Student Evaluation of Teaching: Facts and Myths

Bob Uttl

2019/05/31

Student Evaluations of Teaching (SET)

Formative vs. Summative Use of SETs

Formative uses

- SETs inform professors what students' perceptions were
- Professors may use SETs to modify their instruction

Summative uses

- SETs used as a measure of teaching effectiveness
- SETs used for hiring, firing, merit, promotion, and tenure decisions

Student Evaluation of Teaching (SET)

Fundamentals of Perception

Fundamentals of perception

- Perception is an interpretation of sensory information based on prior knowledge, beliefs, experiences, etc.
  - Two perceivers may interpret the same information differently
  - The same perceiver may interpret the same information differently on different occassions

Students – perceivers of teaching – vary in

- Cognitive ability (e.g., memory, intelligence)
- Prior knowledge
- Interests and motivation
- Prejudices
- ...

Clearly, students are NOT disinterested, objective, and trained evaluators of effective teaching...

Student Evaluations of Teaching (SET)

Effective Teaching

What is effective teaching? Experts

- Do not agree as to what effective teaching is
- Agree that some behaviors are not effective teaching
  - not showing up for your classes
  - using class time to read textbook chapters verbatim
  - ...
- Agree that effective teaching results in learning

Fundamental problem

- How to evaluate something without knowing what it is?
- How to set standards of performance for this unknown?

A simple solution to a complex problem

- Asks students if professor is effective... Use SET...
- Use university or departmental SET averages as "norms"...
- Use SET to make high stakes personnel decisions

Student Evaluations of Teaching (SET)

Typical Process

- SETs are administered within the last few weeks of courses
- Students rate professors on 5-point Likert scale
- Evaluation unit produces summary ratings for each class
- Summaries include means, SDs, frequencies, etc.
- Summaries may include the departmental or university "norms"
- Summaries are distributed to professors, chairs, and deans
- No standards for satisfactory performance are provided
- SETs are key evidence of teaching effectiveness
- Chairs, deans and TPC members (evaluators) do not understand numbers
- Evaluators believe in different satisfactory standards
- Evaluators change periodically and unpredictably
- ...

Student Evaluations of Teaching (SET)

Reasons for Using SETs for Summative Decisions (e.g., Murray, 2005)

- SETs are cheap and convenient means to evaluate faculty
- SETs are useful for public accountability and relations
- SETs allow students to have a say in professors' evaluations
- Students are uniquely positioned to evaluate faculty as they are (sometimes) in class when professors are teaching

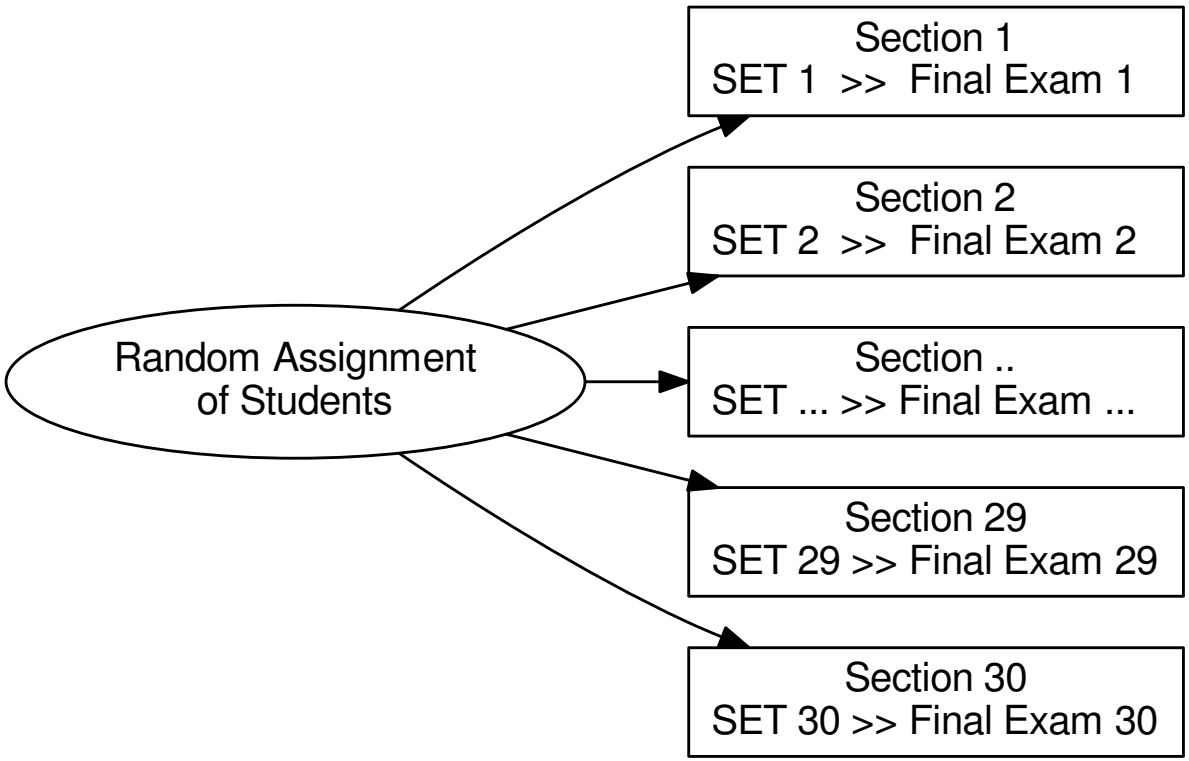
Student Evaluations of Teaching (SET)

Reasons for NOT Using SETs for Summative Decisions

- SET measure student satisfaction
  - "a happy or pleased feeling because of something that you did or something that happened to you" (Merriam-Webster)
- SET are influenced by Teaching Effectiveness Irrelevant Factors (TEIFs)
  - Students' intelligence, interests, motivation...
  - Academic discipline, class level, class size,...
  - Profesors' gender, beauty/hotness, accent, national origin, ...
  - Professors' academic standards, provision of chocolates, ...
  - ...
- SET have insufficient validity to be used in high stake personnel decisions
  - Cohen (1981) reported r = .43 between SETs and learning
  - SETs explain at best 16% of variance in section learning
  - SETs do not explain 84% of variance in section learning

Validity of SETs: Multisection Studies (MSS)

