Student Evaluation of Instruction/Teaching (SEI/SET) Review

Bob Uttl, Antoine Eche, Olive Fast, Bev Mathison, Halia Valladares Montemayor, Alain Morin, Verna Raab

MRFA
Mount Royal University

2012/01/30

Executive Summary

The mandate of MRFA FEC is to monitor evaluation of faculty, keep abreast of the latest findings and practices elsewhere, inform MRFA membership, and make recommendations to MRFA Executive.

In January 2011, MRFA FEC conducted a survey of faculty members' perceptions of MRU SEI. The results indicate that faculty members are concerned about the validity of SEIs, SEI score interpretations, and student ratings per se.

The FEC has reviewed the literature on SEIs as well as faculty teaching evaluation practices elsewhere with the goal of improving reliability and validity of SEI instrument itself, interpretation of SEI scores, and summative decisions based on SEI scores. The FEC has examined many sources including a recent review of SEI research and practice by Higher Education Quality Council of Ontario (Gravestock & Greenleaf, 2008), a recent meta-analysis investigating relationship between SEIs and student learning (Clayson, 2009), numerous research articles, CAUT recommendations, institutional practices (e.g., University of Victoria, University of Alberta), and some of the most well known SEI systems (e.g., IDEA, ETS SIR II).

Executive Summary

Based on the MRFA FEC SEI Survey results and our review of the literature and teaching evaluation practices, the FEC made a number of recommendations to the MRFA Executive. The key principles behind our recommendations are:

- evaluation of faculty be based on faculty's performance only
- evaluation of teaching performance be comprehensive
- assessment of teaching performance be reliable and valid
- processes, criteria, and standards be transparent and efficient
- processes, criteria, and standards be in writing
- evaluators be trained on processes, criteria, and standards
- evaluation be against written processes, criteria, and standards
- processes, criteria, and standards be equitable and uniformly applied across departments
- processes, criteria, and standards be periodically reviewed

This presentation summarizes what we learned from the literature review and the examination of teaching evaluation practices elsewhere.

Notes on Terminology

Surveys of Student Opinions About Instruction/Courses

- Student Evaluation of Instruction (SEI) (term used at MRU)
- Student Evaluation of Teaching (SET) (term used in literature)

SEI and SET are used interchangeably in this review of the literature and SEI practice.

Outline

- SEI basics
- Summarizing SEIs & outliers
- SEIs and evaluation of teaching
- What do SEI measure?
- Teaching effectiveness irrelevant factors (TEIFs)
- Student and faculty SEI-related behaviors
- Standards for satisfactory SEI performance
- Reliability/Measurement precision
- Courses with new modes of/approaches to instruction
- Use of individual items and written comments
- Written policies
- Procedures, criteria, standards, and interpretive guides
- Formative uses of SEIs
- Research on SEIs: Proceed with caution!
- Summary

SEIs: What Are SEIs Used For?

Formative vs. Summative Uses

Formative uses

- to improve instruction
- non-controversial, widely accepted
- universally considered useful

Summative uses

- to make personnel decisions
  - hiring, tenure, promotion, merit pay, awards
- controversial
  - concerns about reliability, validity
  - concerns about negative impact on education
    - grade inflation, workload reduction
    - ...

"By a wide margin, course evaluations are used for summative, as opposed to formative, purposes..." (Gravestock &Gregor-Greenleaf, 2008)

SEI: What Can Students Evaluate?

Students are not qualified to evaluate some aspects of teaching

- appropriate level of course content
- amount and accuracy of course content
- instructor's knowledge
- ...

Students are able to evaluate other aspects of teaching

- presentation clarity
- organization
- ...

Students are uniquely positioned to evaluate

- class time use, class cancellations
- coverage of scheduled topics
- their own motivation, study habits, preparation, workload
- ...

Coren (2001); Theall & Franklin (2001); Green et al., 1998; Cashin, 1998; d'Appolonia & Abrami (1997)

SEIs: What Are the Common Response Scales?

Likert Response Scales

5-Point Likert

Not applicable
Insufficient Info
Strongly Agree
Agree
Neutral
Disagree
Strongly Disagree

Properties

Not directly interpretable
Meaning of "Neutral" unclear
"Neutral" reduces averages/medians

4-Point Likert

Not applicable
Insufficient Info
Strongly Agree
Agree
Disagree
Strongly Disagree

Properties

Not directly interpretable
Avoids problems with "Neutral"

"Not applicable" and "Insufficient Info" options ought to be included in all scales.

SEIs: What Are the Common Response Scales?

Anchored Response Scales

5-Point Anchored

Not applicable
Insufficient Info
Excellent
Very Good
Good
Satisfactory
Poor

Properties

Directly interpretable
Meaning of ratings clear

7-Point Anchored

Not applicable
Insufficient Info
Outstanding
Excellent
Very Good
Good
Satisfactory
Poor
Very Poor

Properties

Directly interpretable
Meaning of ratings clear
Higher discrimination & reliability

SEIs: What Are the Common Response Scales?

Which response scale(s) is the best?

- The Anchored Scales are preferable
  - tell us directly what students thought about various aspects of courses and instruction
- 7-point Scales may be preferable
  - provide greater discriminability
  - have higher reliability

Summarizing SEIs: What Are the Most Appropriate Measures of Central Tendency for Skewed Distributions?

(University of Alberta's Example)

Instructor's ratings: 11 x 5(SA), 8 x 4(A), 4 x 3(N), 1 x 2(D), 1 x 1(SD)

Mean = 4.08

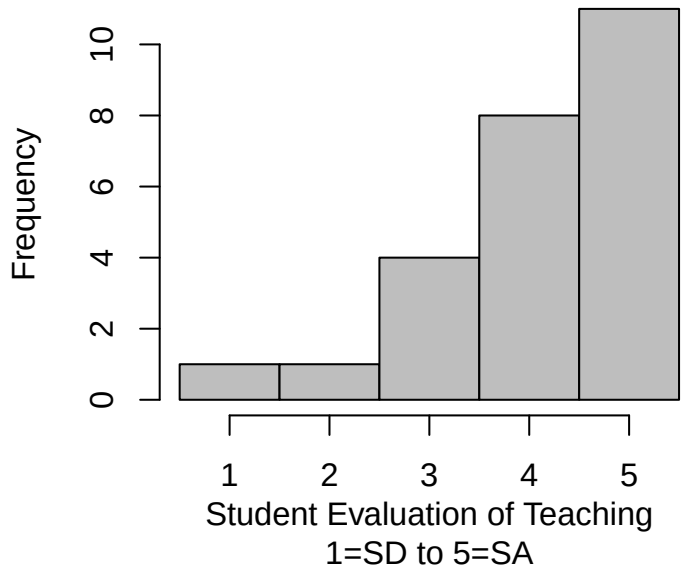
- Inappropriate

Median = 4.00

- Poor resolution w/4 or 5 level categorical data

Interpolated Median = 4.31

- Reflects underlying continuity of opinions



The interpolated medians are the most appropriate. The means are generally poor measures of skewed distribution central tendencies.

The interpolated median: We have 25 scores. The 50% of 25 is 12.5. Six ratings are below "4" and 8 are "4". Thus, the median is within "4 = Agree" or between 3.5 and 4.5. We need to step 6.5 ratings (12.5 - 6 = 6.5) into 8 "Agree" ratings that span 3.5 to 4.5. Accordingly, interpolated median is 3.5 + 6.5/8 = 4.31. (see https://www.aict.ualberta.ca/units/client-services/tsqs/idq/median)

Summarizing SEIs: Should Outliers Be Removed?

Outliers

- scores that deviate from the rest of the score distribution
- mild outliers
- extreme outliers

Outliers

- have substantial effects on means
- may have some effect on medians
- may have small effect on interpolated medians

Identifying Outliers: Tukey's (1977) Analysis

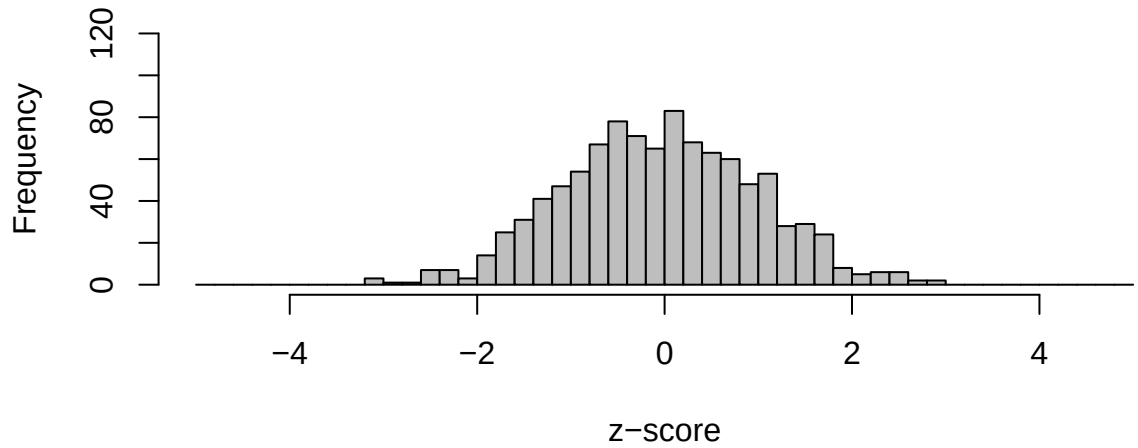
Tukey's Box-and-Whisker Plot is one of the standard ways to identify outliers

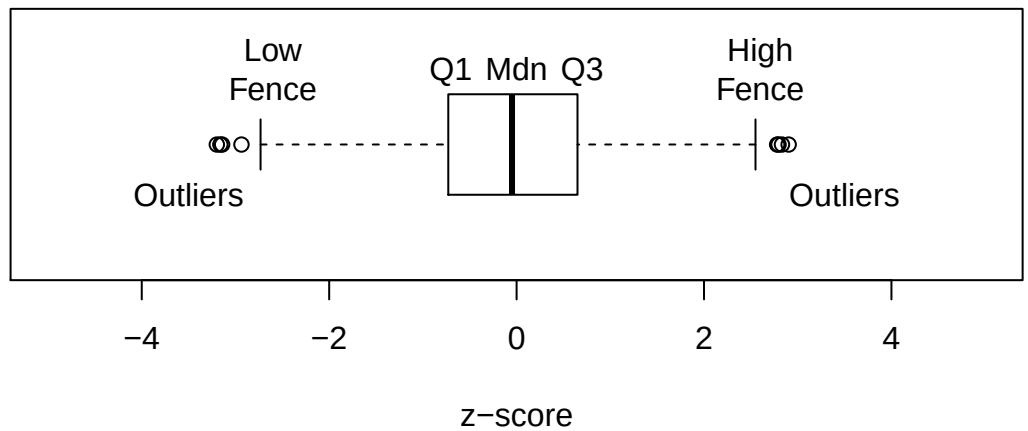
Mean	0
SD	1
Low Fence	-2.70
Q1 (25%)	-0.67
Mdn (50%)	0
Q3 (75%)	0.67
High Fence	2.70
IQR (Q3-Q1)	1.35

Low Fence is 1.5*IQR below Q1

High Fence is 1.5*IQR above Q3

Values below the low fence or above the high fence are considered outliers.



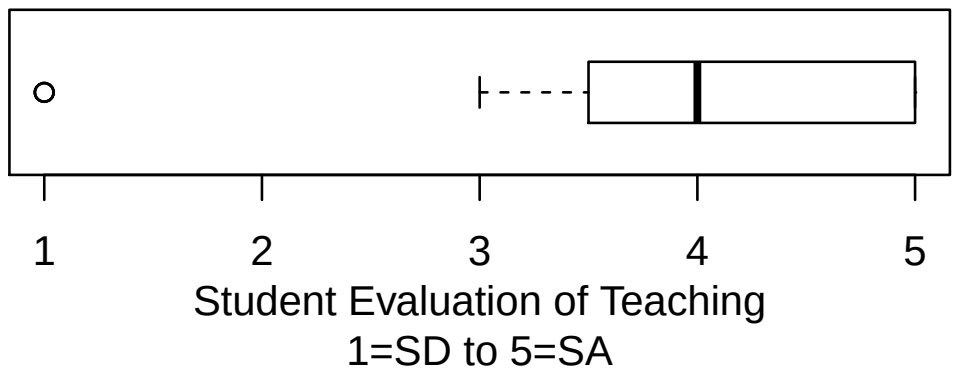


For more information, see Howell (2010), any other statistics textbook, or en.wikipedia.org/wiki/Box_plot.

Influence of Outliers: Example 1 With 3 Outliers

How Much Do Outliers Affect Central Tendency Measures?

