consequences of ceiling effects). Yet, all of the studies we
67   have examined to date do precisely that -- use means, SDs, ds, rs, and r^2 to describe SETs; and to
68   investigate associations between SETs and TEIFs.
69   Second, when making judgments about the practical significance of associations between TEIFs
70   and SETs, researchers typically rely on various parametric effect size indexes such as ds, rs, and r^2 or
71   proportion of variance explained and, after finding them to be small, conclude that TEIFs are ignorable
72   and do not undermine the validity of SETs. However, it has been argued elsewhere that effect size
73   indexes should be chosen based not only on the statistical properties of data but also based on their
74   relationship to practical or clinically significant outcomes (Bond, Wiitala, & Richard, 2003; Deeks,
75   2002). Given that SETs are used to make primarily binary decisions about whether a professor's
76   teaching effectiveness is "satisfactory" or "unsatisfactory", the most appropriate effect size indexes may
77   be relative risk ratio (RR) or odds ratios (OR) of professors passing the "satisfactory" cut off as a
78   function of, for example, them teaching quantitative vs. non-quantitative courses rather than ds, rs, and
79   r^2 (Deeks, 2002).
80   Third, researchers sometimes evaluate the importance of various factors based on correlation
81   and regression analyses of SET ratings given by individual students (individual student SET ratings)
82   rather than on the mean SET ratings given by all responding students in each class (class SET summary
83   ratings). However, the proportion of variance explained by some characteristic in individual student

84 SET ratings is not relevant to the effect the characteristic may have on the class SET summary ratings
85 that are used to make personnel decisions about faculty members. For example, Beran and Violato
86 (Beran & Violato, 2005) based their conclusion that various student and course characteristics "are not
87 important factors" based on regression analyses over individual student SET ratings.
88   Accordingly, we re-examined the influence of one TEIF -- teaching quantitative vs. non-
89 quantitative courses -- on SET ratings and SET-based personnel decision in a large sample of class
90 summary evaluations from a midsize US university. We had two primary objectives. First, what is the
91 relationship between course subject and SET ratings? Specifically, what is the distribution of SET
92 ratings obtained by Math (and Stats) professors vs. professors in other fields such as English, History,
93 and Psychology? Second, what are the consequences of course subject on making judgments about
94 professors' teaching effectiveness? Specifically, what percentage of professors teaching Math vs.
95 professors teaching other subjects pass the satisfactory cut-off determined by the mean SET ratings
96 across all courses or other norm referenced cut-offs that ignore course subject?
97   In addition, we also examined how personnel decisions about professors might be affected if
98 criterion referenced, label-based cut-offs were used instead of norm referenced cut offs. In many
99 universities, SET questionnaires use Likert response scales where students indicate their degree of
100 agreement with various statements purportedly measuring teaching effectiveness. Professors' teaching
101 effectiveness is then evaluated against various norm-referenced cut offs such as the departmental mean,
102 mean minus one standard deviation (e.g., 4.0 on 5-point scale), or perhaps a cut off determined by the
103 20th percentile of all ratings such as 3.5 on 5 point scale. In other universities, SETs use label based
104 response scales where students indicate whether a particular aspect of instruction was, for example,
105 "Poor", "Fair", "Good", "Very Good", and "Excellent". Here, if students rate professors as "Poor", then,
106 arguably, to the extent to which SETs measure teaching effectiveness (a contentious issue on its own), a
107 professor's teaching effectiveness is not satisfactory. If students rate a professor as "Fair", the plain
108 meaning of this term is "sufficient but not ample" or "adequate" (Merriam-Webster online dictionary)
109 or "satisfactory". Presumably, if students rate professors as "Good" or higher, professors should be
110 more than "satisfactory" and those rated as "Excellent" are deserving of teaching awards. In contrast to
111 Likert response scales, label-based response scales directly elicit clearly interpretable evaluation
112 judgments from students themselves.