Logic



Validity of SETs

Cohen (1981): r = .43 between SETs and learning

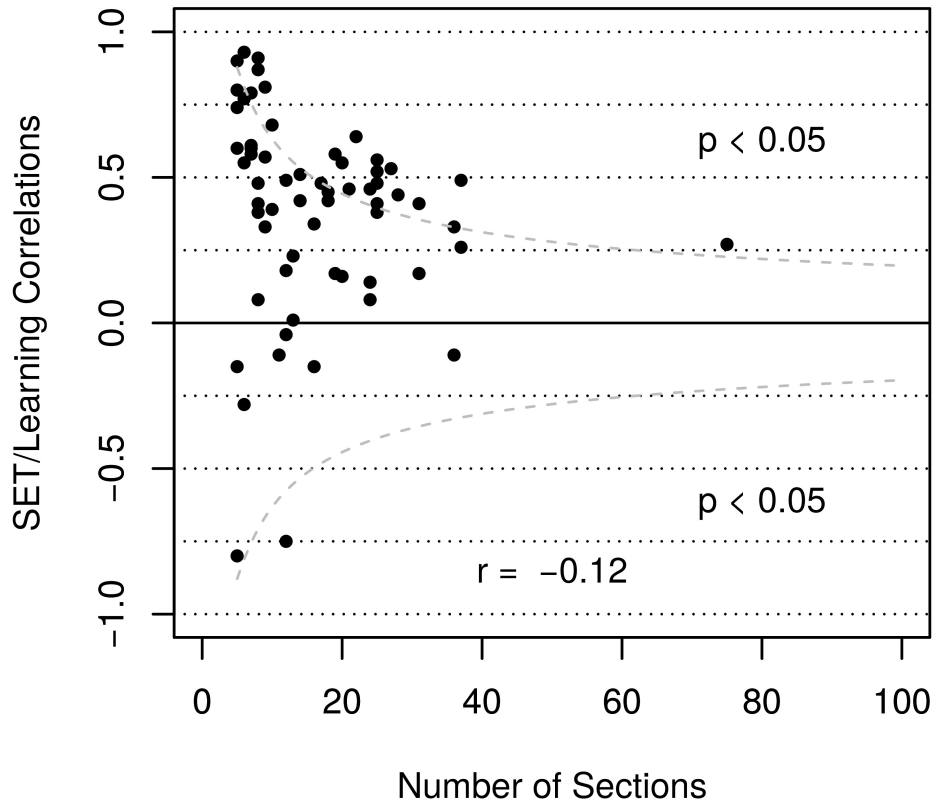
"The results of the meta-analysis provide strong support for the validity of student ratings as a measure of teaching effectiveness" (p. 281)

"... we can safely say that student ratings of instruction are a valid index of instructional effectiveness. Students do a pretty good job of distinguishing among teachers on the basis of how much they have learned." (p. 305)