Example 1 (n=15)	SD	D	N	A	SA	M	Mdn	Int Mdn
Class w/3 outliers	3	0	1	5	6	3.73	4.00	4.20
Outliers removed	0	0	1	5	6	4.42	4.50	4.50



- Outliers lower means much more than interpolated medians
- Mean and interp. median are nearly identical w/outliers removed

Influence of Outliers: Example 2 With 1 Outlier

How Much Do Outliers Affect Central Tendency Measures?

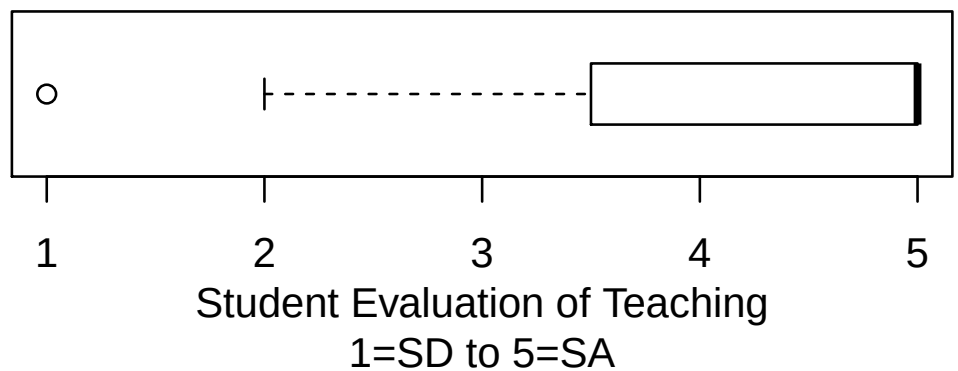
EXAMPLE 2 (n=15)	P	F	G	VG	E	M	Mdn	Int Mdn
Class w/1 outlier	1	1	2	3	8	4.07	5.00	4.56
Outliers removed	0	1	2	3	8	4.29	5.00	4.63

Outliers lower means much more than interpolated medians

Means are dragged down by the tail of skewed distribution

Should faculty focus on teaching to the bottom 10-20% to avoid outliers and tails?

Alternatively, should we remove outliers and use interpolated medians...



SEIs and Evaluation of Teaching

What Weight Should SEI Have In Evaluation of Teaching?

Teaching

- includes many activities (MRFA CA, CAUT, ...)
  - "credit instruction, student consultation, practicum and field supervision, major project supervision, curriculum and course development, pedagogical design and preparation, materials development,..." (MRFA Collective Agreement)
- SEI are relevant to only one teaching activity (credit instruction) and only one aspect of it (student opinions about courses and teaching)

SEIs and Evaluation of Teaching

What Weight Should SEI Have In Evaluation of Teaching?

General Recommendations

- "No evaluation of teaching performance may rely exclusively or primarily upon student questionnaires" (CAUT Model Clause on the Evaluation of Teaching Performance, 2007)
- use multiple sources of data to evaluate teaching
- SEIs should constitute "no more than 30-50% of the final judgment" (IDEA, 2004; assumes SEI actually measure student learning)

What Do SEI Measure? Two Views

Do SEI Measure Instructor Teaching Effectiveness or Student Satisfaction?

Instructor Teaching Effectiveness

- equated with the "amount learned" (Theall, 2001)
- SEI measure characteristics of instructors (not of students)
- key cited evidence for validity are (old) multi-section studies
  - correlation 0.43 b/w SEI and amount learned (Cohen, 1981)

Student Opinions About/Satisfaction With Courses/Instructors

- questions asking student opinions measure student opinions
- SEI measure student satisfaction with courses/instructions
  - different students are satisfied for different reasons
  - SEI measure variables that influence student satisfaction
  - SEI do not necessarily measure characteristics of instructors

Do SEI Measure "Amount Learned"? Old Evidence

Cohen's (1981) Meta-Analysis of Multi-Section Studies

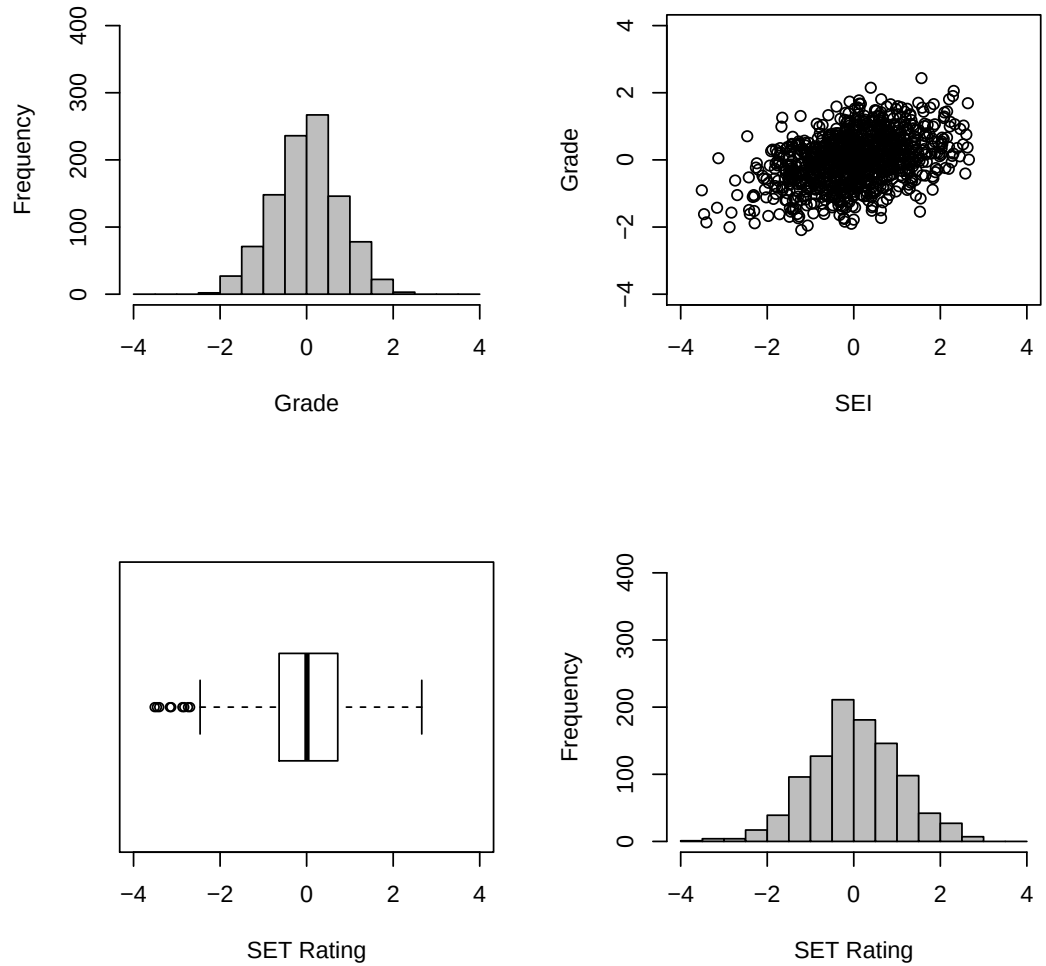
Multi section studies have been used to argue validity of SEIs as measures of learning. Multiple sections of the same course are given the same common exam. If classes taught by instructors with higher SEIs receive higher average grades, we have evidence that students learn more from highly rated instructors.

In one of the first meta-analyses of multi-section studies, Cohen (1981) claimed that SEI and learning are related with r = 0.43 (see the figure for simulated example of r = 0.43) (this study was considered the strongest evidence for validity of SEIs; cited 500+ times).

However, knowing that an instructor received average SEIs (0) tells us little about students' learning/grade in that particular class. This correlation is too weak to be practically useful (Derry, 1979; McCallum, 1984).

Moreover, there were numerous problems with Cohen (1981) meta-analysis.

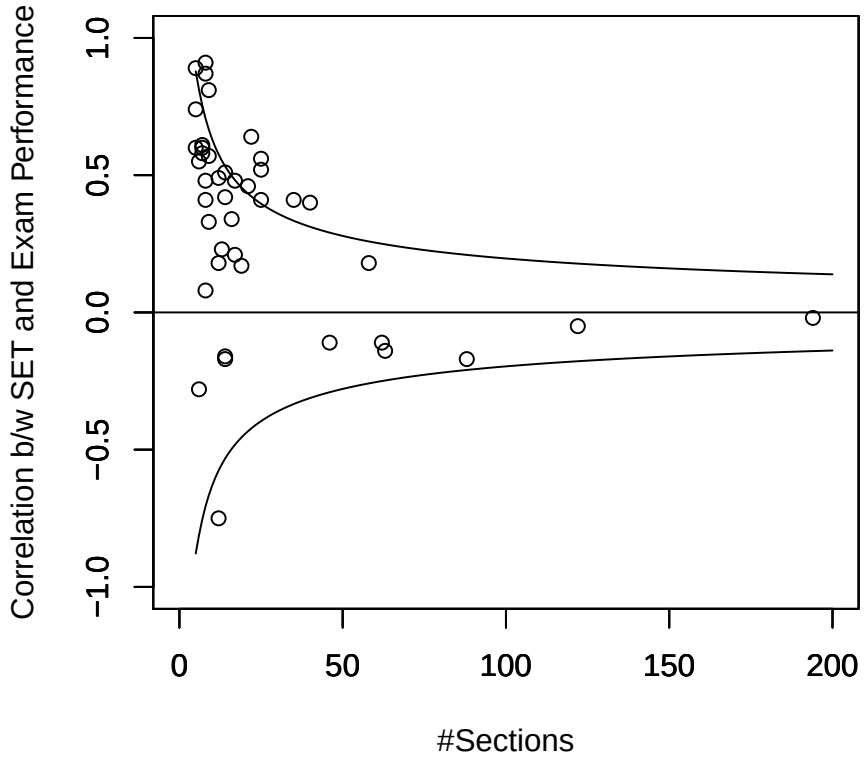
Simulated Example of 0.43 Correlation



Do SEI Measure "Amount Learned"? New Evidence

Clayson (2009) Meta-Analysis of Multi-section Studies

- Conducted meta-analysis of SET and learning relationship
- Found no study after 1990 with positive SET and learning relationship
- Weighted correlations between SET and learning = 0.13
- "Objective measures of learning are unrelated to the SET" (Clayson, 2009)



Uttl & Fruncillo (in preparation; data from Clayson, 2009, Table 1, p. 22): Large sample studies show no significant correlations between SEIs and learning. Students do not learn more from highly rated instructors.

Do SEI Measure "Amount Learned"?

STLHE Conference (University of Windsor, 2008)

Independently, in 2008, the reliability and validity of SEIs was debated at the Society for Teaching and Learning in Higher Education conference, held at University of Windsor. The participants debated the motion: "student course evaluation are a valid and reliable measure of teaching effectiveness for the purposes of summative evaluation." At the end of the debate, "a large majority" voted against the motion (Gravestock, Greenleaf, & Boggs, 2009).

Teaching Effectiveness Irrelevant Factors (TEIFs)

Are SEIs Measuring Factors Other Than Instructor's Teaching Effectiveness?

"administrative, instructor, and course characteristics influence student ratings of instruction" (d'Apollonia & Abrami, 1997)

"Research has confirmed the common belief that instructional outcomes are influenced by 'extraneous variable'..." (Hoyt & Lee, 2003; in "Understanding The IDEA System's Extraneous Variables")

Strength of the Influence: Examples

- "Liberal grading practices increased student ratings, at most, by slightly less than 0.5 on a 5-point scale." (d'Apollonia & Abrami, 1997)
- "Measures of five extraneous circumstances accounted for about 15-20% of the variation in ratings of course outcomes [in IDEA System]. The supposition that student learning is impacted by circumstances beyond the instructor's control was confirmed." (Hoyt % Lee, 2003)
- Quantitative vs. non-quantitative courses are rated lower by 0.2 to 0.5 on a 5-point scale (IDEA, SIR II, MRU data, ...)

Teaching Effectiveness Irrelevant Factors

What Factors Influence SEIs?

Student Variables

- student prior motivation/interest
- student preparation
- student work habits
- attendance
- ...

Course Variables

- discipline
- required vs. elective
- class size
- course level
- ...

Instructor Variables

- personality/popularity
- expressiveness
- ...

Instructor Controlled Variables

- grades
- workload
- ...