Method

114   We obtained 14,872 class summary evaluations, with each representing a summary of SET
115 ratings provided by individual students responding in each class in a US midsize university (New York
116 University or NYU). In total, 325,538 individual student SET ratings contributed to the 14,872 class
117 summary evaluations. The unit of analysis used in this study are the class summary evaluations. The
118 class summary evaluations were posted on the university's website (www.nyu.edu), available to the
119 general public (rather than to registered students only), and were downloaded in the first quarter of
120 2008. Table 1 shows the individual questions on the NYU SET forms used to evaluate teaching
121 effectiveness on a 5-point scale where 1 = *Poor* and 5 = *Excellent*. The mean ratings across all nine
122 items and course subject (e.g., English, Math, History) were extracted from the evaluations and used in
123 all analyses. The SET evaluations included responses to other questions including questions on
124 workload, labs, and course retake that are not considered in this report. No ethics review was required
125 for this research because all data were available to general public in form of archival records.

Results

127   Table 1 shows the means and standard deviations for individual SET items across all 14,872
128 courses as well as the mean overall average (i.e., average calculated for each course across the 9
129 individual items). Item mean ratings ranged from 3.90 to 4.37 with *SD*s ranging from 0.52 to 0.63.

130   The mean overall SET rating was 4.13 with $SD = 0.50$.

131   Figure 1 shows the smoothed density distributions of overall mean ratings for all courses and
132   for courses in selected subjects -- English, History, Psychology, and Math, including the means and
133   standard deviations. This figure highlights: (1) distributions of ratings are negatively skewed for most
134   of the selected subjects due to ceiling effects, (2) distributions of ratings differ substantially across
135   disciplines, and (3) mean ratings vary substantially across disciplines and are shifted towards lower
136   values by ratings in tails of the distributions. The density distributions in Figure 1 were generated using
137   R function density() with smoothing kernel set to "gaussian" and the number of equally spaced points
138   at which the density was estimated set to 512 (R Core Team, 2015).

139   Figure 2 shows the density distributions for Math (representing quantitative courses) and
140   English (representing humanities, non-quantitative courses). The thick vertical line indicates one of the
141   often used norm-referenced standard for effective teaching -- the overall mean rating across all courses.
142   The thinner vertical lines show the overall mean ratings for Math and English, respectively. This figure
143   highlights that although 71% of English courses pass the overall mean as the standard only 21% of
144   Math courses do so. The vast majority of Math courses (79%) earn their professors an "Unsatisfactory"
145   label in this scenario.

146   Figure 3 shows the same density distribution for Math and English but the vertical lines indicate
147   criterion referenced cut-offs for different levels of teaching effectiveness -- Poor, Fair, Good, Very
148   Good, and Excellent -- as determined by students themselves. It can be seen that Math vs. English
149   courses are far less likely to pass the high (Very Good and Excellent) criteria.

150   Figure 4 shows the percentage of courses passing criteria as a function of teaching effectiveness
151   criteria. If the teaching effectiveness criteria are set at 2.5 ("Good"), the vast majority of both Math and
152   English courses pass this bar (96.60 vs. 99.63%, respectively). However, as the criteria are set higher
153   and higher, the gap between Math and English passing rates widens and narrows only at the high
154   criteria end where a few English and no Math courses pass the criteria.

155   Table 2 shows the percentages of course SETs passing and failing different commonly-used
156   norm-referenced teaching effectiveness criteria as well as label-based criterion-referenced standards,
157   for Math and English courses. The table includes the relative risk ratios of Math vs. English courses
158   failing the standards. Math vs. English courses are far less likely to pass various standards except the
159   label-based, criterion-referenced "Fair" and "Good" standards.

160   Finally, the mean overall SET rating for English courses was 4.29 ($SD = 0.42$) whereas it was
161   only 3.68 ($SD = 0.56$) for Math courses, $t(828.62) = 22.10$, $p < .001$, $d = -1.29$ with $95\% CI = (-1.18,$
162     $-1.40)$. Critically, the correlation between the course subject (Math coded as 1, English coded as 0) and
163   the overall mean rating was $r = -.519$, $95\% CI = (-.553, -.482)$, $p < 0.001$, indicating that Math
164   professors received lower ratings than English professors. In contrast, the correlations between the
165   course subject (Math coded as 1, non-Math courses as 0) and the overall mean ratings when all non-
166   math courses are included, regardless of the degree of their quantitative nature, was relatively smaller, r
167     $= -.172$, $95\% CI = (-.188, -.156)$, $p < 0.001$, with $r^2 = 0.030$.