Cohen (1981): Instructor Rating

Scatterplot

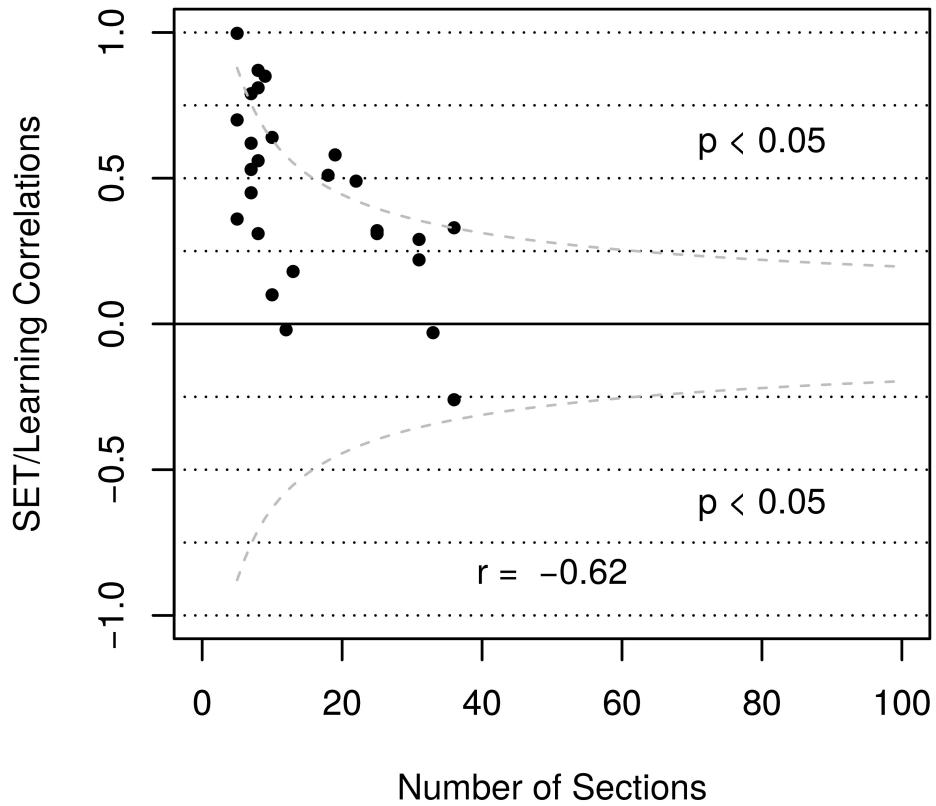
SET/Learning r = .43



Feldman (1989): Preparation and Organization

Scatterplot

SET/Learning r = .55

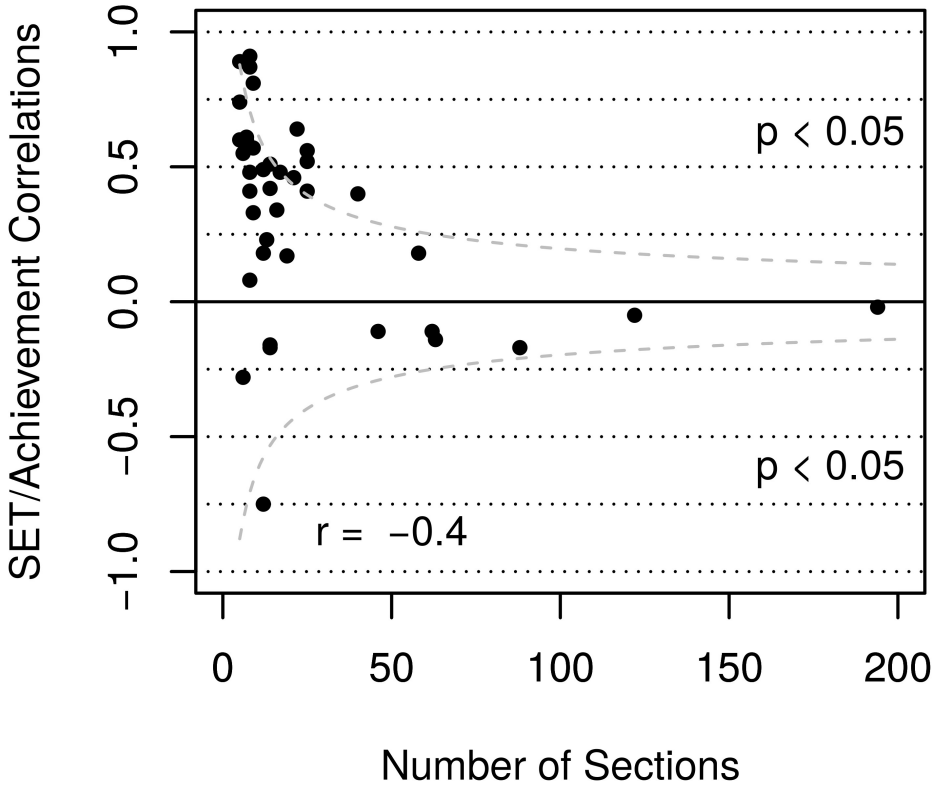


Clayson (2009): Instructor Rating

Scatterplot

SET/Learning r = .33
SET/Learning r_w = .13

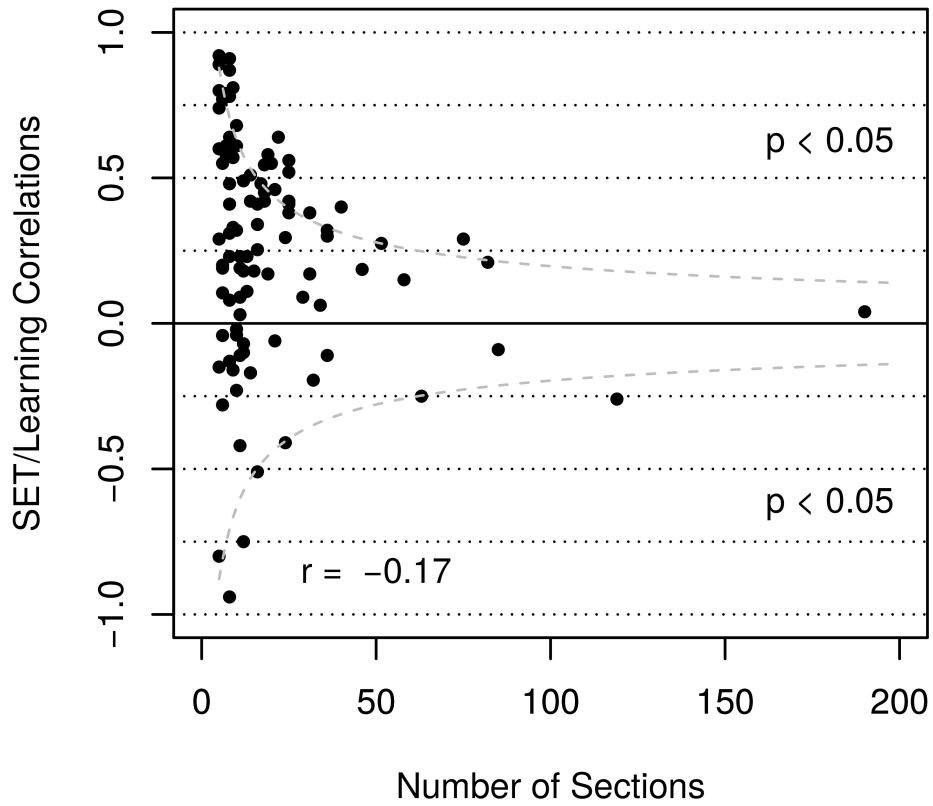
Inexplicably, used Cohen's (1981) meta-analysis r = .41 (corrected) as if it were a single study with 35 sections.



Uttl et al. (2016): Instructor SET/Learning Correlations

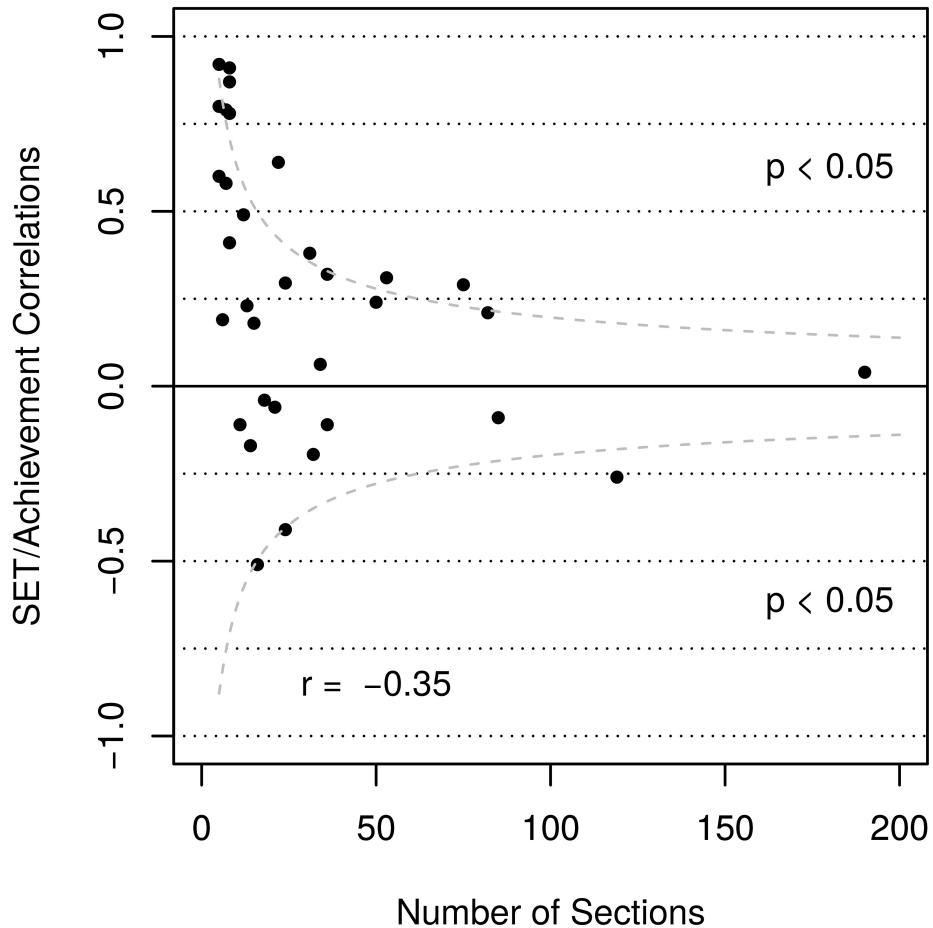
Scatterplot

$k = 97$



Uttl et al. (2016): Instructor Rating (Ability Adjusted rs)

