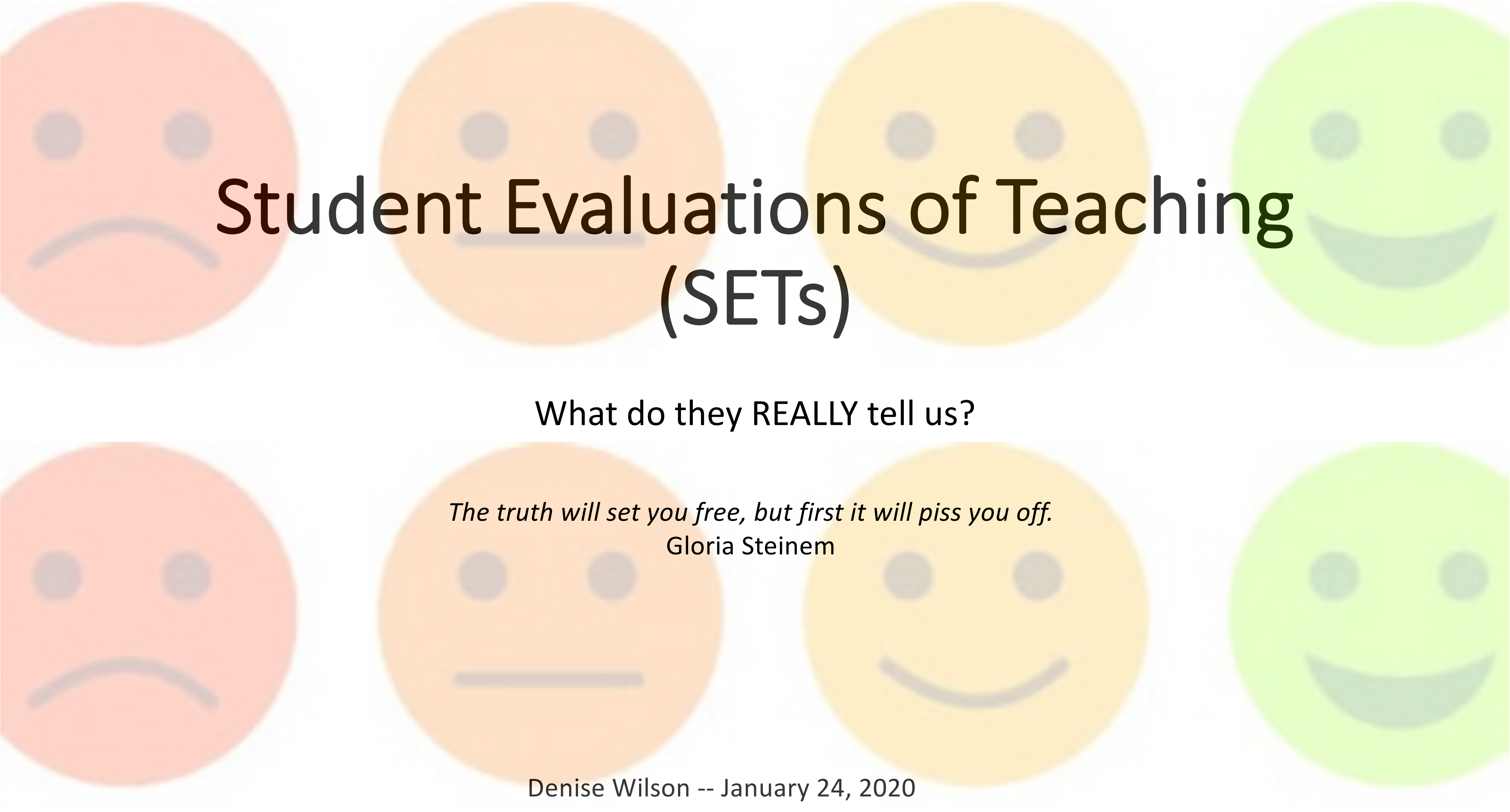


Student Evaluations of Teaching (SETs)