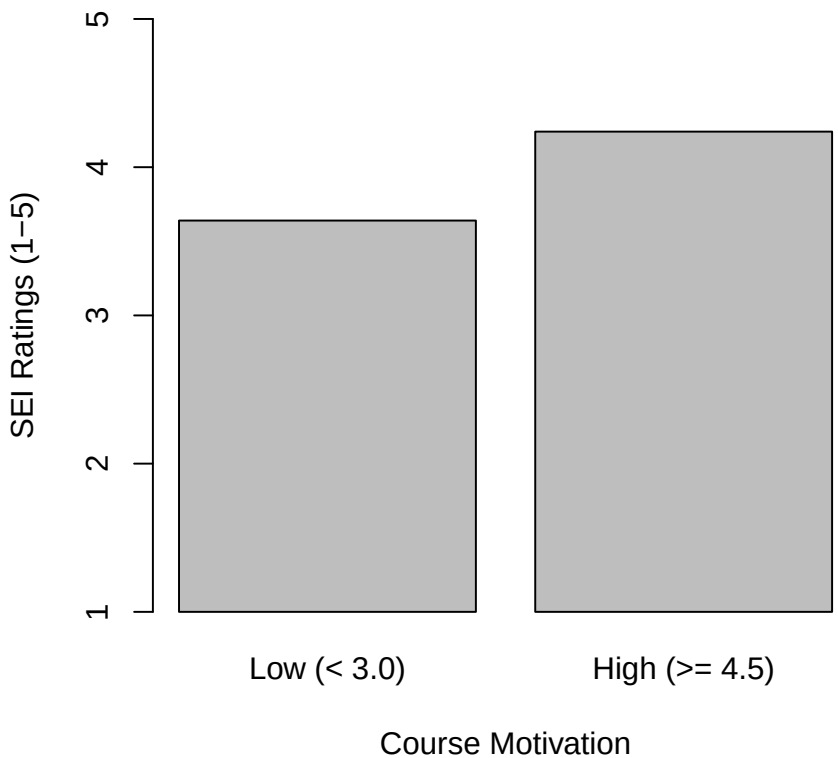
Administrative Variables

- timing of SEIs
- anonymity
- presence of instructor
- ...

Student Variables: Student Motivation

Is student motivation to take a course related to SEIs?

IDEA System Motivation Question: "I really wanted to take this course regardless of who taught it" (1=Definitely False, 2=More false than true, 3=In between, 4=More true than false, 5=Definitely true)



CM correlation with "excellence of course" was .47.

"In summary, instructors in classes with highly motivated students had a considerable advantage over those teaching classes with poorly motivated students." (Hoyt & Lee, 2003)

Student Variables: Student Motivation & Class Size

IDEA System Example (Hoyt & Lee, 2002)

Item: Explained course material clearly and concisely.

	Class Size			
Motivation	Small	Medium	Large	Very Large
Low	3.93	3.89	3.84	3.80
Low Average	4.07	4.05	3.99	3.97
Average	4.16	4.16	4.13	4.10
High Average	4.29	4.23	4.25	4.15
High	4.37	4.33	4.29	4.30

- Students with high vs. low motivation give much higher ratings (0.44 on 5-point scale, for small classes).
- Students in large classes tend to give somewhat lower ratings.

Class size: Small (10-14), Medium (15-34), Large (35-49), Very large (50 or more); Motivation: High (upper 10%), High Average (next 20%), Average (middle 40%), Low Average (next 20%), Low (lowest 10%).

Student Variables: Student Motivation & Class Size

IDEA System Example (Hoyt & Lee, 2002)

Item: Inspired students to set and achieve goals that really challenged them.

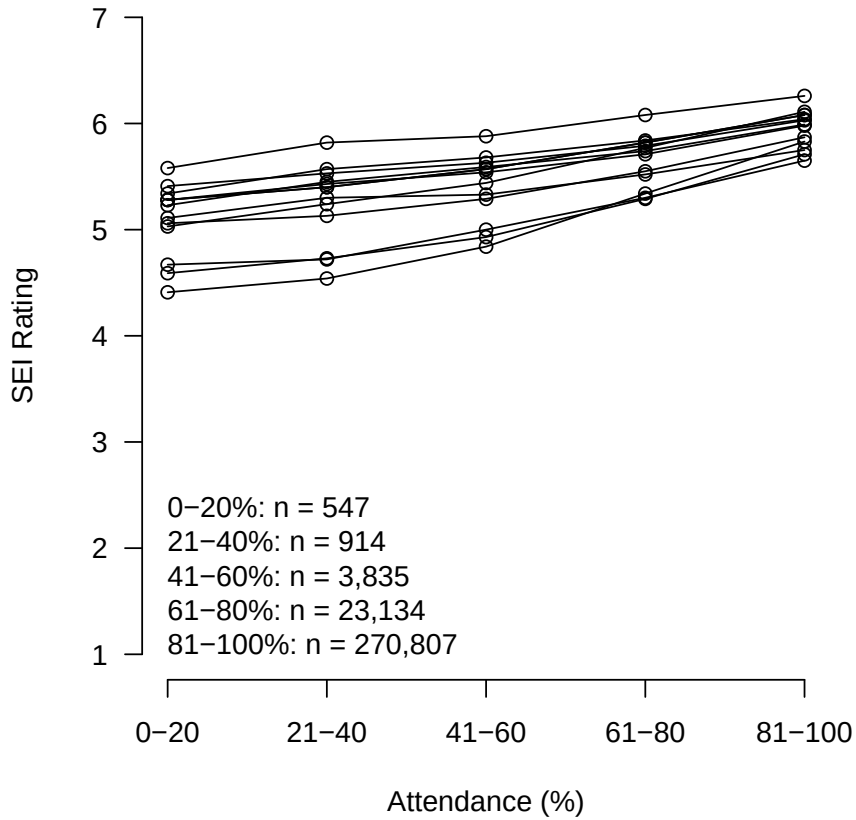
	Class Size			
Motivation	Small	Medium	Large	Very Large
Low	3.70	3.52	3.28	3.16
Low Average	3.83	3.66	3.47	3.33
Average	3.92	3.82	3.64	3.52
High Average	4.06	3.95	3.86	3.75
High	4.21	4.14	4.03	4.07

- Students with high vs. low motivation give much higher ratings (0.51 on 5-point scale, for small classes).
- Students in large classes give much lower ratings.
- Effect of class size depends on student motivation.

Class size: Small (10-14), Medium (15-34), Large (35-49), Very large (50 or more); Motivation: High (upper 10%), High Average (next 20%), Average (middle 40%), Low Average (next 20%), Low (lowest 10%).

Student Variables: Student Attendance

Is student attendance related to SEIs?



Students who attend classes give higher SEI ratings (raw difference 0.86)

UC USRI Review Committee (2003)

Course Variables: Discipline

Are Some Disciplines Rated Higher?

"Professors in fine arts, humanities, and health-related professions are more highly rated than their science, engineering and math-related colleagues" (Franklin & Theall, 1995)

University of Alberta Example (1=SD to 5=SA)

Department	Low Fence	25%	Median	75%
Physics	2.4	3.7	4.1	4.5
Math & Stats	2.8	3.9	4.2	4.6
English	2.8	4.0	4.4	4.7
Elementary Education	2.7	4.0	4.5	4.8
Drama	2.9	4.1	4.7	4.9

There are substantial disciplinary differences in median SEI ratings.

USRI Reference Data (University of Alberta)

Course Variables: Discipline

MRU 2008/2009 Data

Department	Mean SEI
Math, Physics, & Engineering	4.19
English	4.31
Humanities	4.44
Phys. E. & Recreation	4.54
Music Performance	4.55

There are substantial disciplinary differences in SEI ratings in MRU.

MRU Office of Institutional Analysis and Planning

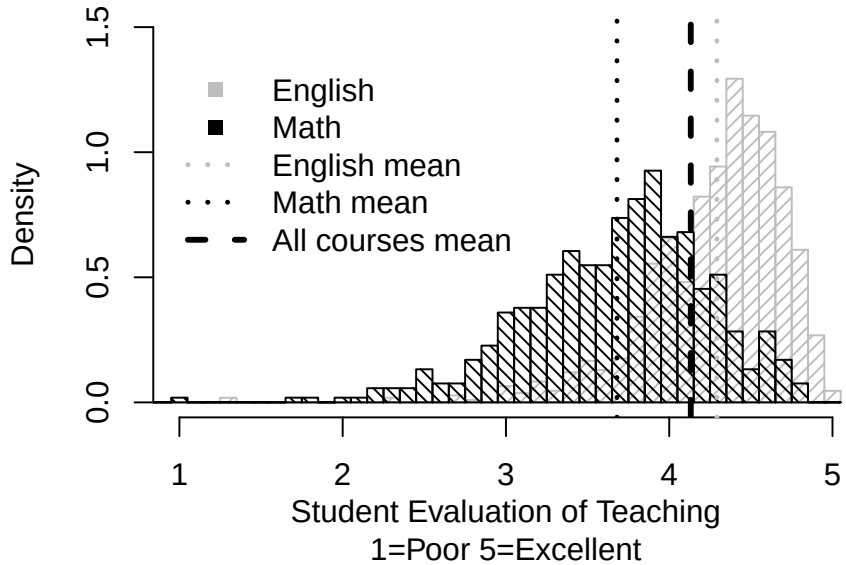
Course Variables: Discipline

How Large Are Disciplinary Effects? (Uttl & Smibert, in preparation)

Percentage of Professors in Math & Stats vs. English Passing Various Standards for "Satisfactory" Performance

Standard	**Math**	**English**
Mean	21.4	71.3
Mean - 1 SD	58.0	93.1
Excellent (4.50)	5.8	35.4
Very Good (3.50)	66.0	94.9
Good (2.50)	96.6	99.6
Fair (1.50)	99.8	99.8

Disparity between percentages of professors passing various standards depends on the standard level. The disparities are the highest at around the mean of the distributions (Mean as standard) and relatively low at "Good" standard and disappear at "Fair"standard.

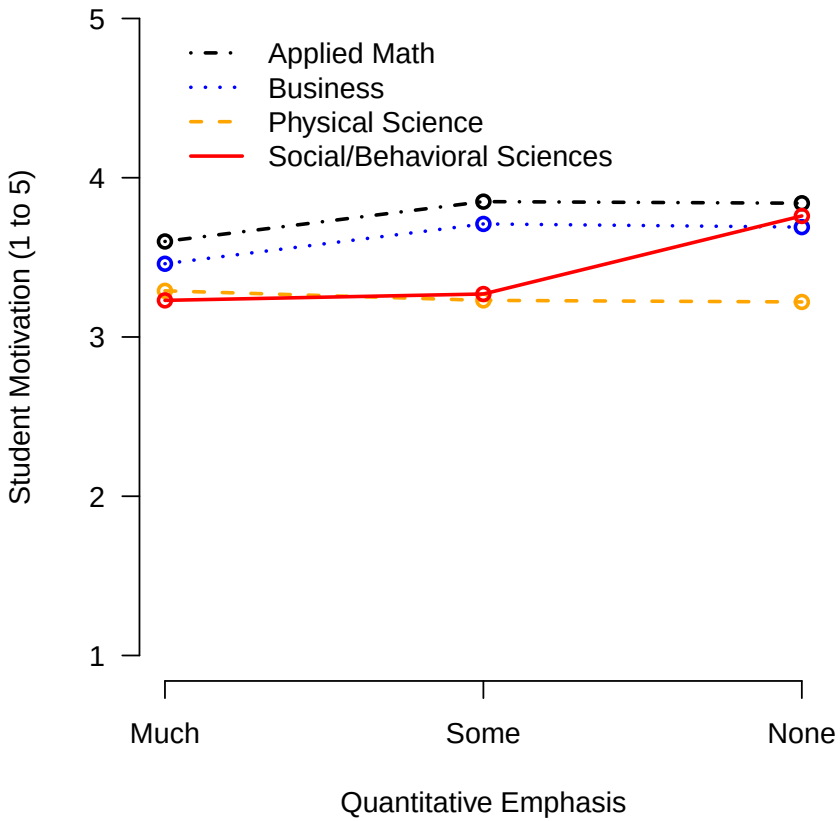


Distribution of SEI ratings for Math & Stats profs is lower than the distribution of English profs. Using the same cut-offs for "Satisfactory" performance results in much higher percentages of Math & Stats profs failing the standard. (from Uttl & Smibert, in preparation)

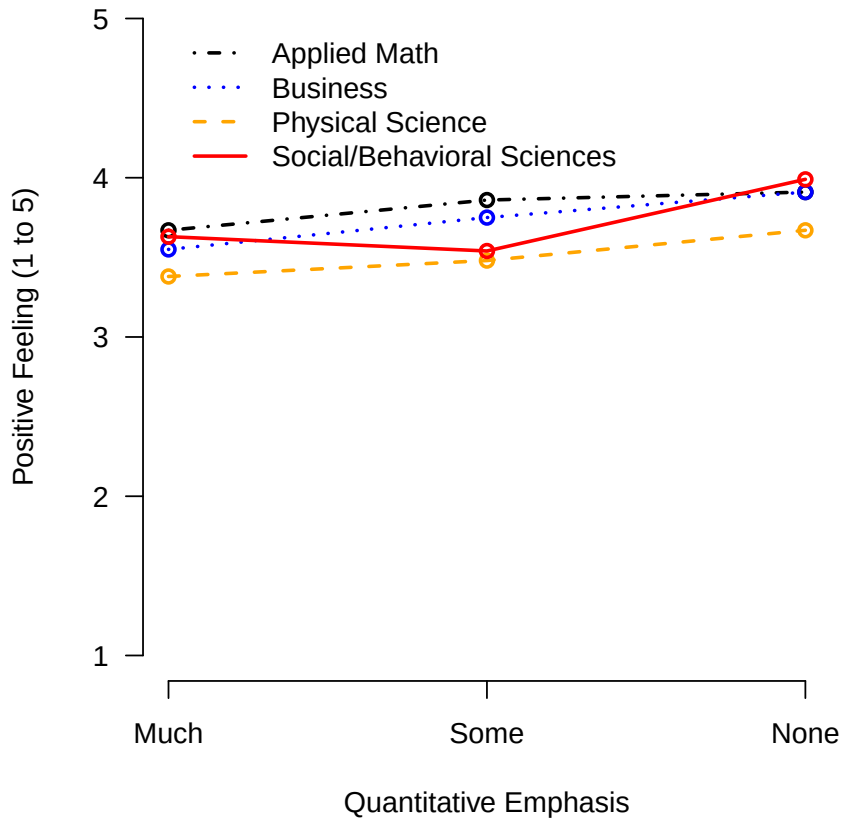
Course Variables: Disciplines Within Disciplines

Are there effects of discipline within disciplines? (Hoyt & Perera, 2001)

Business, social/behavioral sciences, physical sciences and other disciplines include courses that have a quantitative emphasis... What are the disciplinary effects within these disciplines?