Discussion

169   Our results show that Math classes received much lower average class summary ratings than
170   English, History, Psychology or even all other classes combined, replicating previous findings showing
171   that quantitative vs. non-quantitative classes receive lower SET ratings (Beran & Violato, 2005; Centra,
172   2009). More importantly, the distributions of SET ratings for quantitative vs. non-quantitative courses
173   are substantially different. Whereas the SET distributions for non-quantitative courses show a typical
174   negative skew and high mean ratings, the SET distributions for quantitative courses are less skewed,
175   nearly normal, and have substantially lower ratings. The passing rates for various common standards
176   for "effective teaching" are substantially lower for professors teaching quantitative vs. non-quantitative

177 courses. Professors teaching quantitative vs. non-quantitative courses are far more likely to fail norm-
178 referenced cut-offs -- 1.88 times more likely to fail the Overall Mean standard, 2.89 times more likely
179 to fail Overall Mean minus 1 SD standard – and far more likely to fail criterion-referenced standards –
180 1.27 times more likely to fail the "Excellent" standard, 3.17 times more likely to fail the "Very Good"
181 standard, and 6.02 times more likely to fail the "Good" standard. Clearly, professors who teach
182 quantitative vs. non-quantitative classes are not only likely to receive lower SETs but they are also at a
183 substantially higher risk of being labeled "unsatisfactory" in teaching, and thus, more likely to be fired,
184 not re-appointed, not promoted, not tenured, and denied merit pay.
185   Regarding norm-referenced vs. criterion referenced standards, our results show that criterion-
186 referenced standards label fewer professors as unsatisfactory than norm-referenced standards. Table 2
187 suggests that, in part due to substantially negatively skewed distributions of SET ratings, the norm-
188 referenced cut-offs Overall Mean standard will result in 43.0% of classes failing to meet the standard,
189 the Overall Mean minus 1 SD standard will result in 15.5% of classes not meeting it, and the Overall
190 Mean minus 2 SD standard will result in 4.3% of classes failing this standard. In contrast, using
191 students' judgments on the anchored scale, 99.3% of courses are considered "Good", "Very Good", or
192 "Excellent" and only 0.7% of courses fail to meet "Good" standards in students' opinion. In other
193 words, use of norm referenced standards results in labeling much greater percentages of professors as
194 unsatisfactory than students themselves label as unsatisfactory. Moreover, professors teaching
195 quantitative vs. non-quantitative courses are less likely to pass the standard under both types of
196 standards.
197   Why has previous research often concluded that TEIFs, such as the courses one is assigned to
198 teach, do not relate to SETs in any substantive way and were ignorable in evaluating professors for
199 tenure, promotion, and merit pay? There are several methodological explanations: First, SET ratings
200 often have non-normal, negatively skewed distributions due to severe ceiling effects. In turn, ds, rs and
201 r^2 based effect size indexes are attenuated, invalid, and inappropriately suggest that influence of course
202 subject on SETs is minimal. Second, parametric effect size indexes such as ds, rs, and r^2 assume normal
203 distributions and are inappropriate for binary "meets standard"/"does not meet standard" decision
204 situations such as tenure, promotion, and merit pay decisions (Deeks, 2002). Third, some researchers
205 used individual student SET ratings rather than class summary evaluations as the unit of analysis.
206 However, using individual student SET ratings as the unit of analysis is inappropriate in this context
207 because summative decisions are made based on class summary evaluations rather than on individual
208 student evaluations.
209   In terms of inappropriate effect sizes such as d or r^2, our results are generally larger than those
210 reported by Centra (2009), who used class summary evaluations from numerous institutions, and to
211 those reported by Beran and Violatto (2005), who used individual SET ratings from a single university.
212 We found $d = 1.29$ between Math vs. English SET ratings, Centra (2009) found $d = .30$, and Beran and
213 Violatto (2005) found $d = 0.60$ between "natural sciences" vs. "social sciences" SET ratings. Our
214 correlational analysis showed $r = 0.18$ ($r^2 = 0.04$) between Math vs. Non-Math and SET ratings,
215 whereas Beran and Violatto (2005) found that this and other factors accounted together for less than 1%
216 of the variance (i.e., $r^2 < 0.01$).
217   However, in contrast to previous research, we examined the impact of courses one is assigned to
218 teach on the likelihood that one is going to pass the standard, and be promoted, tenured, and/or given
219 merit pay and we found the impact to be substantial. Professors teaching quantitative courses are far
220 less likely to be tenured, promoted, and/or given merit pay when their class summary ratings are
221 evaluated against common standards, that is, when the field one is assigned to teach is disregarded.
222 They are also far less likely to receive teaching awards based on their class summary SET ratings. The
223 impact of using common standards may vary depending on whether a university uses the standards
224 based on SET ratings of all professors across the entire university (university based standards) or the
225 standards based on SET ratings of all professors within each department only (department based