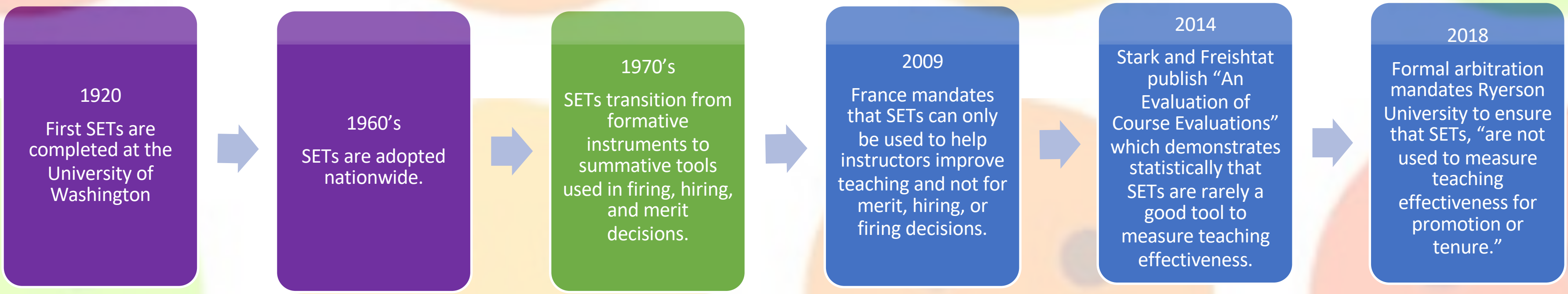
What do they REALLY tell us?

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

Denise Wilson -- January 24, 2020

An Overview of SETs

SETs have been around for a hundred years and have evolved from their original intent as formative instruments (to help instructors improve their teaching) to summative tools (to judge teaching quality).



1920

First SETs are completed at the University of Washington

1960's

SETs are adopted nationwide.

1970's

SETs transition from formative instruments to summative tools used in firing, hiring, and merit decisions.

2009

France mandates that SETs can only be used to help instructors improve teaching and not for merit, hiring, or firing decisions.