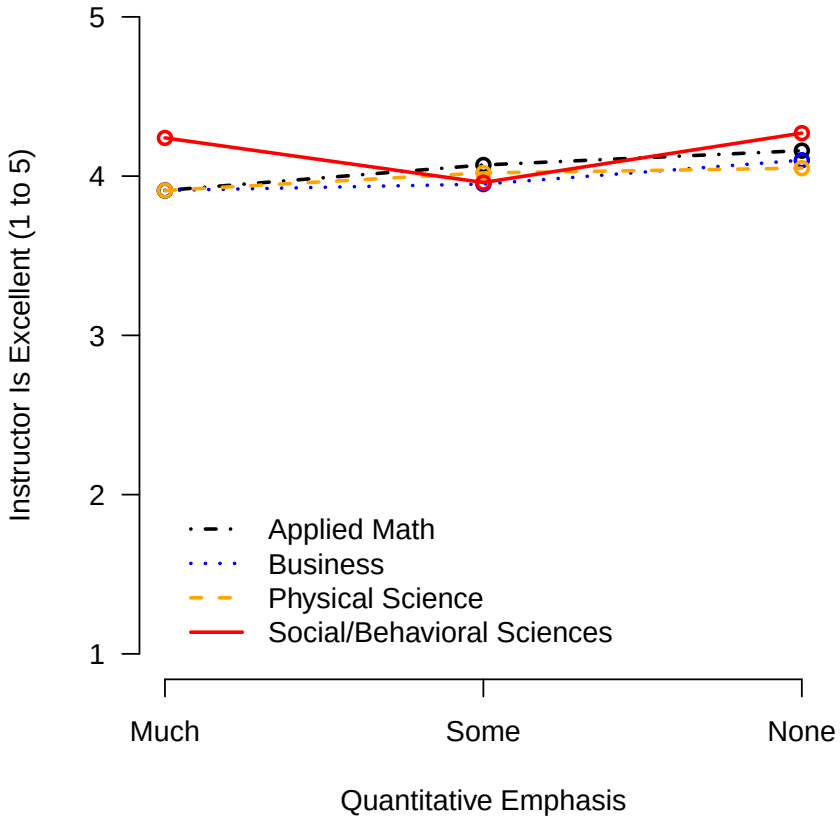
Students are less likely to have "strong desire" to take quantitative courses, although patterns differ across the disciplines.



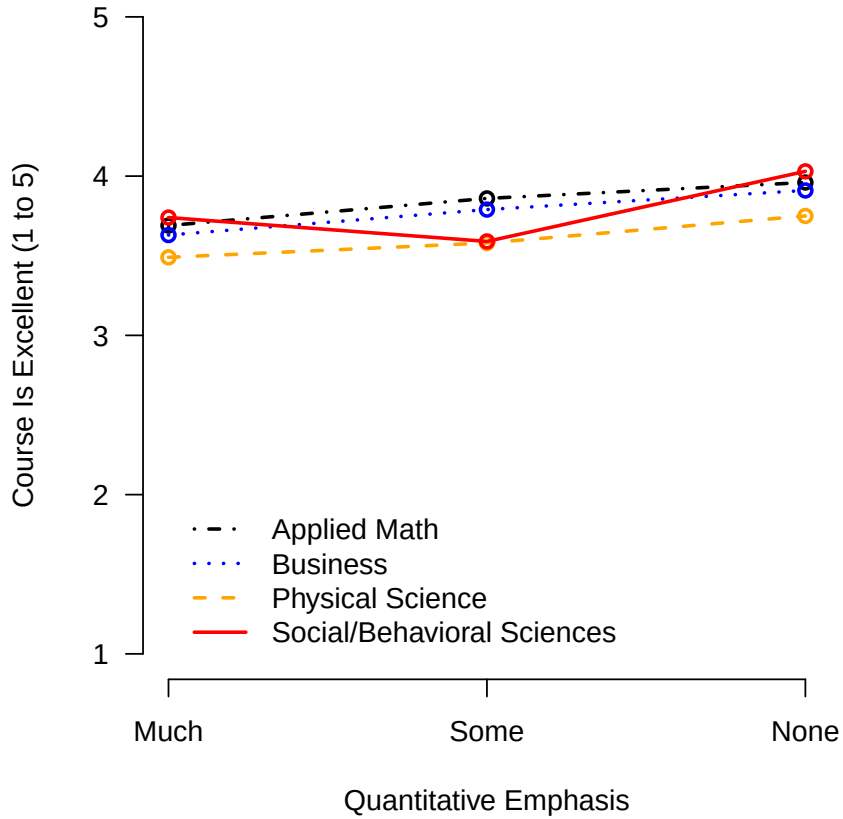
Students are less likely to have "more positive feelings toward this field of study" after taking quantitative courses.

Course Variables: Disciplines Within Disciplines

Are there effects of discipline within disciplines? (Hoyt & Perera, 2001)



Students tend to rate instructors lower in quantitative courses.



Students tend to rate quantitative courses even lower than instructors in quantitative courses.

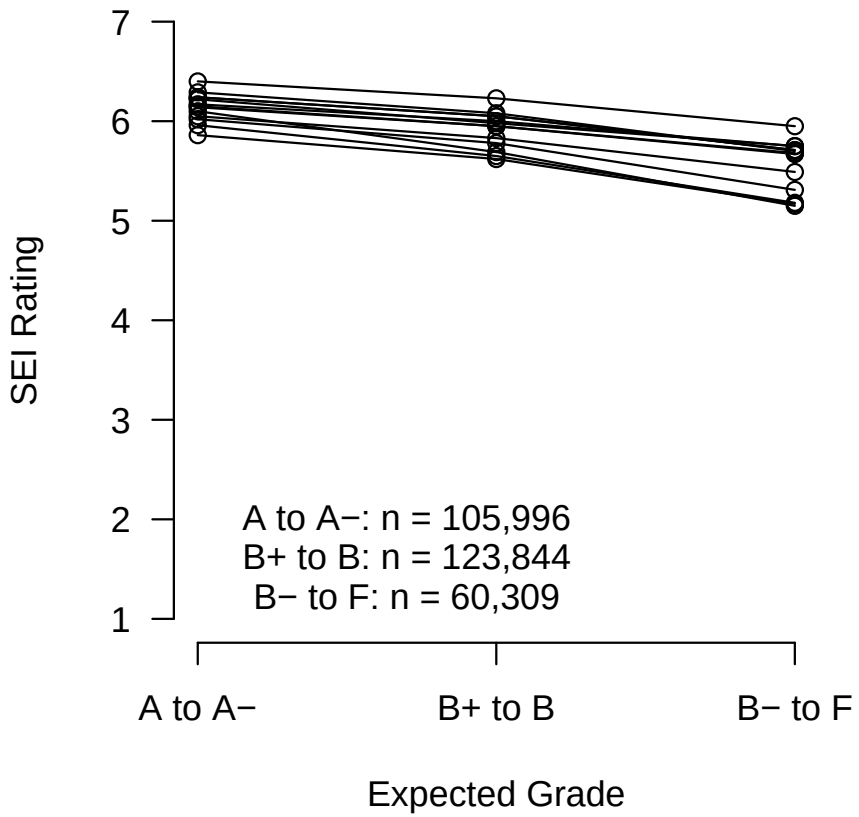
Course Variables: Disciplines Within Disciplines

Are there effects of discipline within disciplines? (Hoyt & Perera, 2001)

"Overall, those teaching classes emphasizing math/quantitative skills are more likely than their colleagues in similar disciplines to have poorly motivated students who regard their courses as difficult and demanding and who offer relatively negative evaluations of the course, its instructor, and course outcomes. However, these conclusions vary with discipline." (Hoyt & Perera, 2001)

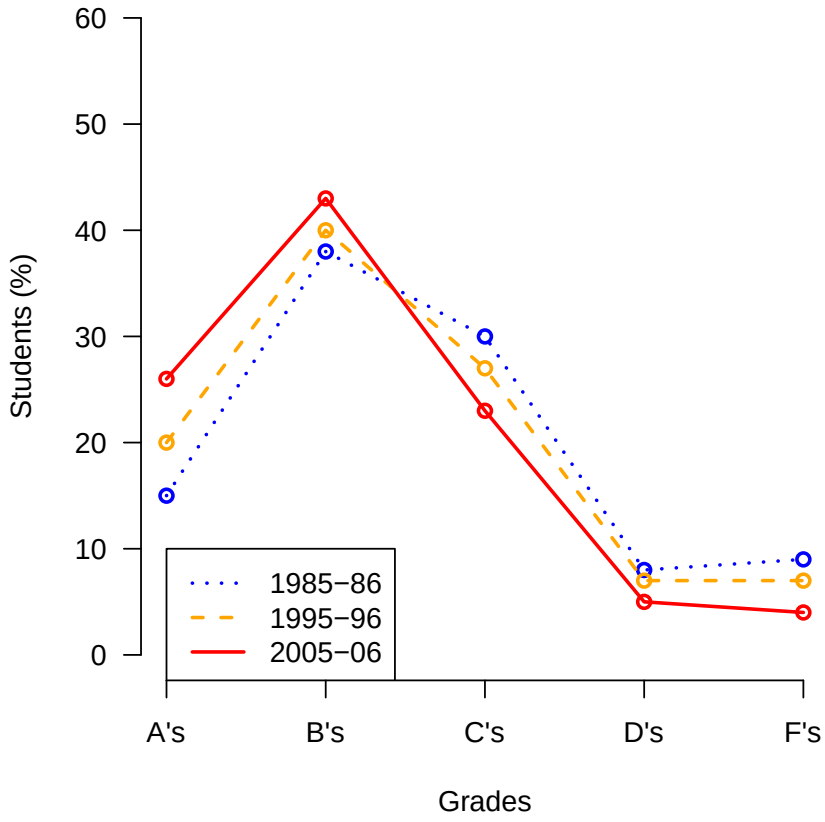
Instructor Controlled Variables: Expected Grades

Are expected grades related to SEIs?