Scatterplot



Meta-Analyses of Multisection Studies (MSS)

Conclusions

- The findings reported in previous meta-analyses (Clayson, 2009; Cohen, 1981; Feldman, 1989) are artifacts of poor meta-analytic methods
- MSS typically included very few sections
- Scatterplots etc. show strong small study size effects
- Analyses of all rs show very weak SET/Learning correlations (up to 1% variance explained)
- Analyses of only rs adjusted for prior ability/knowledge show zero SET/Learning correlations
- Above holds both for overall instructor ratings as well as for averages of all SET ratings
- **SET do not measure faculty's teaching effectiveness; students do not learn more from more highly rated professors.**

Meta-Analyses of Multisection Studies (MSS)

Implications

- Universities and colleges focused on student learning and student future success may need to give minimal or no weight to SET ratings.
- Universities and colleges focused on students' perception or satisfaction rather than learning may want to evaluate their faculty's teaching effectiveness using primarily or exclusively SET ratings
  - terminate all faculty who do not exceed the average SET department or university ratings...
  - expunge courses students do not like from university curricula
  - give students As
  - provide chocolate chip cookies
  - pay students for attendance
  - ...

Validity of SET: Effects of TEIFs

TEIFs NOT Attributable to Professors

SETs are influenced by TEIFs

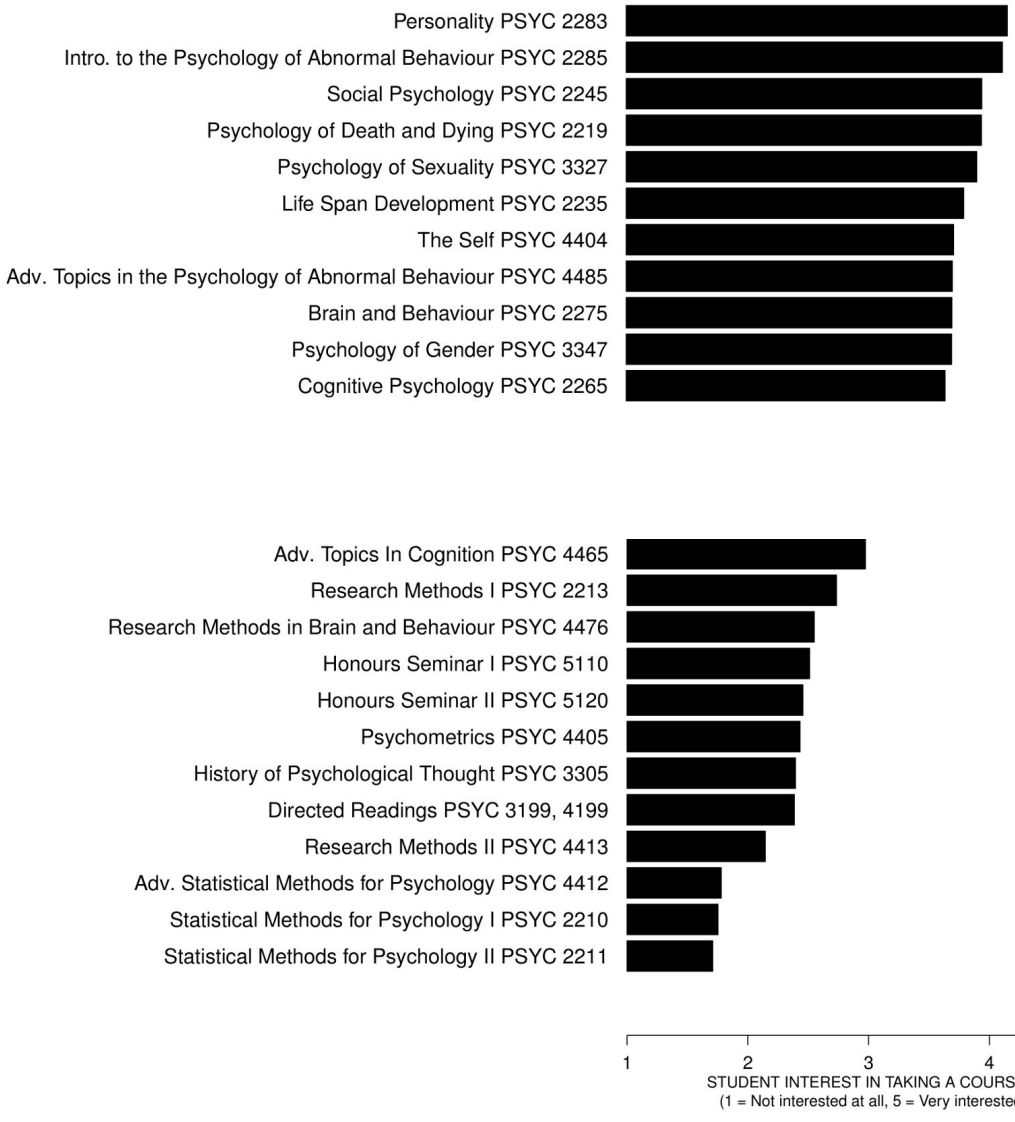
- Students' ability & prior knowledge
- Students' interest in a course
- Students' motivation
- Students' prejudice
- ...
- Course type (e.g, quantitative vs. non-quantitative)
- Class size
- Class level
- ...

These TEIFs are not attributes of professors. Ergo, if these TEIFs influence SET, SET are less valid as a measure of something about professors.

TEIFs: Student Interest

Quantitative vs. Non-quantitative Courses (Uttl, White,& Morin, 2013)

- 340 students
- rating interest
  - 1 = Not at all interested
  - 5 = Very interested
- 44 courses
  - 3 high QC
  - 6 moderate QC
  - 34 low QC



TEIFs: Student Course Interest

Conclusions

- Students have minimal interest in taking courses with any substantive quantitative content.
- Out of 340 students, fewer than 10 were "very interested" in taking any of the three statistics courses.
- Out of 340 students, nearly half – 159 – were "very interested" in taking abnormal psychology
- Faculty teaching quantitative courses find themselves facing students who do not want to be in their courses.