2014

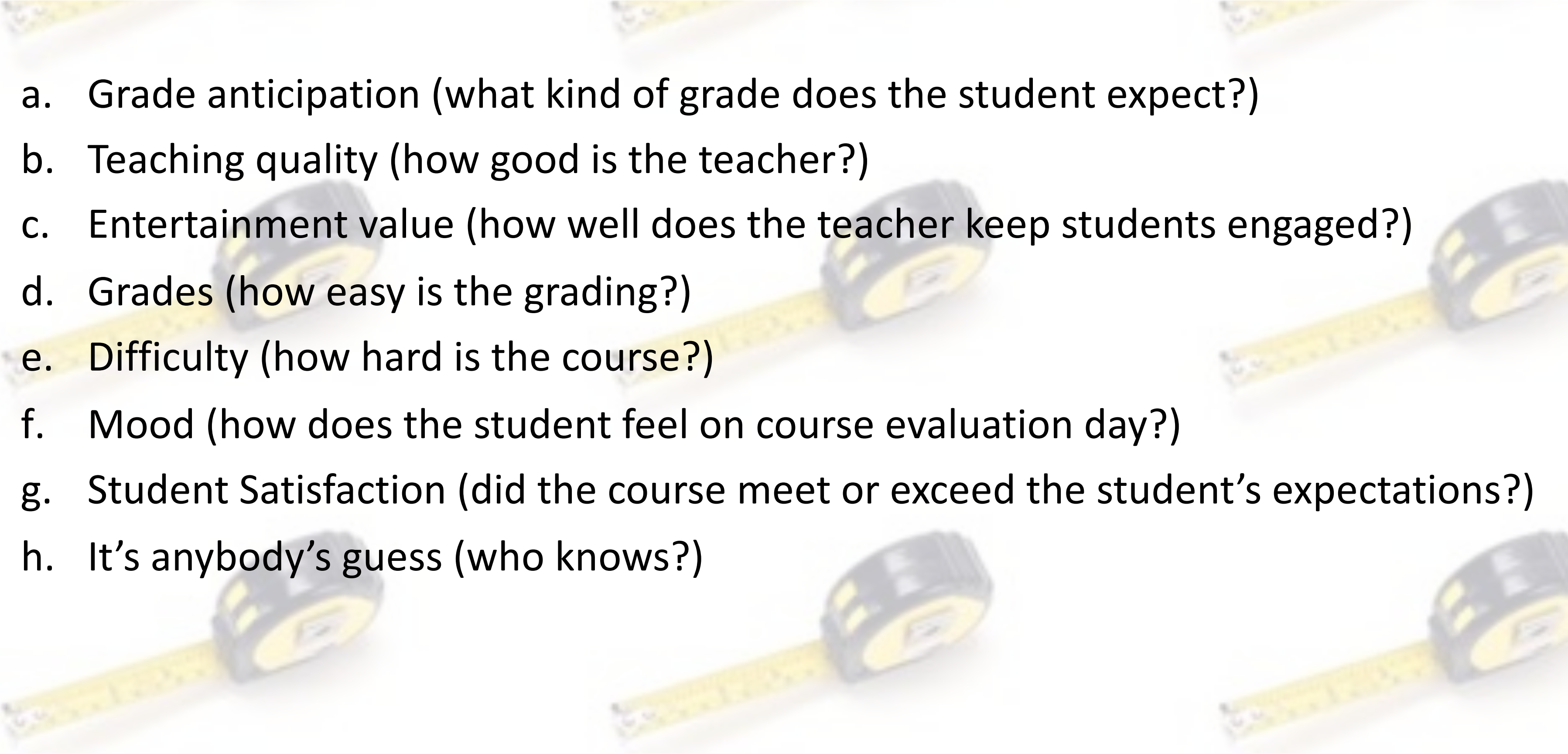
Stark and Freishtat publish "An Evaluation of Course Evaluations" which demonstrates statistically that SETs are rarely a good tool to measure teaching effectiveness.

2018

Formal arbitration mandates Ryerson University to ensure that SETs, "are not used to measure teaching effectiveness for promotion or tenure."