226 standards). If all or nearly all professors within the same department teach quantitative courses (e.g.,
227 math and statistics departments), the impact of using common vs. course-type specific department
228 based standards to evaluate professors teaching quantitative vs. non-quantitative courses may be
229 minimal. In contrast, if a few professors teach quantitative courses and the majority of professors teach
230 non-quantitative courses within the same department (e.g., psychology, sociology), the impact of using
231 common vs. course-type specific, department based standards may be as large or even larger than if
232 university based standards were used.
233   Of course the finding that professors teaching quantitative vs. non-quantitative courses receive
234 lower SET ratings is not evidence, by itself, that SETs are biased, that use of the common standards is
235 inappropriate and discriminatory, and that more frequent denial of tenure, promotion, and/or merit pay
236 to professors teaching quantitative vs. non-quantitative courses is in any way problematic. The lower
237 SET ratings of professors teaching quantitative vs. non-quantitative courses may be due to real
238 differences in teaching, that is, due to to professors teaching quantitative vs. non-quantitative courses
239 being ineffective teachers.
240   However, lower SET ratings of professors teaching quantitative vs. non-quantitative courses
241 may be due to a number of factors unrelated to professors' teaching effectiveness, for example,
242 students' lack of basic numeracy, students' lack of interest in taking quantitative vs. non-quantitative
243 courses, students' math anxiety, etc.. Numerous research studies, task forces, and government
244 sponsored studies have documented steady declines in numeracy and mathematical knowledge of
245 populations worldwide. For example, (Orpwood & Brown, 2015) cite the 2013 OECD survey showing
246 that numeracy among Canadians declined over the last decade and that more than half of Canadians
247 now score below the level required to fully participate in a modern society. We (Uttl, White, & Morin,
248 2013) found that students' interest in taking quantitative courses such as introductory statistics was six
249 standard deviations below their interest in taking non-quantitative courses. Fewer than 10 out of 340
250 students indicated that they were "very interested" in taking any of the three statistics courses. In
251 contrast, 159 out of 340 were "very interested" in taking the *Introduction to the Psychology of*
252 *Abnormal Behavior* course. Moreover, this effect was stronger for women than for men: women's
253 interest in taking quantitative courses relative to their interest in non-quantitative courses was even less
254 than that of men. This lack of interest in quantitative courses propagates to lack of student interest in
255 pursuing graduate studies in quantitative methods and lack of quantitative psychologists to fill all
256 available positions. For example, the American Psychological Association noted that in the 1990s
257 already there were on average 2.5 quantitative psychology positions advertised for every quantitative
258 psychology PhD graduate (APA, 2009). If SETs are biased or even perceived as biased against
259 professors teaching quantitative courses, we may soon find out that no one will be willing to teach
260 quantitative courses if they are evaluated against the common standard set principally by professors
261 who teach non-quantitative courses.
262   Thus, the critical question is: Are SETs valid measures of teaching effectiveness, and if so, are
263 they equally valid when used with quantitative vs. non-quantitative courses or are they biased?
264 Although SETs are widely used to evaluate faculty's teaching effectiveness, their validity has been
265 highly controversial. The strongest evidence for the validity of SETs as a measure of professors'
266 teaching effectiveness were so called multi-section studies showing small-to-moderate correlations
267 between class summary SET ratings and class average achievement (Uttl, White, & Gonzalez, 2016).
268 Cohen (1981) conducted the first meta-analysis of multi-section studies and reported that SETs
269 correlate with student learning with $r = .43$ and concluded "The results of he meta-analysis provide
270 strong support for the validity of student ratings as a measure of teaching effectiveness" (p. 281).
271 Cohen's (1981) findings were confirmed and extended by several subsequent meta-analyses (Uttl et al.,
272 2016). However, our recent re-analyses of the previous meta-analyses of multi-section studies found
273 that their findings were artifacts of small study bias and other methodological issues. Moreover, our up-
274 to-date meta-analysis of 97 multi-section studies revealed no significant correlation between the class