Students who expect higher grades give higher SEI ratings (raw difference = 0.59)

UC USRI Review Committee (2003)

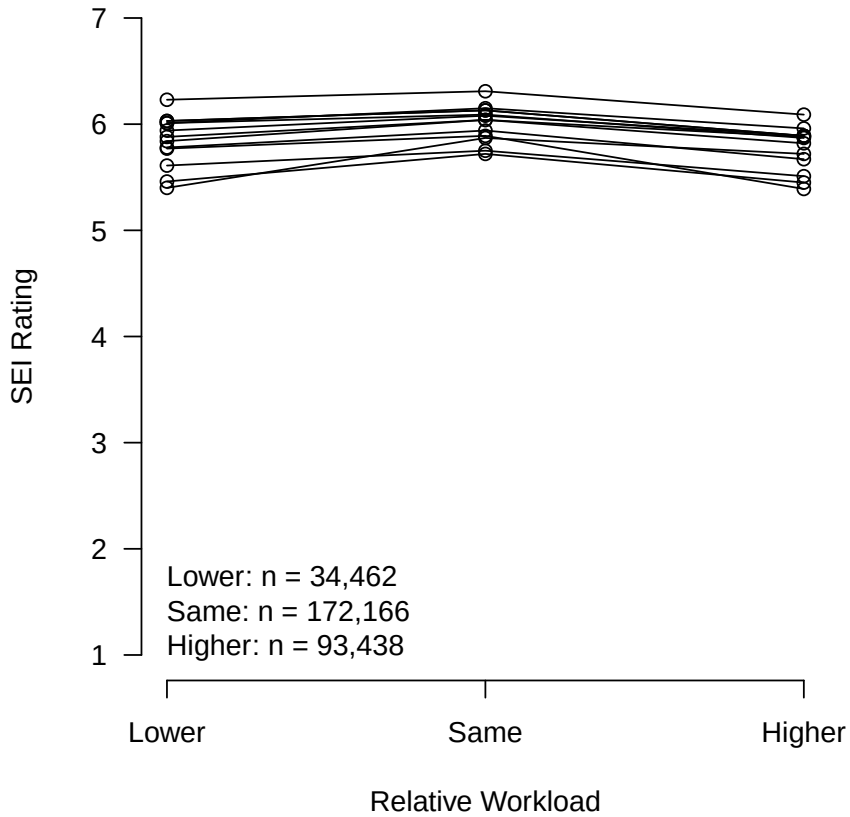


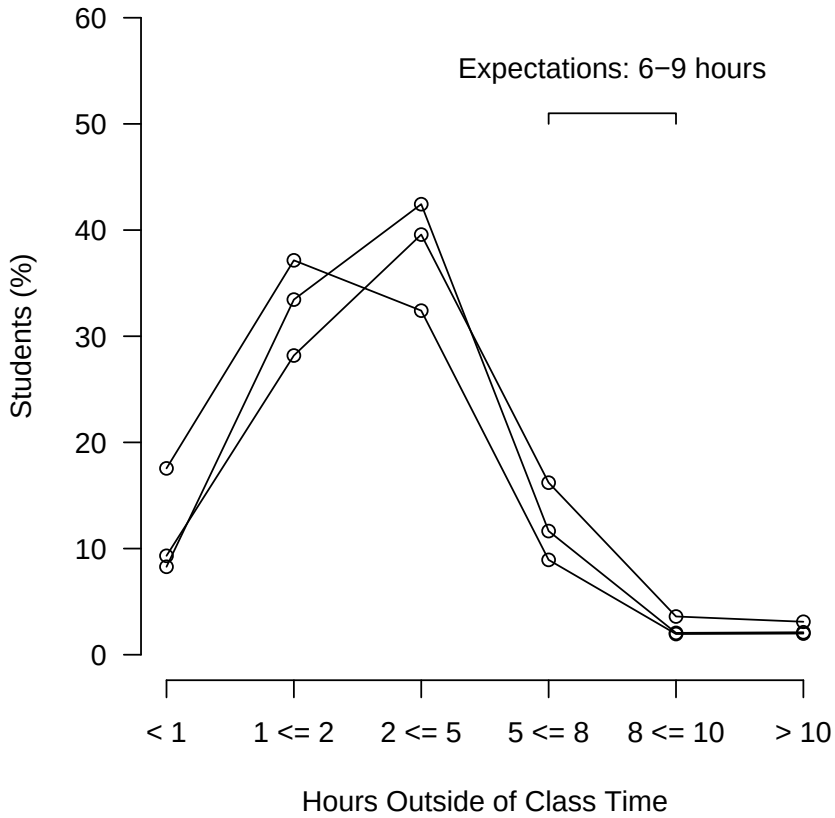
Grade distribution has shifted towards higher grades over the years

UC OIA (2006)

Instructor Controlled Variables: Workload

Is student workload related to SEIs?





Students who perceive workload as higher or lower give lower SEI ratings (raw difference 0.24 and 0.17, respectively)

UC USRI Review Committee (2003)

Students spent about 3 rather than 6-9 hours per 3 credit class (outside of class time)

UVic data for Enginering, Education, and Social Science (see also NSSE data)

Instructor Controlled Variables: One Interpretation

SEIs Caused Grade Inflation and Workload Reduction

- Instructors feel pressure to increase SEIs
- Instructors give higher grades
- Instructors reduce workload, become less demanding

(e.g., Emery et al., 2003; Clayson, 2009; Crumbley & Reichelt, 2009; Zabaleta, 2007)

Teaching Effectiveness Irrelevant Factors (TEIFs)

What should be done about the TEIFs?

"when the focus is on the *evaluation of teaching effectiveness*, it is important to separate the contributions of the *teacher* from the contribution of *extraneous factors* to student learning. That is the purpose of the IDEA system's **adjusted ratings**." (Hoyt & Lee, 2003; in "Understanding The IDEA System's Extraneous Variables")

Mean/median adjustments

- adjust scores for various TEIFs such as student motivation, work habits, class size

Comparison groups (same or similar courses)

- compare scores only to the same or similar courses

Teaching Effectiveness Irrelevant Factors (TEIFs)

Adjusting for TEIFS: Examples

IDEA System (www.theideacenter.org)

- adjusts for student motivation, work habits, class size
- uses relevant comparison group

SIR 2 (www.ets.org)

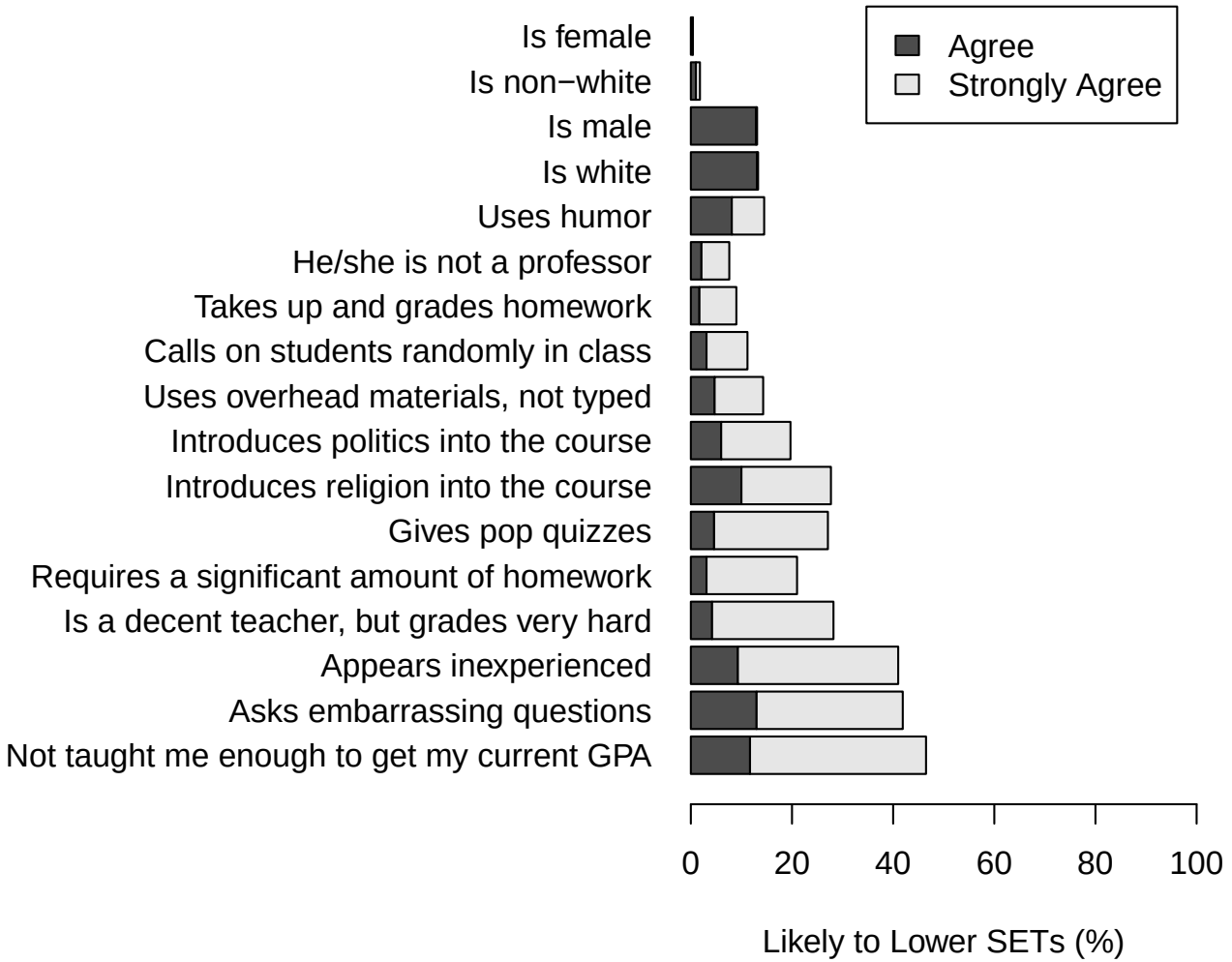
- uses relevant comparison group

University of Alberta
(www.aict.ualberta.ca/units/client-services/tsqs/usri)

- uses relevant comparison group down to the same course

Students Lower SEIs for Teaching Irrelevant Reasons

Crumbley et al. (2001): Survey of 530 accounting students



Substantial percentages of students state that they are likely to lower SETs for various behaviours and professor characteristics irrelevant to teaching effectiveness

Students Use SEIs as Reward and Revenge Tool

Lin, 2008: Survey of 693 students

Q: "Honestly did you ever intentionally use student evaluations of teaching to reward or revenge on professors?"
YES: 40.6% of 693 responding students

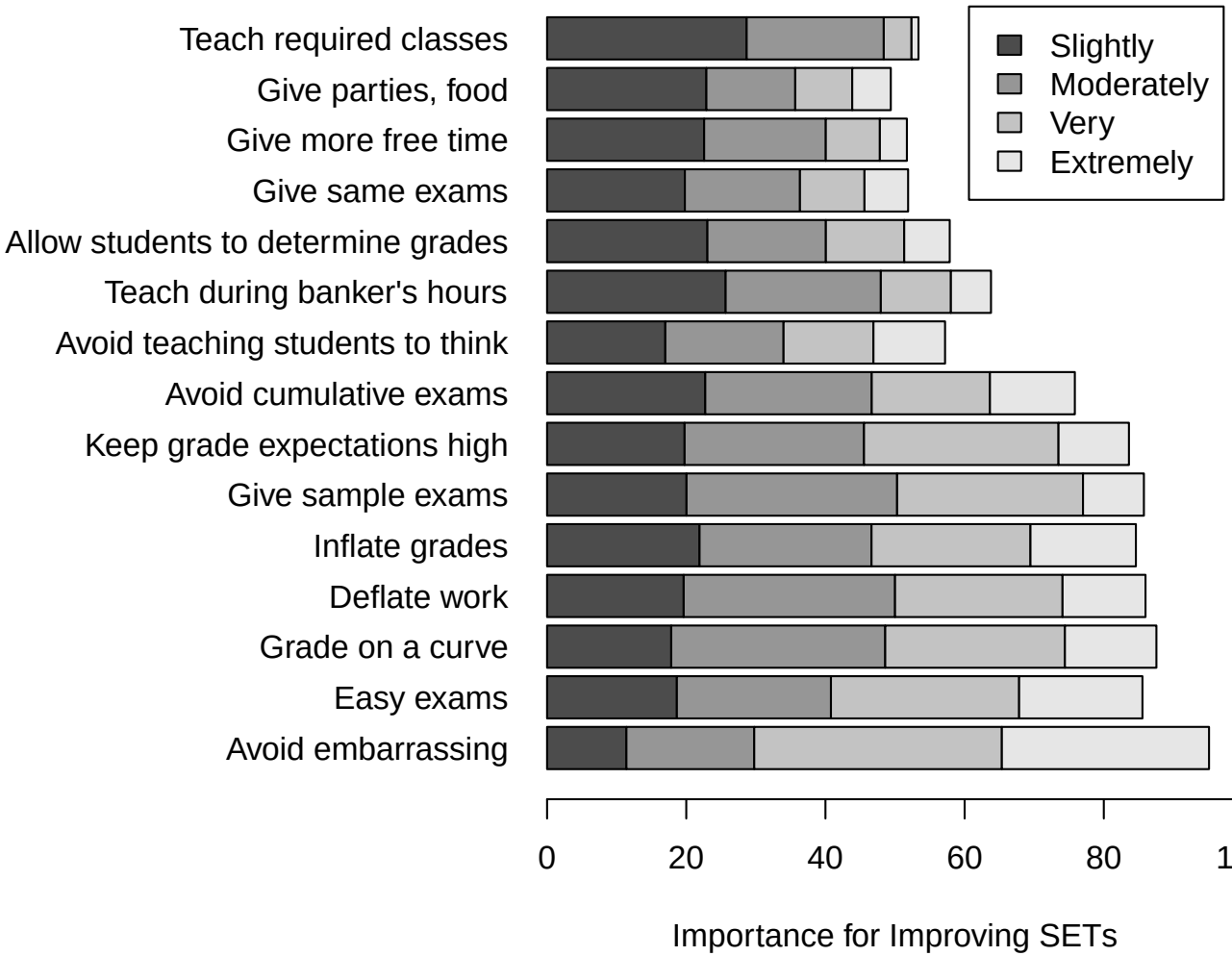
Students themselves admit that they use SETs for other purposes than to evaluate teaching.