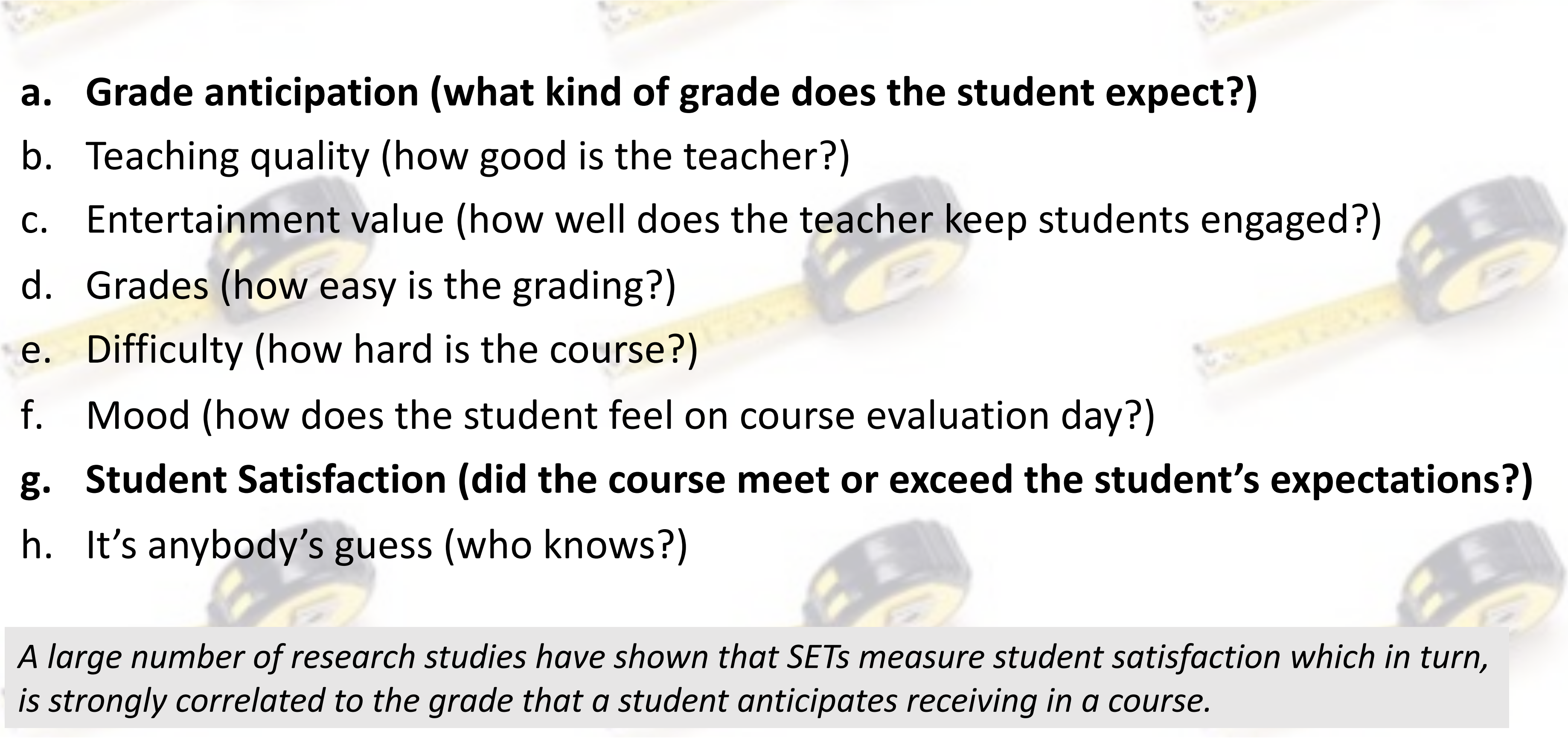
A large body of research has argued that their use as a summative instrument to measure teaching quality for personnel decisions is at best misguided and at worst unethical or illegal.

If they don't measure teaching effectiveness, What DO SETs Measure?



a. Grade anticipation (what kind of grade does the student expect?)
b. Teaching quality (how good is the teacher?)
c. Entertainment value (how well does the teacher keep students engaged?)
d. Grades (how easy is the grading?)
e. Difficulty (how hard is the course?)
f. Mood (how does the student feel on course evaluation day?)
g. Student Satisfaction (did the course meet or exceed the student's expectations?)
h. It's anybody's guess (who knows?)