TEIFs: Student Course Interest

Implications

- Using the same SET standards for faculty teaching quantitative vs. non-quantitative courses is inappropriate
- If the same SET standards are used, faculty are under pressure to dumb down quantitative courses
- Universities focused on student learning may want to abolish SETs and focus on evaluation of student learning
- Universities focused primarily on student satisfaction may want to expunge quantitative courses from their curriculum

TEIFs: SET of Quantitative vs. Non-Quantitative Courses

Claim: Efffects of TEIFs are not important, ignorable

- SET correlate with various TEIFs
- Some have argued that those correlations are small and do not undermine validity of SETs (Beran & Violatto, 2005)
  - d = .61 b/w natural vs. social science
  - Regression analyses [over individual rather than course SETs] showed that TEIFs including the discipline were not important ($< 1\%$ var. explained)

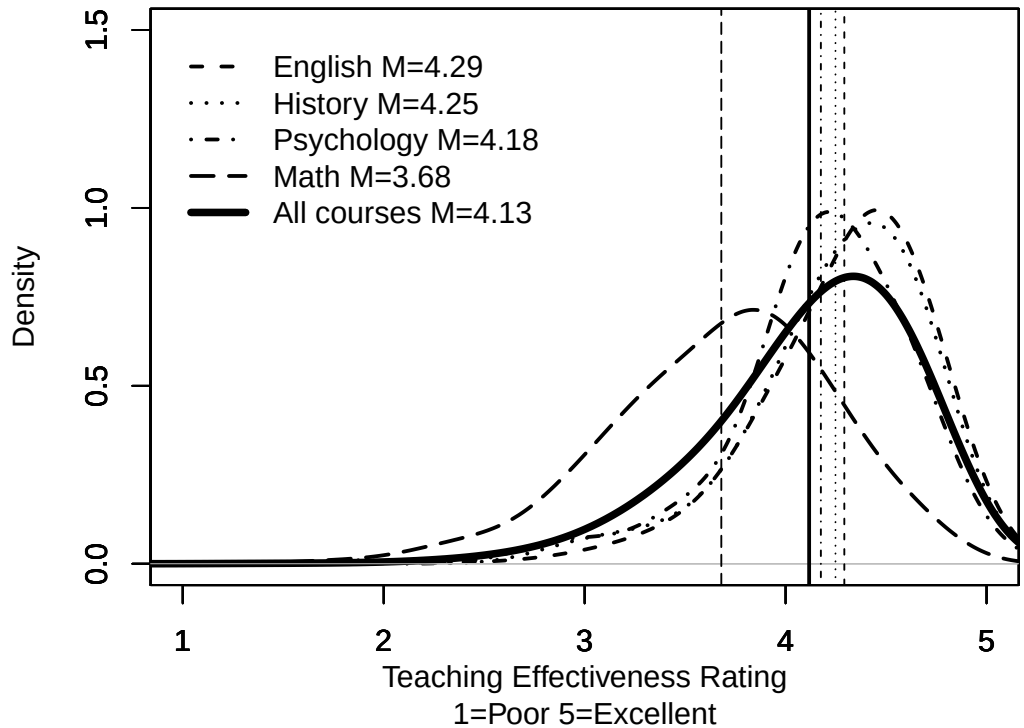
"From examining numerous student and course characteristics as possible correlates of student ratings, results from the present study suggest that they are not important factors." (Beran & Violatto, 2005, p. 599)

"... a third standard deviation does not have much practical significance." (Centra, 2009)

TEIFs: SETs of Quantitative vs. Non-Quantitative Courses

Distributions and Means (Uttl & Smibert, 2017)

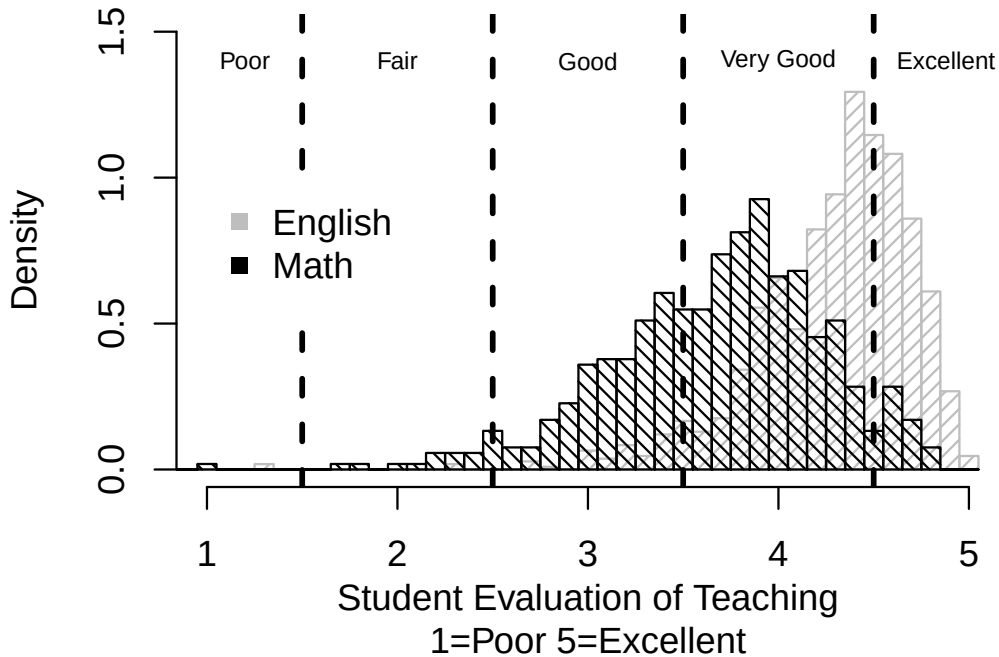
- 14,872 courses
- 1,082 English courses
- 529 Math courses
- midsize university



TEIFs: SETs of Quantitative vs. Non-Quantitative Courses

Math vs English Profs SETs and Standards (Uttl & Smibert, 2017)

Fewer Math Profs pass any given standards than English Profs.



TEIFs: SET of Quantitative vs. Non-Quantitative Courses

Conclusions (Uttl & Smibert, 2017)

- Course subject has strong association with SET
- Course subject has substantial impact on professors being labeled satisfactory vs. non-satisfactory, excellent vs. non-excellent
- The impact varies depending on the criteria used for classification
- Professors teaching quantitative vs. non-quantitative courses are far more likely not to receive tenure, promotion, and/or merit pay when performance is evaluated against common standards

TEIFs: SETs of Quantitative vs. Non-Quantitative Courses

Implications (Uttl & Smibert, 2017)

- To evaluate whether TEIFs are ignorable, one should use effect sizes that closely correspond to how SETs are used to make high stakes personnel decisions
- If SET are to be used for evaluation, regardless of what they measure, professors teaching a specific subject should be evaluated against professors teaching the same subject rather than against common standards

SET measure students' perceptions or students' satisfaction with something but this satisfaction depends on many TEIFs.