Consider: Are students likely to lower a professor's SET rating if he/she:

- denies them an extension on a project
- reports them for academic dishonesty violation
- ...

Faculty Engage In Dysfunctional Behaviors to Improve SEIs

Crumbley & Reichelt (2009): Survey of 447 accounting professors (44.7% response)



Large percentages of professors believe that many dysfunctional behaviours are important for improving SETs...

Are SEIs Valid?

Definition of test validity

- The validity of a test is the extent to which it measures what it is supposed to measure.
- "A test is valid to the extent that inferences made from it are appropriate, meaningful, and useful" (AERA, APA, & NCME, 1985; Standards for Educational and Psychological Testing)

Are SEIs Valid?

Definition of effective teaching

- Effective teaching is teaching that results in learning (e.g., Marsh & Roche, 1997)
- "[no] universal set of characteristics of effective teachers and courses that should be used as a target... appear to exist" (Ory and Ryan, 2001, p. 32)
- "faculty members and administrators have stereotypes about what good teaching involves... teachers who do not conform to the stereotype are likely to be judged ineffective despite other evidence of effectiveness" (McKetchie, 1997)

Are SEIs Valid?

Are SEIs valid measures of student learning/teaching effectiveness AND/OR student opinions about teaching?

SEIs are not valid measure of student learning/teaching effectiveness

- Do not correlate with learning
- Influenced by wide variety of TEIFs
- Students admit using SEI for other purposes
- Faculty admit engaging in variety of impression management techniques including grade inflation, workload reduction, etc.

SEIs are (somewhat) valid measure of student opinions about teaching

- SEIs ask students to provide their opinions
- SEIs sometimes do not reflect student opinions about courses/instruction

Standard for Satisfactory SEI Performance

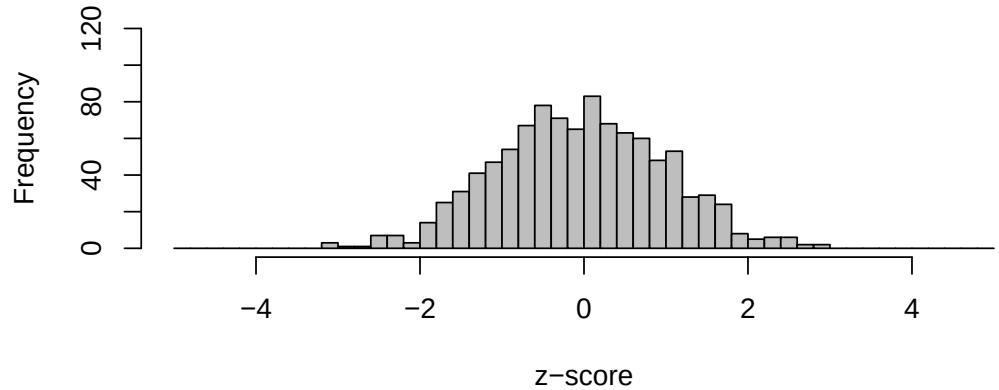
How can we set the standard for satisfactory performance?

- Case by Case "Standard"
  - Standard is determined by evaluators at the time of evaluation
  - e.g., the standard is 4.5 for this person but 3.0 for that person
- Norm Referenced Standard
  - Standard depends on a normative group's performance
  - e.g., only scores at or above the group mean are satisfactory
- Criterion Referenced Standard
  - Standard is a priori meaningful cut-off
  - e.g., scores above 3.0 on a 1 to 5 scale are satisfactory
- Distribution Referenced Standard
  - Standard is a fit within a group performance distribution
  - e.g., scores that are not low outliers are satisfactory

Example 1: University of Victoria Model

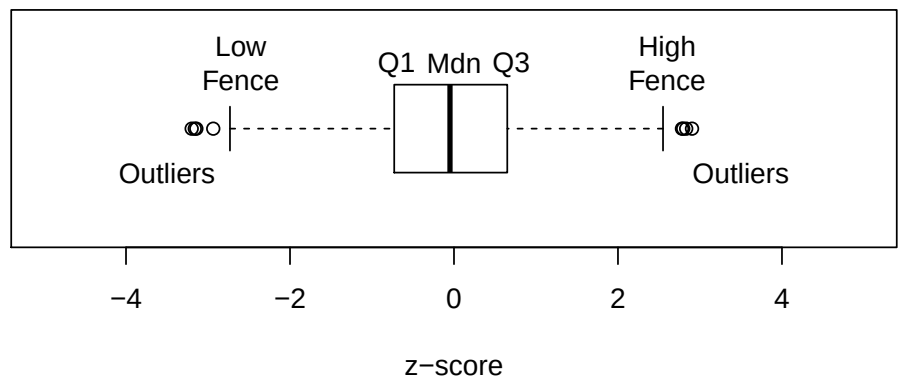
Norm-referenced standard set at M - 2 SDs (assumes normal distribution)

Mean	0
SD	1.00
Low Fence	-2.70
Q1 (25%)	-0.67
Mdn (50%)	0
Q3 (75%)	0.67
High Fence	2.70
IQR (Q3-Q1)	1.35



Low Fence is 1.5*IQR below Q1
High Fence is 1.5*IQR above Q3

Imagine: If this were the distribution of SEI scores in MRU, where should we place the cut off for "unsatisfactory performance"? What percentage of instructors should we deem "unsatisfactory" each year? 50%? 16%? 2%?

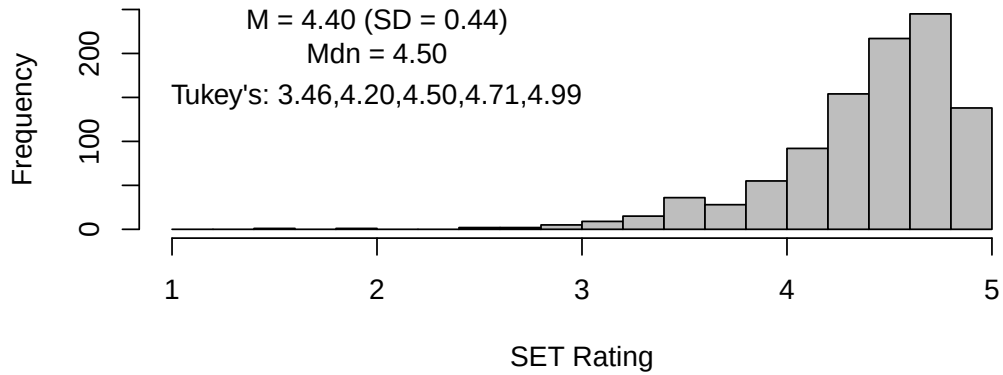


"It is likely that the majority of instructors will fall within 2 standard deviations of the means for their department or faculty. Those at the extremes might possibly be considered, with additional information, as candidates for more intensive mentoring or for teaching awards." (Using the New Course Experience Survey... at UVic; Dawson & Wall, 2009, p. 5)

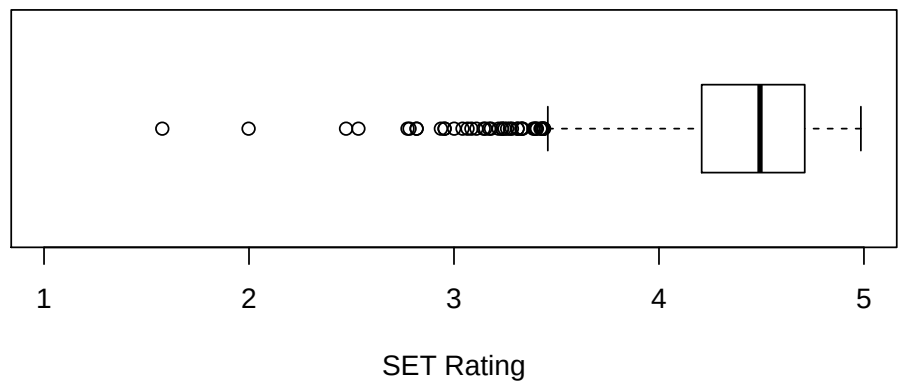
Example 2: University of Alberta Model

Distribution-referenced standard (acknowledges non-normal, skewed SEIs distributions)

Percent	Cut off
1%	2.95
2%	3.22
3%	3.33
4%	3.44
6%	3.59
8%	3.71
10%	3.81
12%	3.88
14%	3.95
16%	4.00



Typical SET distribution is skewed. In skewed distributions, the median better describes the central tendency of scores than the mean. Tukey's low fence (3.46) identifies SET scores that may be considered outliers. Table shows that in this distribution about 4% of scores are low outliers. In contrast, setting cut off to 4.00 would classify as "unsatisfactory" 16% of faculty each year/evaluation period.



"Since 25% of the classes obtain medians above the 75th percentile and 25% obtain medians below the 25th percentile **by definition**, these values should not be used to determine whether a particular score is *good* or *bad*... This value [low fence] identifies a point below which scores (medians) may be considered outliers, i.e., scores which appear to be outside the usual distribution of scores for the reference group being tabulated."(IDQ Reports, University of Alberta)

Norm Referenced Standards

Key Features

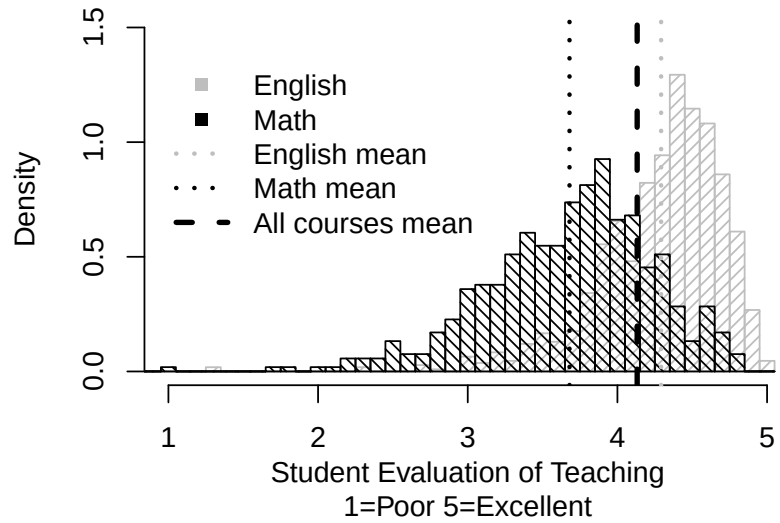
- performance of others determines the standard
- cut-offs typically determined from means and SDs

Pros

- ??

Cons

- SEI have skewed distributions (Ms & SDs invalid)
- focuses on competition among faculty
- demoralizing because, by definition, some faculty will always be unsatisfactory



Satisfactory Standards Examples

Standard	Percentage of Unsatisfactory
Mean	50%
Mean minus 1 SD	16%
Mean minus 2 SD	2%

By definition, large percentages of faculty fail these standards even if they are all excellent. If applied consistently, substantial proportion of faculty would be declared "unsatisfactory" at the end of each review period.