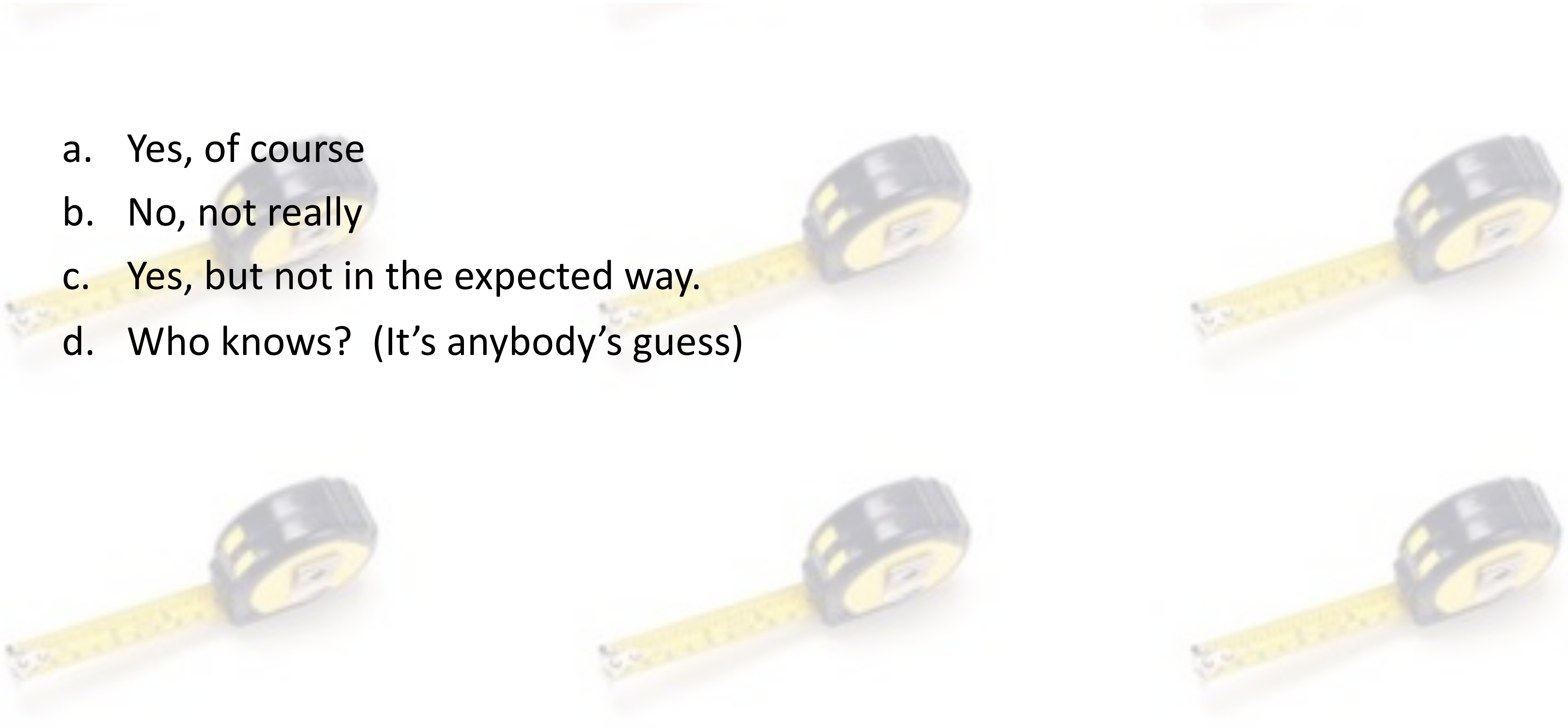
If they don’t measure teaching effectiveness, What DO SETs Measure?



a. **Grade anticipation (what kind of grade does the student expect?)**
b. Teaching quality (how good is the teacher?)
c. Entertainment value (how well does the teacher keep students engaged?)
d. Grades (how easy is the grading?)
e. Difficulty (how hard is the course?)
f. Mood (how does the student feel on course evaluation day?)
g. **Student Satisfaction (did the course meet or exceed the student’s expectations?)**
h. It’s anybody’s guess (who knows?)