275 summary SET ratings and learning/achievement (Uttl et al., 2016). Thus, the strongest evidence of SET
276 validity – multisection studies – turned out to be evidence of SETs having zero correlation with
277 achievement/learning. Moreover, to our knowledge, no one has examined directly whether SETs are
278 equally valid or biased measures of teaching effectiveness in quantitative vs. non-quantitative courses.
279 Even the definition of effective teaching implicit in multi-section study designs – a professor whose
280 students score highest on the common exam administered in several sections of the same courses is the
281 most effective teacher – has been agreed on only for lack of a better definition.
282   The basic principles of fairness require that the validity of a measure used to make high-stakes
283 personnel decisions ought to be established before the measure is put into widespread use, and that the
284 validity of the measure is established in all different contexts that the measure is to be used in (AERA,
285 APA, & NCME, 2014; APA, 2004). Given the evidence of zero correlation between SETs and
286 achievement in multi-section studies, SETs should not be used to evaluate faculty's teaching
287 effectiveness. However, if SETs are to be used in high stakes personnel decisions – even though
288 students do not learn more from more highly rated professors and even though we do not know what
289 SETs actually measure – fairness requires that we evaluate a professor teaching a particular subject
290 against other professors teaching the same subject rather than against some common standard. Used
291 this way, SET ratings can at least tell us where a professor stands within the distribution of other
292 professors teaching the same subjects, regardless of what SETs actually measure.

Conclusion

294   Our results demonstrate that course subject is strongly associated with SET ratings and has a
295 substantial impact on professors being labeled satisfactory/unsatisfactory and excellent/non-excellent.
296 Professors teaching quantitative courses are far more likely not to receive tenure, promotion, and/or
297 merit pay when their performance is evaluated against common standards. Moreover, they are unlikely
298 to receive teaching awards. To evaluate whether the effect of some TEIFs is ignorable or unimportant
299 should be done using effect size measures that closely correspond to how SETs are used to make high
300 stakes personnel decisions such as passing rates and relative risks of failures rather than ds or rs. A
301 professor assigned to teaching introductory statistics courses may find little solace in knowing that
302 teaching quantitative vs non-quantitative courses explain at most 1% of variance in some regression
303 analyses of SET ratings (Beran & Violato, 2005) or that in some experts' opinion $d = .30$ is ignorable
304 (Centra, 2009) when his or her chances of passing the department's norm based cut off for
305 "satisfactory" teaching may be less than half of his colleagues passing the norms.

306 Acknowledgements

307 We thank Amy L. Siegenthaler and Alain Morin for careful reading and comments on the manuscript.

308
309
310

311 References

312

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington,

D.C: American Educational Research Association.

APA. (2004). *Code of Fair Testing Practices in Education*. Washington, D.C: American Psychological

Association.

APA. (2009). *Report of the Task Force for Increasing the Number of Quantitative Psychologists*.

Washington, D.C: American Psychological Association.

Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and

literature. IDEA Center.

Beran, T., & Violato, C. (2005). Ratings of University Teacher Instruction: How Much Do Student and

Course Characteristics Really Matter? *Assessment and Evaluation in Higher Education, 30*(6),

593–601.

Bond, C. F. J., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences.

Psychological Methods, 8(4), 406–418. https://doi.org/10.1037/1082-989X.8.4.406

Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Princeton,

NJ: Educational Testing Service.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of

multisection validity studies. *Review of Educational Research, 51*(3), 281–309.

https://doi.org/10.2307/1170209

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with

binary outcomes. *Statistics in Medicine, 21*(11), 1575–1600. https://doi.org/10.1002/sim.1188

Felton, J., Koper, P., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues:

Student evaluations of professors on ratemyprofessors.com. [References]. *Assessment &*

Evaluation in Higher Education, 33(1), 45–61.

https://doi.org/http://dx.doi.org/10.1080/02602930601122803

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The

relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher*

Education, 29(1), 91–108.