Criterion Referenced Standards

Key Features

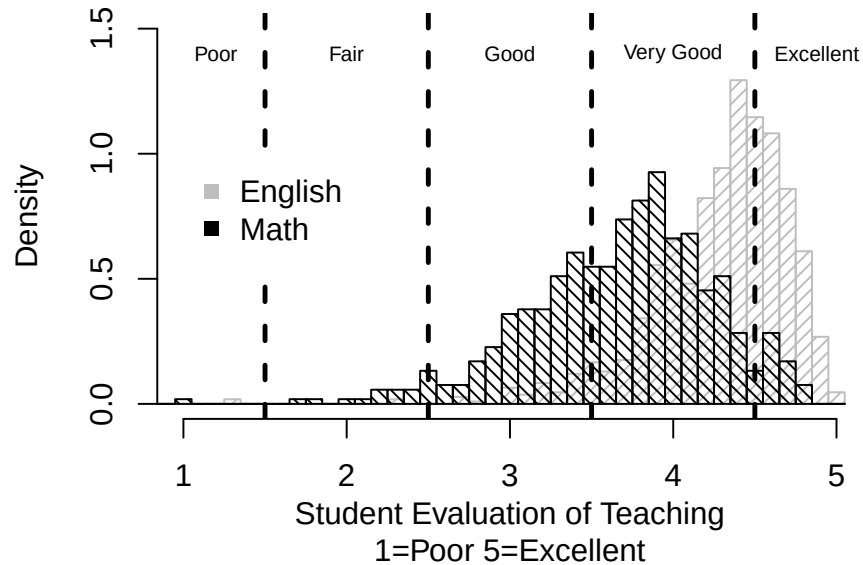
- performance of others does not influence the standard
- cut-offs determined as a priori meaningful performance levels

Pros

- focuses on competency
- no one has to fail
- does not hinder cooperation

Consider

- Some may be only "Fair/Satisfactory"



Standard Example

Below 1.4	Poor
1.5-2.4	Fair/Satisfactory
2.5-3.4	Good
3.5-4.4	Very Good
4.5 or higher	Excellent

A few faculty receive "Poor" from students. Nearly all faculty are considered "Good", "Very good", and "Excellent".

See www.cnu.edu/facresc/pdf/IDEAInterpretiveGuideShortForm.pdf for another example

Distribution Referenced Standard

Key Features

- Performance of a group determines the standard
- Fit within the group distribution indicates satisfactory performance
- Low outliers do not fit/indicate unsatisfactory performance

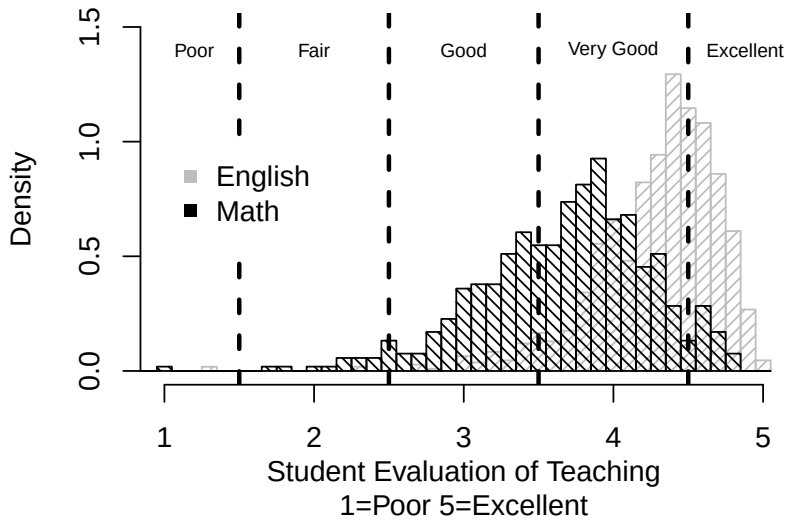
Pros

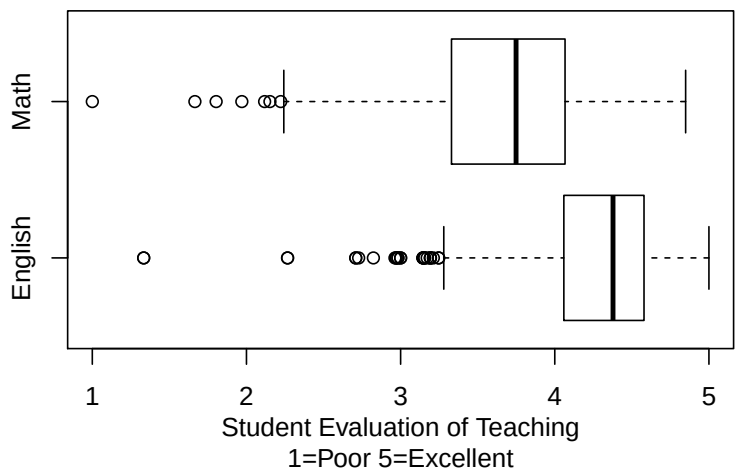
- No one has to fail (outliers are not statistical necessity)

Cons

- ??

See http://www.aict.ualberta.ca/units/client-services/tsqs/usri/usri-reference-data





Reliability

The reliability of a test is the extent to which the test scores are consistent across testing occasions.

The reliability of SEI is determined primarily by the extent to which individual students give identical ratings to the same professor. Students are assessing (presumably) the same measurable quantity (e.g., instructor's teaching effectiveness) and should all give the same rating, if the SEIs were perfectly reliable.

In reality, different raters rate differently; different raters register in courses; different raters are present on different days; different raters had different amounts of sleep; different raters were dealing with different kinds of problems; different raters arrived to different ratings of whatever they were rating...

Reliability: SEI Score IS NOT Instructor's True Score

Observed scores and reliability only allows us to set some limits on likely true scores

Observed SEI Scores vs. True SEI Scores

- SEIs are not perfectly reliable
- Ratings include random error
  - Observed = True + Random Error
- Ratings may include systematic error
  - Observed = True + Random Error + Systematic Error

Reliability: Validity of Score Interpretation

When can we conclude that someone failed the standard?

Validity of Score Interpretations

- One CANNOT conclude that an instructor failed the standard if his or her observed SEI score is numerically below the standard
- One CAN only conclude (tentatively, with specified degree of confidence) that an instructor failed the standard IF AND ONLY IF (for example) 95% Confidence Interval surrounding his or her observed SEI score is COMPLETELY below the standard

What Are the Consequences of Ignoring Test Reliability? Ms. Leilani Muir Example

Background

- born July 15, 1944, Calgary
- unwanted child, average in schools
- at age of 9, her mother applied to have her placed in Provincial Training School for Mental Defectives (a.k.a., Michener Center) in Red Deer, Alberta
- required to take IQ test and undergo compulsory sterilization (if mentally deficient, if IQ less 70)

Ms. Leilani Muir: Two Assessments, Two Scores

One justified sterilization (by the laws of the time), the other did not

First IQ Assessment

- scored 64 on IQ test, labeled "Mental Defective Moron"
- sterilized in 1959 (age 15) (believing she was undergoing appendectomy)

Second IQ Assessment

- in 1989, depressed
- given IQ test, scored 89 (normal; IQ mean $= 100$, SD $= 15$)

Lawsuit

- sued Alberta for wrongful sterilization
- on January 25, 1996, awarded $740,780 in damages plus $230,000 for legal costs

Source: Muir v. Alberta, 1996 CanLII 7287 (AB QB)

Ms. Leilani Muir: Neither of the Two Assessments Justified Sterilization

First Assessment

- standard error of measurement (SEM)
- SEM = SD * sqrt(1 - rxx) where rxx = test reliability
- SD = 15, rxx = .90, SEM = 4.80
- 95%CI = 64 +/- 9.60 = (54,74)
- 95%CI includes the standard/cut-off (70), therefore, there was no evidence that Ms. Muir was "Mental Defective Moron"

Second Assessment

- 95% CI = (79,99)
- the 2nd IQ score and 95% CI were above the standard/cut-off
- Ms. Muir was not "Mental Defective Moron"

A Lesson from Ms. Muir's Story

- measurement error and test reliability matter
- one ought not to make important decisions about people based on random nor systematic error
- decisions made must be legally defensible

Reliability: SEI Example

Let's assume that if 100 students took a course, 50 would give 5=SA, 35 would give 4=SA, 10 would give 3=N, 3 would give 2=D, and 2 would give 1=SD. The raters are not agreeing very much. Now, take 10 random samples of 10 students/raters (R1...R10) and observe how mean and median ratings vary (depending on which 10 raters ended up rating the course/instructor)...

Sample	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	M	Mdn*
1	5	5	3	4	5	5	5	5	5	4	4.6	4.8
2	4	4	4	5	5	4	5	5	5	5	4.6	4.7
3	4	4	5	4	5	3	5	4	4	1	3.9	4.1
4	5	5	1	5	3	4	5	5	4	5	4.2	4.7
5	5	4	4	5	5	5	5	3	5	4	4.5	4.7
6	5	4	4	5	4	5	4	3	2	5	4.1	4.2
7	4	5	4	3	4	2	4	4	5	4	3.9	4.0
8	3	5	5	5	3	5	5	5	5	5	4.6	4.9
9	5	5	5	5	5	5	5	3	5	5	4.8	4.9
10	4	5	4	4	3	1	3	3	5	3	3.5	3.5

Statistics	Range	Mean	True Mean/Median
Mean	(3.5,4.8)	4.27	4.28
Median*	(3.5,4.9)	4.45	4.50

- A single class rating can be off by as much as 1.0 on a 5-point scale.
- Medians tend to describe the center of skewed distributions better.

Mdn* = Interpolated Median (see explanation below)

Reliability: ETS SIR II Example

ETS SIR II uses 95% Confidence Intervals to flag low or high SEIs

SIR II calculates 95% confidence interval around the class mean and if the interval is below 10th percentile, the score is flagged as "reliably at or below the 10th percentile." (see Abrami, 2001, for similar approaches)

"Specifically, the scores have been compared against the score values corresponding to the 10th percentile and 90th percentile in the comparative group. If the results indicate a score is sufficiently reliable and is below the 10th percentile or above the 90th percentile, it will be flagged in the report as follows:

This class mean is reliably at or above the 90th percentile.
This class mean is reliably at or below the 10th percentile.

Scores above the 90th percentile or below the 10th percentile are flagged when there is appropriate statistical confidence that the 'true score' (i.e., the scores that would be obtained if there were no measurement error) fall within these ranges." (ETS SIR II Form)

Reliability: Drawing Valid Inferences About Instructors

An Example from McGhee (2002)

In the example below, Instructor A was rated in 10 classes and received a mean class median on an item of 3.8. Using SEM = SD * sqrt(1- rxx), we can calculate 95% Confidence Interval within which the Instructor A's true score is likely to lie. In this case, SEM is .19, Margin of Error is .38, and 95% Confidence Interval ranges from 3.4 to 4.2. For Instructor B, we have ratings available from 30 classes. Accordingly, SEM is smaller, Margin of Error is smaller, and 95% Confidence Interval ranges from 4.5 to 4.9.

Instructor	Item Mean	#Classes	SEM	Error	95% CI
A	3.8	10	.19	.38	(3.4,4.2)
B	4.7	30	.12	.23	(4.5,4.9)
C	4.1	15	.16	.32	(3.8,4.4)
D	4.6	30	.12	.23	(4.4,4.8)

McGhee (2002) (www.washington.edu/oea/pdfs/reports/OEAReport0205.pdf)

Gilmore (2000) (www.washington.edu/oea/pdfs/reports/OEAReport0002.pdf)

Reliability: Drawing Valid Inferences About Instructors

An Example from McGhee (2002)

We continue with the previous example but now we want to know whether the four instructors differ. We can determine the so called minimum significant difference (MSD) from SEM1, SEM 2, desired confidence level (95%), and numbers of ratings. If the difference in ratings of two instructors is larger than the MSD, we are justified in concluding that one instructor is better than the other (with 5% error rate). With ratings based on 10 classes, we need about 0.50 difference to conclude that one instructor is better or worse than the other.

Instructor	**#Classes**	**SEM1**	**SEM2**	**SEdiff**	**MSD**
A vs. B/D	10 v. 30	.19	.12	.23	.45
A vs. C	10 v. 15	.19	.16	.25	.50
C vs. B/D	15 v. 30	.16	.12	.20	.39

McGhee (2002) (www.washington.edu/oea/pdfs/reports/OEAReport0205.pdf)

Gilmore (2000) (www.washington.edu/oea/pdfs/reports/OEAReport0002.pdf)

Reliability: 3-Year Running Averages Increase Reliability

Reliability of Single Course Ratings

- reliability generally insufficient
- very unreliable for small classes (such as MRU's)

CAUT Recommendations

- "An evaluation of teaching performance shall consider a minimum of three years, unless it is for renewal of a contract with duration less then three years." (CAUT Model Clause on the Evaluation of Teaching Performance, 2007)

Reliability: Aggregate Report Across Multiple Courses

University of Alberta Example

One faculty's performance over number of courses (% and Mdn*)

	1=SD	2=D	3=N	4=A	5=SA	Mdn*
Well prepared	0	6	18	56	20	4.0
Explained concepts clearly	6	12	25	43	14	3.7
Spoke clearly	5	19	23	39	14	3.6
Was enthusiastic	3	14	25	43	15	3.7
Overall, was excellent	3	9	20	49	18	3.9

Reference group data

	Low Fence	25%	50%	75%	Number of Mdns
Well prepared	3.1	3.9	4.0	4.2	16
Explained concepts clearly	1.9	3.2	3.7	4.1	16
Spoke clearly	1.3	2.9	3.7	3.9	16
Was enthusiastic	2.2	3.3	3.7	4.1	16
Overall, was excellent	2.7	3.6	3.8	4.2	16

This faculty's aggregate median ratings (Mdn*) are all above Low Fences, hence not different from the faculty comparison group.