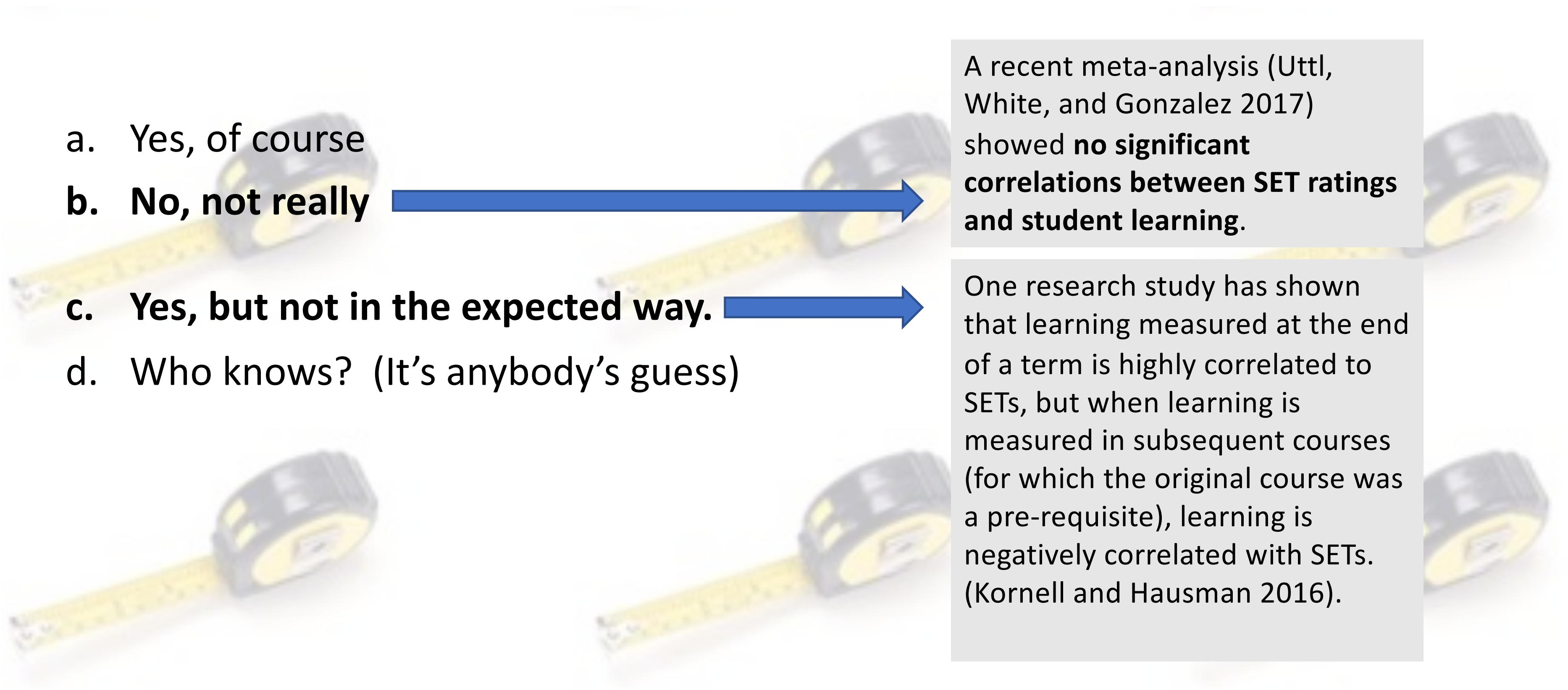
A large number of research studies have shown that SETs measure student satisfaction which in turn, is strongly correlated to the grade that a student anticipates receiving in a course.

If they don't measure teaching effectiveness, DO SETs Measure Student Learning?



a. Yes, of course
b. No, not really
c. Yes, but not in the expected way.
d. Who knows? (It's anybody's guess)

If they don't measure teaching effectiveness, Do SETs Measure Student Learning?