Orpwood, G., & Brown, E. S. (2015). *Closing the numeracy gap*. Toronto, ON, Canada.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R

Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Uttl, B. (2005). Measurement of Individual Differences: Lessons From Memory Assessment in

Research and Clinical Practice. *Psychological Science, 16*(6), 460–467.

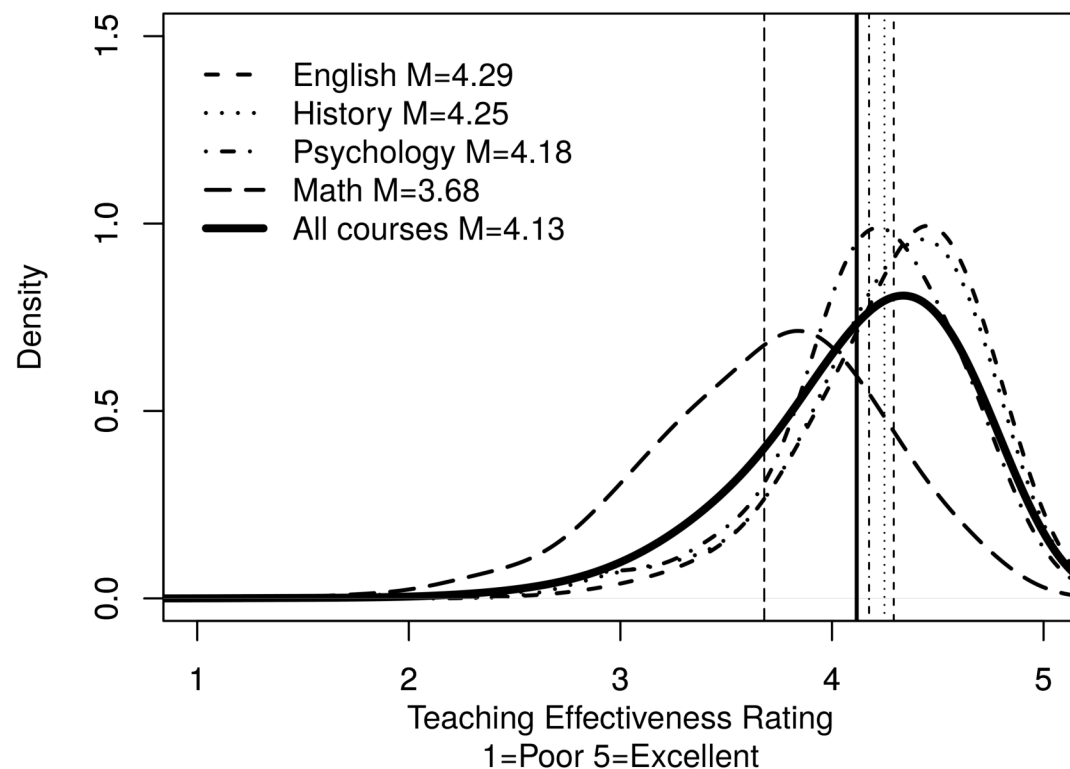
Uttl, B., White, C. A., & Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness:

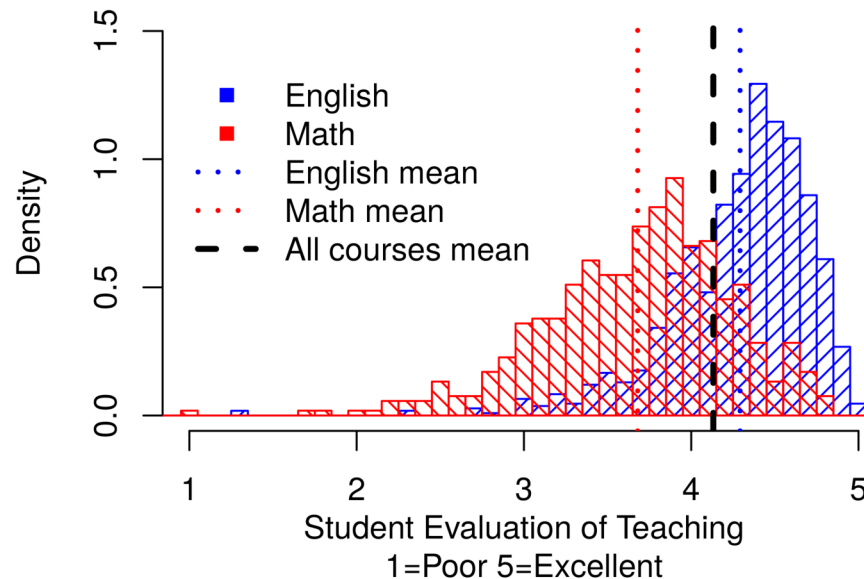
Student evaluation of teaching ratings and student learning are not related. *Studies in*