TEIFs: Attributes of Professors

Use of some TEIFs is ill-advised and/or illegal

SETs are influenced by TEIFs attributable to professors

- Gender
- Accent
- National origin
- Ethnicity
- Aproachability
- Beauty/hotness
- ...
- Provision of chocolates and cookies

Use of SET in high stake personnel decisions is (a) violating various human rights legislations and (b) at minimum unwise with respect to cookies and chocolates.

TEIFs: Attributes of Professors

Availability of cookies (Hessler et al.. 2018)

Anectodal evidence suggests that professors can influence SETs by providing their students with chocolates, cookies, etc..

Hessler et al. (2018) reported on the first randomized controlled trial (RCT) investigating whether availability of **chocolate** cookies improves SETs. It does:

- Cookie group evaluated teachers significantly better than the control group (d = .68)
- Cookie group though the course material was better (d = .66)
- Cookie group evaluated the course overall as better (d = .51)

SETs: Summation 1

Summary

- Contrary to popular beliefs, multisection studies show zero correlation betweeen SETs and learning/achievement. Students do not learn more from more highly rated professors.
- TEIFs have substantial associations with SETs and substantial impact on classifying professors as satisfactory vs. non-satisfactory. Many of these TEIFs are not attributes of professors (e.g., student interest, ability, prior knowledge), rendering SETs invalid measures of professors, regardless of what they actually measure.
- Other TEIFs with substantial associations with SETs – sex/gender, accent, national origin, ethnicity, beuty/hotness, distribution of chocolate, etc. – are attributes of professors but their use triggers anti-discrimination laws and/or is ill-advised.

Conflict of Interest and SET-Learning Correlations

Money, jobs, career, ... (Uttl, Cnudde, & White (in review)

> *It is difficult to get a man to understand something when his salary depends upon his not understanding it.*
> *(Upton Sinclair)*

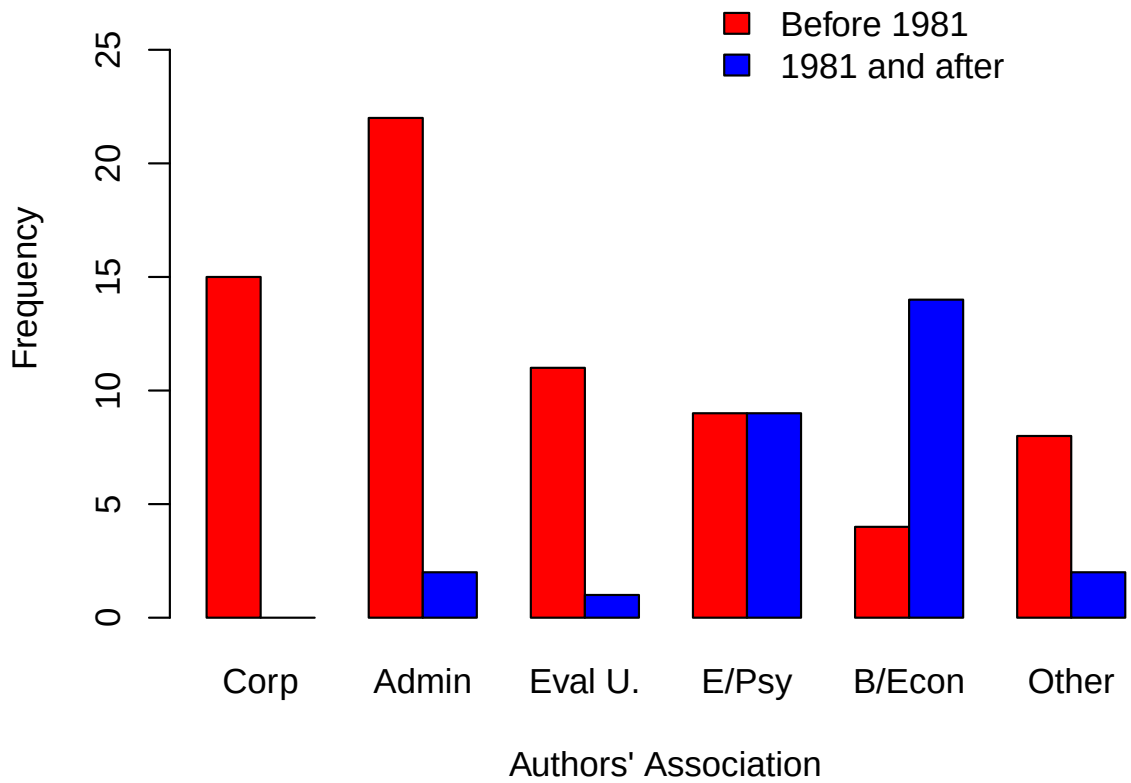
Are authors' financial and other ties (profits, salaries, career, etc.) related to the size of SET-Learning correlations reported in multisection studies?

- Corporate interests
- Administrative interests
- Evaluation unit interests
- SET author interests

Conflict of Interest and SET-Learning Correlations

Early vs. late studies (Uttl, Cnudde, & White, in review)