a. Yes, of course

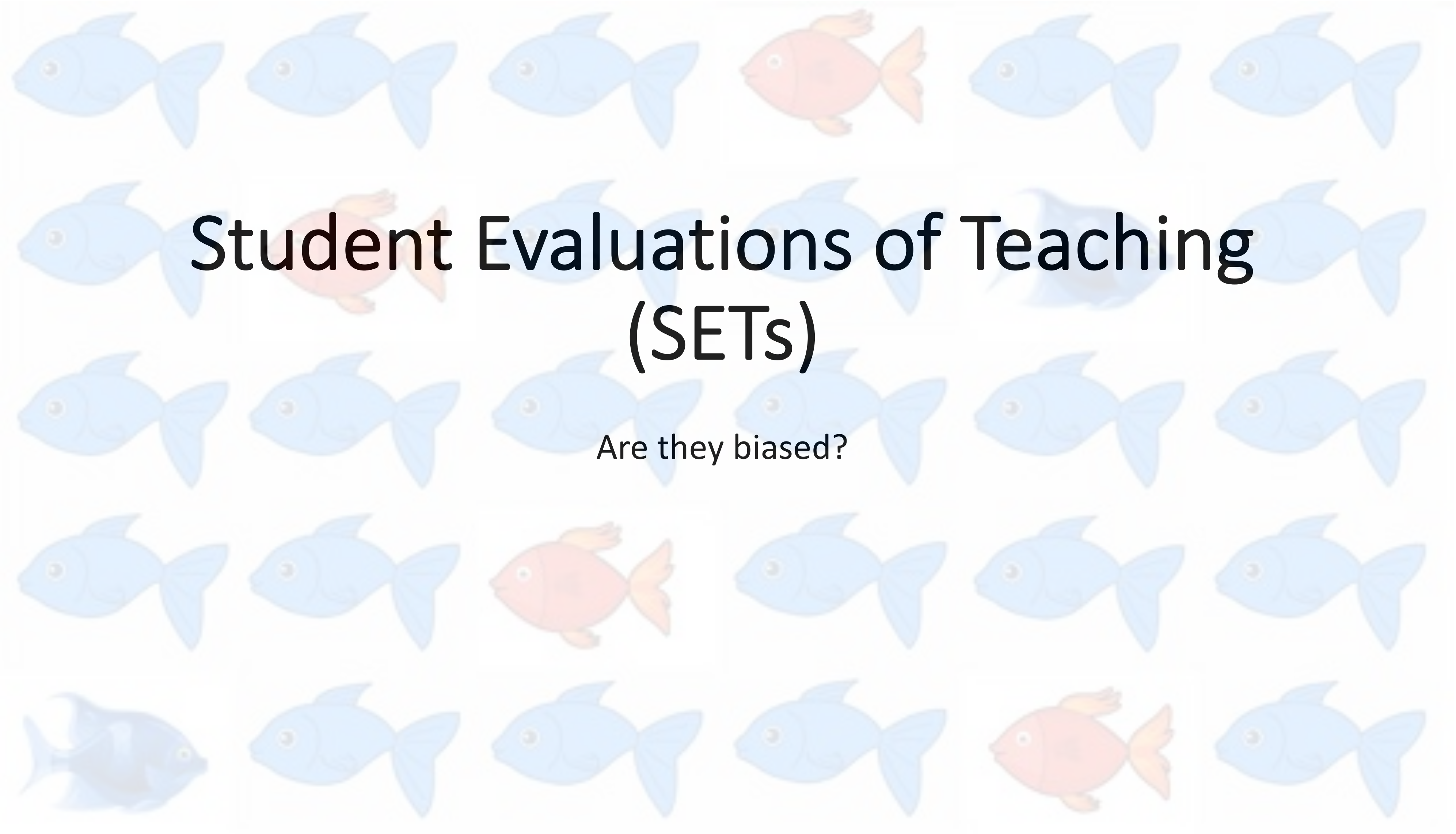
b. No, not really →

A recent meta-analysis (Uttl, White, and Gonzalez 2017) showed **no significant correlations between SET ratings and student learning**.