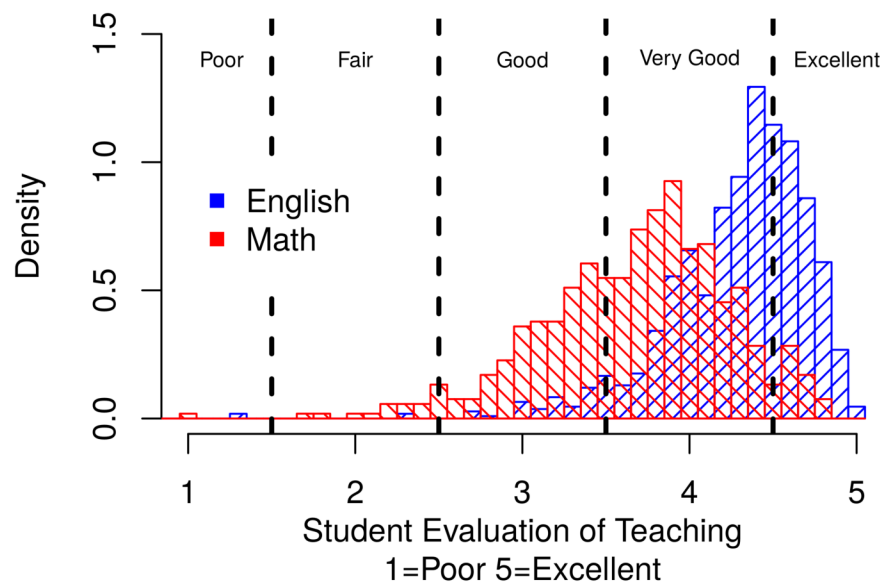
Educational Evaluation. https://doi.org/10.1016/j.stueduc.2016.08.007

Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: students don't like numbers! *Plos*

One, 8(12), e83443–e83443. https://doi.org/10.1371/journal.pone.0083443

313 **Figures and Figure Legends**
314
315
316
317
318

319 *Figure 1*. **Distributions of overall mean ratings for all courses and for courses in selected subjects.**
320 The figure shows the smoothed density distributions of overall mean ratings for all courses and for
321 courses in English, History, Psychology, and Math, including the means and standard deviations. The
322 figure highlights: (1) distributions of ratings are negatively skewed for most of the selected subjects
323 due to ceiling effects, (2) distributions of ratings differ substantially across disciplines, and (3) mean
324 ratings vary substantially across disciplines and are shifted towards lower values by ratings in tails of
325 the distributions.
326

327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353

354 *Figure 2*. **Distributions of overall mean ratings for Math vs. English.** The thick vertical line
355 indicates one of the often used norm-referenced standard for effective teaching -- the overall mean
356 rating across all courses. The thinner vertical lines show the overall mean ratings for Math and English,
357 respectively. Although 71% of English courses pass the overall mean as the standard, only 21% of
358 Math courses do so. The vast majority of Math courses (79%) earn their professors an "Unsatisfactory"
359 label in this scenario.
360

361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387

388 *Figure 3.* **Distribution of overall mean ratings for Math vs. English with criterion referenced cut**
389 **offs.** The vertical lines indicate criterion referenced cut-offs for different levels of teaching
390 effectiveness -- Poor, Fair, Good, Very Good, and Excellent -- as determined by students themselves. It
391 can be seen that Math vs. English courses are far less likely to pass the high (Very Good and Excellent)
392 criteria.
393