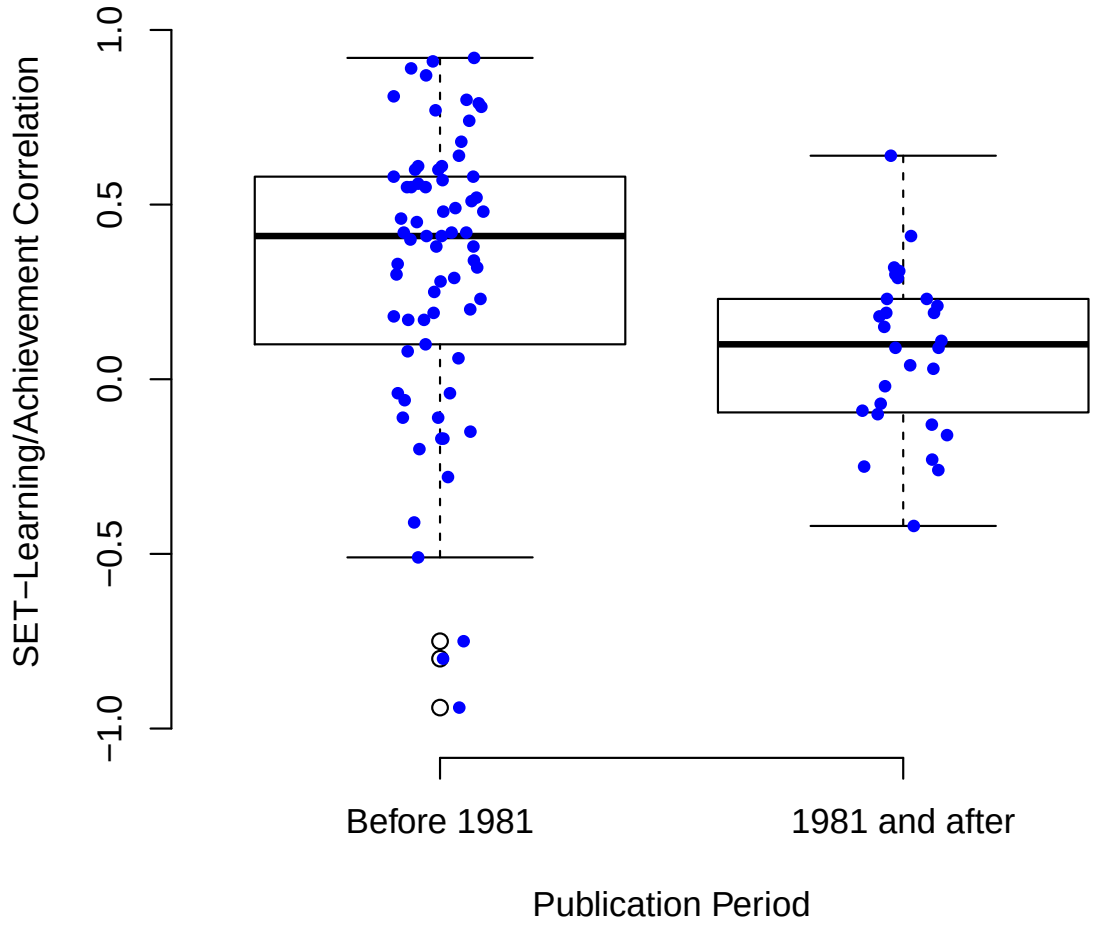
Earlier studies were done primarily by corporate, administrative, and evaluation unit authors...



Conflict of Interest and SET-Learning Correlations

Early vs. late studies (Uttl, Cnudde, & White, in review)

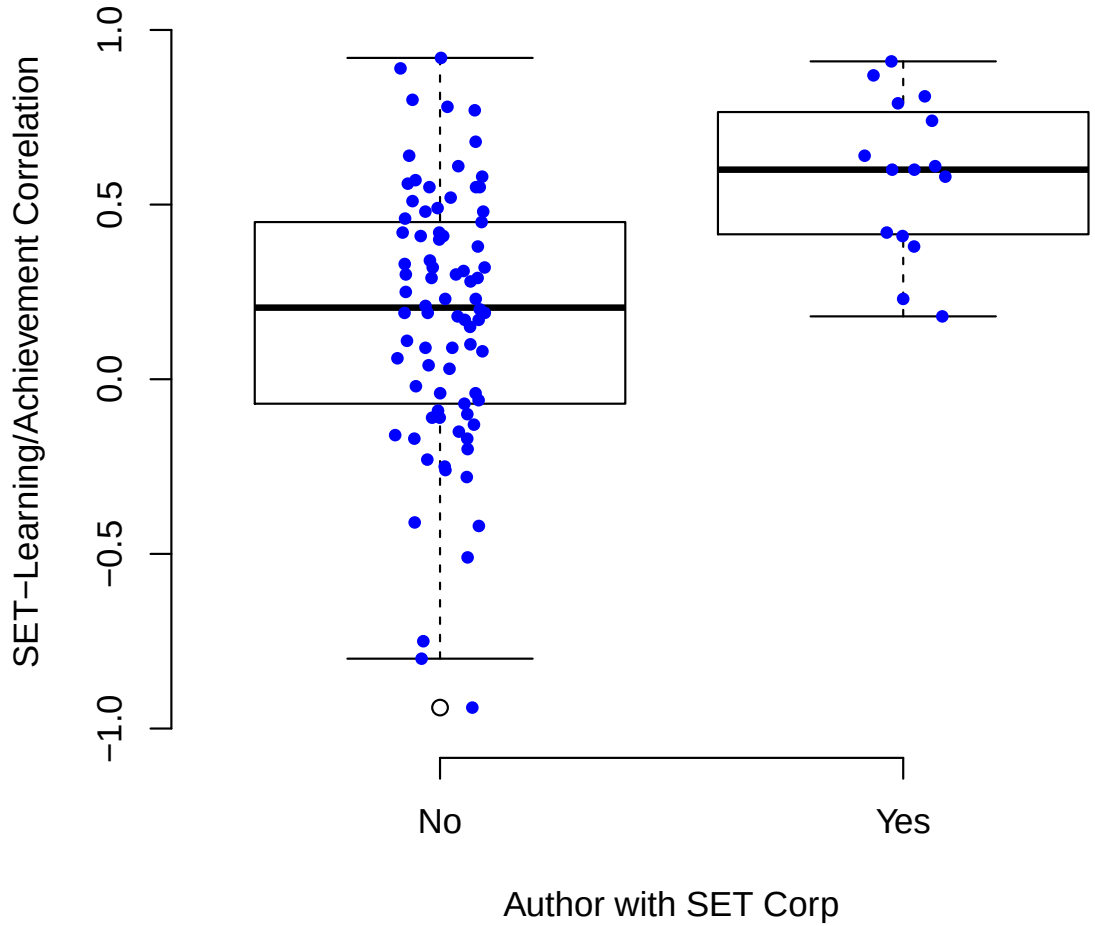
Earlier studies found larger SET-Learning correlations than later studies...