Mdn* =interpolated median

https://www.aict.ualberta.ca/units/client-services/tsqs/idq/idq-req/540-aggregate-report

Courses With New Modes of/Approaches to Instruction

"**5. Aministrators support faculty experimentation to enhance teaching and learning, and they ensure that faculty are not punished if student evaluations are uneven**
In general, learning will be improved when faculty try out different modes of instruction, with the goal of selecting appropriate modes that depend on the propensities of individual faculty (e.g., some may prefer to do mostly lecturing; others mostly group work around problems they pose in class), the nature of the students in their classes, the types of classes, and the learning goals of faculty members and students. Faculty will use various modes of instruction only when there is a clear understanding that if student ratings of faculty effectiveness temporarily drop as a result of the experimentation (e.g., an attempt at cooperative learning does not work as well as expected), there will be no negative consequences for their promotion, tenure or salary decisions. Faculty members should be free to experiment with different teaching and learning approaches without fear of reprisal." (p. 17, APA, 2011; APA Principles for Quality Undergraduate Education in Psychology)

Use of Individual SEI Items

Individual SEI Item Use

- suitable for formative purposes
- should not be used for summative (satisfactory/unsatisfactory) decisions

Overall SEIs

- suitable for summative (unidimensional) decisions

Abrami (2001), d'Apollonia and Abrami (1997), Cashin and Downey (1992)

Use of SEI Written Comments

Written Comments

- idiosyncratic
- potentially biasing
  - evaluators often focus on a few negative comments and ignore many more positive comments
- time consuming if systematically coded
- nearly impossible to develop valid interpretive guidelines for

Recommendation and Use

- should be used only for formative purposes (Cashin, 1990)
- should not be used in personnel decisions by a number of collective agreements (McGill, Brandon, SFX)
- should be used only by the instructor (CAUT, 2007): "Anonymous commentary, regardless how it is collected, shall not be seen or used by individuals other than the member" (CAUT Model Clause on the Evaluation of Teaching Performance, 2007)

Written Policy on SEI

Should policies regarding procedures, criteria, and SEI standards be in writing?

Policies on procedures, criteria, and standards for evaluation of teaching and interpreting SEIs should be in writing and available to faculty.

- Cashin (1996)
- CAUT Model Clause on the Evaluation of Teaching Performance (2007)
- U of A GFC Policies

"Committee members first understand the required standards and then assess the candidate's performance" (MRU and MRFA Tenure and Promotion Workshops 2010-2011)

Procedures, Criteria, Standards, and Interpretive Guide

Should faculty know the procedures, criteria, standards, and interpretative guidelines?

Evaluated faculty

- should know evaluation procedures
- should know criteria
- should know standards for performance
- should know how SEI scores will be interpreted and used
- should be provided with a guide detailing all of the above

Examples

- IDEA System Interpretive Guide
- SIR2 (Education Testing Service)
- U of Alberta
- U Vic

Are Evaluators Knowledgeable in Interpreting SEI Ratings?

Should evaluators be trained?

A major threat to the validity of student ratings is administrators' lack of relevant knowledge and training in interpreting the SEI results (e.g., Cashin, 1996; Menges, 2000; Abrami, 2001; Theall & Franklin, 2001; Wachtel, 1998).

- "[do not] assume that those who will examine these ratings have the necessary skills and knowledge to use them within the guidelines recommended by ratings experts... In one multi-institutional study, more than half of the faculty using ratings of the colleagues could not answer basic questions about the common statistics that appear on typical ratings reports, such as means and standard deviations" (Franklin, 2001, p.86)
- evaluators should be trained in interpreting SEIs
- evaluators should be provided with written guidelines, standards, interpretive guides, warnings, ...

Example Recommendations: Cashin (1996)

- "...
- 9. Define *major* faculty responsibilities at the *beginning* of the evaluation period... [e.g., teaching, research, service]
- 10. Define faculty sub-responsibilities at the beginning of the evaluation period and determine their weighting... [e.g., for teaching: credit instruction, student supervision, course development,...]
- 11. Define the sources of data to be used to evaluate each subresponsibility at the *beginning* of the evaluation period...
- 12. Use *multiple* sources of data...
- 13. Ensure that the data/measures are *technically* acceptable, i.e., are reliable and valid...
- 14. *Specifically* define the criteria and the standards for each subresponsibility... [e.g., 3.0 on SEIs]
- 15. Train the evaluators to evaluate...
- ..."

http://www.theideacenter.org/sites/default/files/Idea_Paper_33.pdf

Information for Students

Should students be informed about purposes and uses of SEIs?

Students may provide more constructive, accurate, and positive evaluations when informed about their purpose and uses (Ory, 2001). Students should be provided with information on

- uses of evaluations
- role of evaluations in personnel decisions
- who has access to evaluation data

See Gravestock and Gregor-Greenleaf (2008)

Formative Uses of SEIs

Should faculty be able to add targeted questions?

SEIs are mostly used for summative purposes (Gravestock & Gregor-Greenleaf, 2008). However, one stated purpose of SEIs is to improve teaching. A number of universities and evaluation systems allow faculty to add additional questions (often from a question pool) to be used for formative purposes (in addition to mandatory questions).

- IDEA System
- University of Alberta (https://www.aict.ualberta.ca/images/stories/aict/tsqs/gfcshort.pdf)
- University of Michigan
- Ryerson University

See Gravestock and Gregor-Greenleaf (2008)

Research on SEIs: Proceed with Caution!

Why are there so many contradictory statements regarding SEIs?

- 10,000+ studies on SEIs and no consensus on what SEIs actually measure (teaching effectiveness/student learning vs. student satisfaction)
- many reviews cite other reviews rather than primary sources
- many studies use correlations ignoring non-linear nature of relationships (e.g., class size and SEIs studies)
- many studies use inappropriate effect sizes (r-squared) and conclude, incorrectly, that some effect on SEIs is "small"
- many studies are methodologically poor (e.g., positive but non-significant correlations are interpreted as meaningful)
- Cohen (1981) meta-analysis (claiming correlation of 0.43 between SEI and learning) is methodologically unsound (ignored sample sizes, did not provide relevant data, etc.)
- many decades old, still cited reviews, are superseded by new findings
- ...

Summary

Things to Consider

- What do SEI measure?
- What weight should SEI have in evaluation of teaching?
- Should we develop a new SEI form? What response scales should we use?
- How should we summarize SEI data for valid interpretations? Should we include 95% confidence intervals?
- Should we control for TEIFs and how?
- What SEI performance standards should we use? Norm-referenced, criterion-referenced, or distribution referenced?
- Should procedures, criteria, & standards be known to faculty?
- Should evaluators be trained?
- Should students be informed about SEIs and their use?
- Should we facilitate formative uses of SEIs?
- ...

Links to Referenced Evaluation Systems

- IDEA System (http://www.theideacenter.org)
- ETS SIR II (http://www.ets.org/sir_ii/about)
- U of A (http://www.aict.ualberta.ca/units/client-services/tsqs/usri)
- U of Victoria (http://ltc.uvic.ca/about/index.php)

References

- Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. New Directions for Institutional Research, 109, 59-87.
- AERA, APA, & NCME (1999). The standards for educational and psychological testing. AERA, APA, & NCME.
- APA (2011). APA Principles for Quality Undergraduate Education in Psychology. APA. (http://www.apa.org/education/undergrad/principles.aspx)
- Cashin, W. E. (1996). Developing an effective faculty evaluation systems. IDEA Paper 33.
- CAUT (2007). CAUT Model Clause on the Evaluation of Teaching Performance (http://www.caut.ca/pages.asp?page=385&lang=1)
- CAUT (????). CAUT Policy on the Use of Anonymous Student Questionnaires in the Evaluation of Teaching (http://www.caut.ca/pages.asp?page=300&lang=1)
- Cashin, W.E. (1990). Students do rate different academic fields differently. In Theall, M. & Franklin, J. (Eds.), Student ratings of instruction: Issues for improving practice [Special issue]. New Directions for Teaching and Learning, 43, 113-121.
- Cashin, W.E., & Downey, R.G. (1992). Using global student rating items for summative evaluation. Journal Of Educational Psychology, 84(4), 563-572.
- Clayson (2009). Student Evaluations of Teaching: Are they related to what students learn? A Meta-analysis of the literature. Journal of Marketing Education, 31, 16-30.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.
- Coren, S. (2001). Are course evaluations a threat to academic freedom? In S.E Kahn & D. Palbih (Eds.), Academic Freedom and the Inclusive University (pp. 104-117). Vancouver: University of British Columbia Press.
- Crumbley, L, Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. Quality Assurance in Education, 9, 197-207.
- Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior. Quality Assurance in Education, 17, 377-392.

References

- Dawson, T., & Wall, M. (2009). Using the new Course Experience Survey to Assess and Improve Teaching at UVic: A manual of best practices. University of Victoria. (http://ltc.uvic.ca/initiatives/documents/CES_manual_for_chairs_09_v8.pdf)
- d'Appolonia, S., & Abrami, P.C. (1997). Navigating student ratings of instruction. American Psychologist, 52(11), 1198-1208.
- Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. Quality Assurance in Higher Education, 11, 37-46.
- Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluation of your teaching accurately. New Directions in Teaching and Learning, 87, 85-100.
- Gilmore, G. M. (2000). Drawing inferences about instructors: The inter-class reliability of student ratings of instruction. OEA Report 00-02, University of Washington. (www.washington.edu/oea/pdfs/reports/OEAReport0002.pdf)
- Gravestock & Gregor-Greenleaf (2008). Student Course Evaluations: Research, Models and Trends. Toronto: Higher Education Quality Council of Ontario.
- Green, B.P., Calderon, T.G., & Reider, B.P. (1998). A content analysis of teaching evaluation instruments used in accounting departments. Issues in Accounting Education, 13(1), 15-30.
- Gravestock, P., Greenleaf, E., & Boggs, A. M. (2009). The validity of student course evaluations: An eternal debate?
- Hoyt, D. P., & Lee, E. (2003). Understanding The IDEA System's Extraneous Variables. IDEA Research Report 6 (http://www.theideacenter.org/sites/default/files/research6.pdf)
- Hoyt, D. P., & Perera, S. (2001). Are quantitatively-oriented courses different? IDEA Research Report 3.(http://www.theideacenter.org/sites/default/files/research3.pdf)
- Hoyt, D. P., & Lee, E. (2002). Basic data for the revised IDEA System. IDEA Technical Report No. 12. (http://www.theideacenter.org/sites/default/files/techreport-12.pdf)
- Hoyt, D. P., & Pallett, W. H. (1999). Appraising teaching effectiveness: Beyond student ratings. IDEA Paper 36.
- Marsh, H.W., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. American Psychologist, 52(11), 1187-1197.
- McGhee, D. E. (2002). Drawing inferences about instructors: Constructing confidence intervals for student ratings of instruction. OEA Report 02-05, University of Washington. (http://www.washington.edu/oea/pdfs/reports/OEAReport0205.pdf)
- McKeachie, W. J. (1997). Student ratings: The validity of use. American Psychologist, 52, 1218-1225.

References

- Menges, R.J (2000). Shortcomings of research on evaluating and improving teaching in higher education. In K.E. Ryan (Ed.), Evaluating teaching in higher education: A vision for the future [Special issue]. New Directions for Teaching and Learning, 83, 5-11.
- McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. Research in Higher Education, 21, 150-158.
- MRFA (2010). MRU and MRFA Tenure and Promotion Workshops 2010-2011. Mount Royal University. Calgary, AB: MRU.
- MRFA (2011). MRU Collective Agreement. Calgary, AB: MRU. (http://www.mtroyal.ca/wcm/groups/public/documents/pdf/cafaculty.pdf)
- Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. In K. G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations. New Directions for Teaching and Learning, no. 87 (pp. 3-15). San Francisco: Jossey-Bass.
- Ory, JC., & Ryan, K .(2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. New Directions for Institutional Research, 109, 27-44.
- Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C Abrami, & L.A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special Issue]. New Directions for Institutional Research, 109, 45-56.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.
- UC USRI Review Committee (2003). The Universal Student Ratings of Instruction instrument at the University of Calgary: A review of a three year pilot project. Calgary, AB: University of Calgary. (http://www.ucalgary.ca/usri/files/usri/usri-review.pdf)
- University of Calgary Office of Institutional Analysis (2006). Distribution of grades 2005-2006. OIA Report 643. Calgary, AB: University of Calgary.
- Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: A brief review. Assessment & Evaluation in Higher Education, 29(2), 191-121.

Q & A

Questions?