394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429

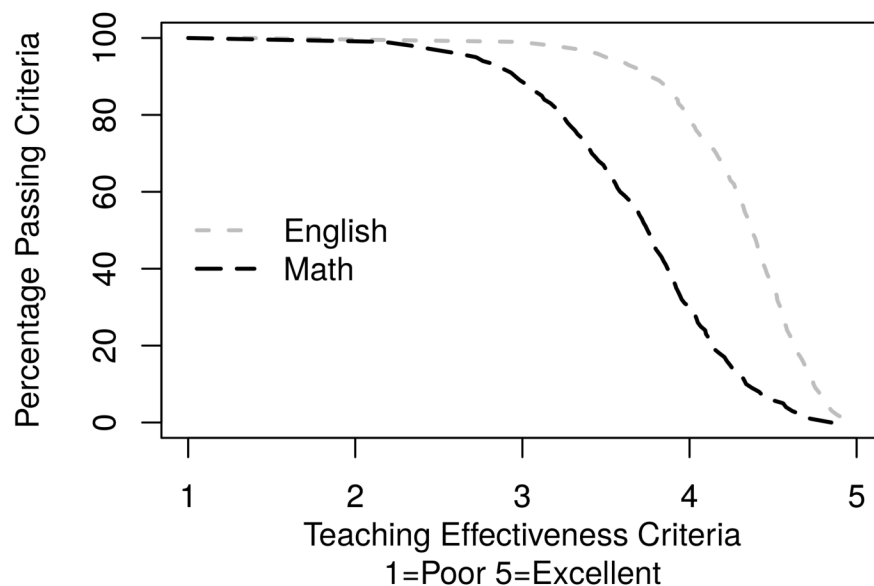



Figure 4. **Percentage of courses passing criteria as a function of teaching effectiveness criteria.** If the teaching effectiveness criteria are set at 2.5 ("Good"), the vast majority of both Math and English courses pass this bar (96.60 vs. 99.63%, respectively). However, as the criteria are set higher and higher, the gap between Math and English passing rates widens and narrows only at the high criteria end where a few English and no Math courses pass the criteria.

430
431 Table 1
432 *SET Questions, Mean Ratings, and Standard Deviations.*

Question	M	SD
1. How would you rate the instructor overall?	4.37	0.55
2. How informative were the classes?	4.26	0.52
3. How well organized were the classes?	4.19	0.55
4. How fair was grading?	4.14	0.55
5. How would you rate this course overall?	4.09	0.57
6. How clear were the objectives of this course?	4.12	0.52
7. How well were these objectives achieved?	4.10	0.53
8. How interesting was the course?	4.01	0.63
9. To what extent were your own expectations met?	3.90	0.58
Mean overall rating (across all items)	4.13	0.50

433 Note: $N = 14,872$
434

435 Table 2
436 *Percentages of Course SETs Passing vs. Failing Different SET Standards and Relative Risk of Failing*
437 *to Achieve the Standards for Math Courses.*

Criteria Cut-offs	All Pass (%)	All Fail (%)	Math Pass (%)	Math Fail (%)	English Pass (%)	English Fail (%)	Math v. Non-Math RR of Failure incl. 95% CI	Math vs. English RR of Failure incl. 95% CI
Norm-referenced								
Mean (4.13)	57.0	43.0	21.4	78.6	71.3	28.7	1.88* (1.80,1.98)	2.74* (2.47,3.05)
Mean Minus 1 SD (3.63)	84.5	15.5	58.0	42.0	93.1	6.9	2.89* (2.60,3.22)	6.05* (4.76,7.70)
Mean Minus 2 SD (3.13)	95.7	4.3	83.9	16.1	98.6	1.4	4.12* (3.34,5.09)	11.59* (6.76,19.87)
Criterion-referenced								
Excellent (4.50)	25.4	74.6	5.9	94.1	35.4	64.6	1.27* (1.24,1.30)	1.46* (1.39,1.53)
Very Good (3.50)	88.5	11.5	66.0	34.0	94.9	5.1	3.17* (2.80,3.61)	6.69* (5.04,8.89)
Good (2.50)	99.3	0.7	96.6	3.4	99.6	0.4	6.02* (3.64,9.97	9.20* (3.13,27.06)
Fair (1.50)	99.9	0.1	99.8	0.2	99.8	0.2	4.52 (0.54,37.47)	1.02 (0.09,11.25)

438 *Note:* All courses N = 14,872; English courses n = 1082; Math courses n = 529; Non-Math courses n =
439 14,343. *p < .001
440
441