Conflict of Interest and SET-Learning Correlations

SET Corporations vs. Not SET Corporation (Uttl, Cnudde, & White, in review)

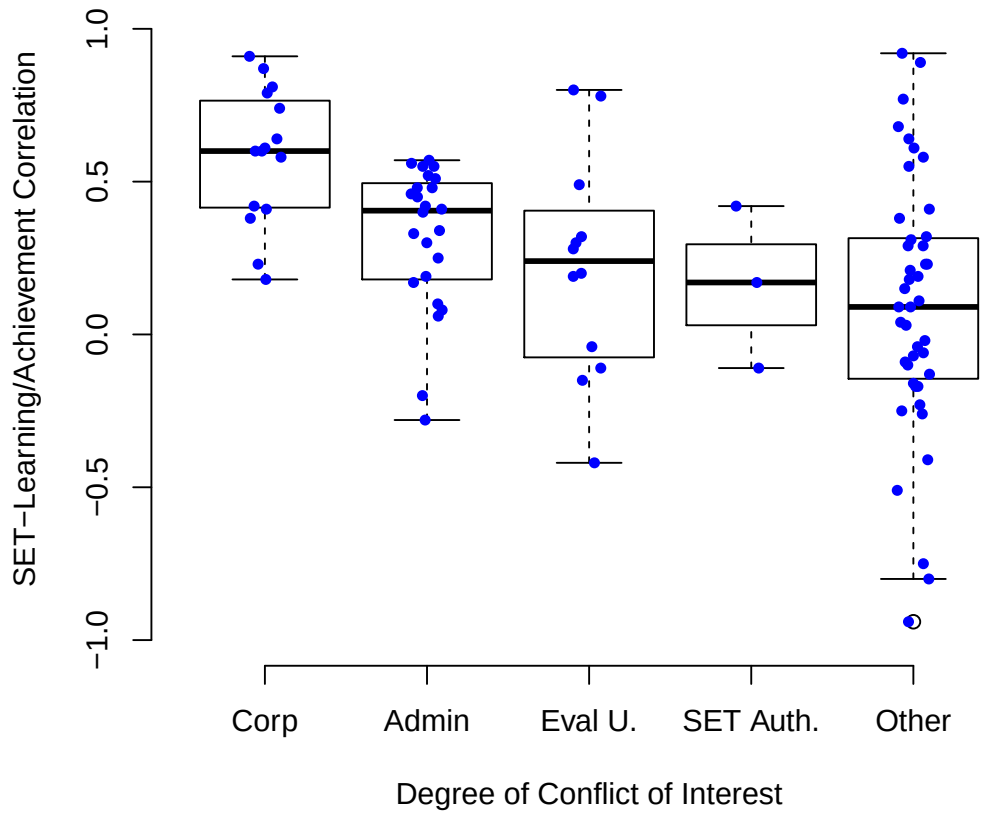
Authors from SET corporations found larger SET-Learning correlations than other authors...



Conflict of Interest and Size of SET-Learning Correlations

Degree of Conflict of Interest (Uttl, Cnudde, & White, in review)

Conflict of interest is strongly associated with SET-Learning Correlations



Conflict of Interest

Summary

- Conflicts of interests explain historical pattern of SET-Learning correlations in MSS
  - Authors with COIs report large SET-Learning correlations
  - Authors with no identifiable COIs report minimal SET-Learning correlations

SET: Legal/Employment Law

Ryerson University v. The Ryerson Faculty Association (2018 CanLII 58446)

> *"That evidence, as earlier noted, was virtually uncontradicted. It establishes, with little ambiguity, that a key tool in assessing teaching effectiveness is flawed, while the use of averages is fundamentally and irreparably flawed. It bears repeating: the expert evidence called by the Association was not challenged in any legally or factually significant way...*
>
> *"The collective agreement is to be amended to ensure that FCS [Ryerson University SETs] results are not used to measure teaching effectiveness for promotion and tenure...."*

See the following for links to the decision and to the expert reports:
https://ocufa.on.ca/blog-posts/significant-arbitration-decision-on-use-of-student-questionnaires-for-teaching-evaluation/

SETs: Summation 2

Fatally flawed and not suitable for evaluation of faculty

- do not measure teaching effectiveness
- measure student satisfaction (happy pleased feeling...)
- are influenced by TEIFs not attributable to professors
- are influenced by TEIFs attributable to professors but illegal/discriminatory or not advisable
- are likely illegal, violating human rights codes
- are ultimately unsuitable for evaluation of faculty

SET and Ethics

Canadian Code of Ethics for Psychologists (4th Edition)

- Principle I: Respect for the Dignity of Persons and Peoples
  - non-discrimination, fairness,..
- Principle II: Responsible Caring
  - "benefit members of society, or at least, do no harm"
- Principle III: Integrity in Relationships
  - "... committment to truthfulness..."
- Principle IV: Responsibility to Society
  - call out "incompetent and unethical behavior, including misinterpreations or misuses of psychological knowledge and techniques..."

SET and Ethics

AERA/APA/NCME (2013). The Standards for Psychological and Educational Testing

The Standards

- 1.1 Validity
- 1.2. Reliability
- 1.3. Fairness in testing
- 2.1 Test design and development
- 2.2 Scores, scales, score linking and cut scores
- ...
- 3.4 Uses of tests for program evaluation, policy studies and accountability

Big Question

How do we evaluate teaching? Alternatives to SET?

- teaching dossiers including peer-reviews (CAUT, OCUFA)
  - However, if no one knows what effective teaching is, can peer-reviews be valid? No...
    - condescending fontsize
    - intimidating Socratese method
    - stressful pop-quizzes
    - time consuming and costly frequent testing
    - ...
- ineffective teaching
  - not showing up for classes
  - reading textbook verbatim during class time
  - not providing feedback to students
  - ...

Professor's Assessment of Student Success (PASS)

In my opinion, this student

- SD D N A SA 1. was organized
- SD D N A SA 2. valued intellectually stimulating courses
- SD D N A SA 3. was genuinely interested in learning
- SD D N A SA 4. valued being encouraged to think
- SD D N A SA 5. was well prepared for each class
- SD D N A SA 6. was genuinely interested in getting help
- SD D N A SA 7. was open to learn a great deal
- SD D N A SA 8. facilitated atmosphere conducive of learning
- SD D N A SA 9. communicated clearly
- SD D N A SA 10. learned a lot
- SD D N A SA 11. was friendly and approachable

Overall, I would rate this student as

- Excellent – Very Good – Good – Poor – Very Poor

Note: Standards for Letter Grades will be determined at the time of the evaluation. Standards vary from faculty to faculty, from year to year, from season to season,... In general, however, students with PASS score below the average of the class get "F".

References

- Uttl, White, & Wong Gonzalez (2016, 2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation.
- Uttl, White, & Morin (2013). The numbers tell it all: Students don't like numbers! PLOS ONE, 8, e83443
- Uttl & Smibert (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. PeerJ
- Uttl, Cnudde, & White (in review). Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: A meta-analysis