c. Yes, but not in the expected way. →

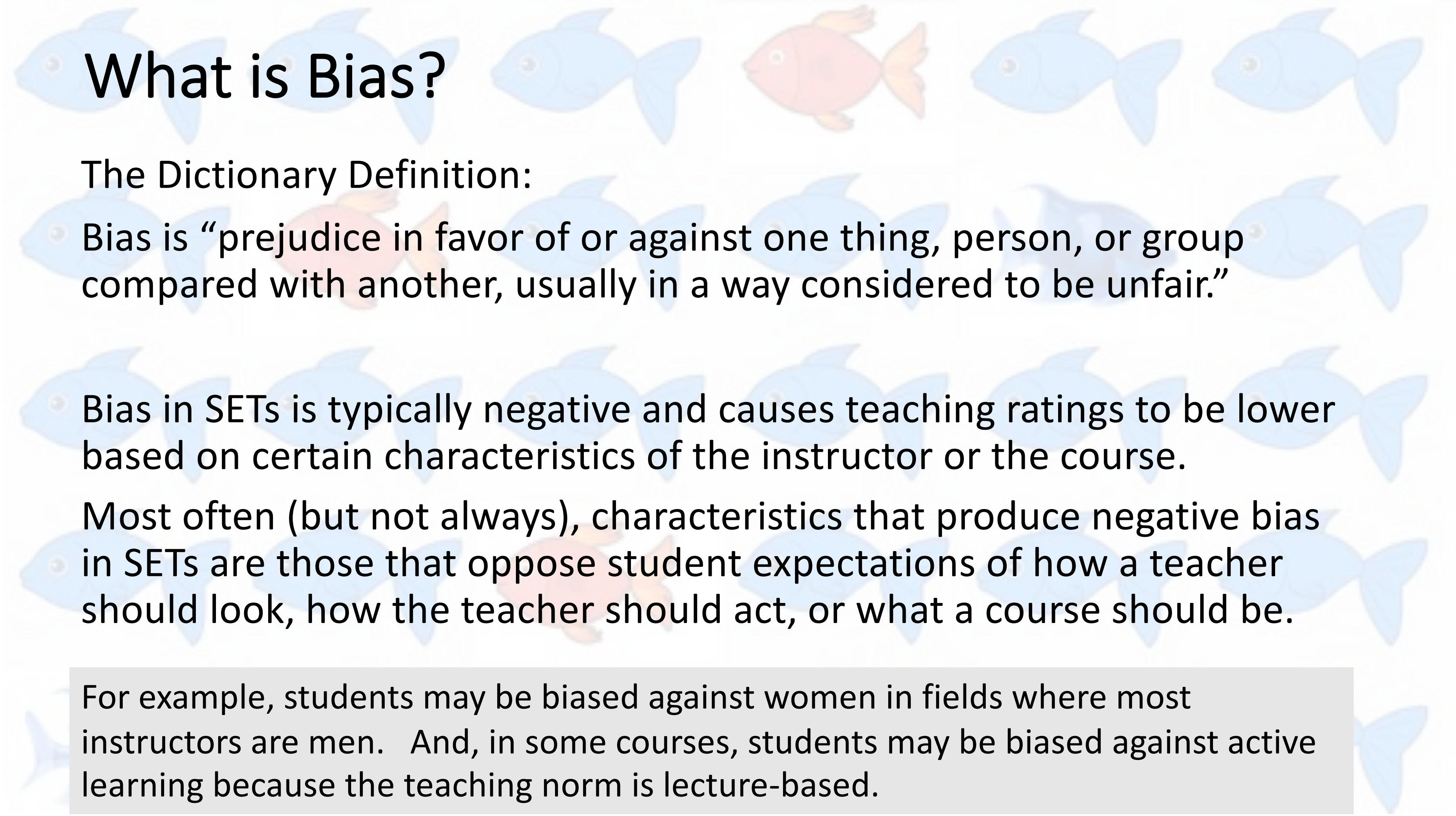
One research study has shown that learning measured at the end of a term is highly correlated to SETs, but when learning is measured in subsequent courses (for which the original course was a pre-requisite), learning is negatively correlated with SETs. (Kornell and Hausman 2016).

d. Who knows? (It's anybody's guess)



Student Evaluations of Teaching (SETs)

Are they biased?



What is Bias?

The Dictionary Definition:

Bias is “prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.”

Bias in SETs is typically negative and causes teaching ratings to be lower based on certain characteristics of the instructor or the course.

Most often (but not always), characteristics that produce negative bias in SETs are those that oppose student expectations of how a teacher should look, how the teacher should act, or what a course should be.

For example, students may be biased against women in fields where most instructors are men. And, in some courses, students may be biased against active learning because the teaching norm is lecture-based.

Do SETs get worse with:

a. Being female?

b. Being non-white?

c. Teaching with higher rank (e.g. full vs. associate professor)?

d. Teaching a larger class?

e. Teaching in Quantitative Fields (e.g. math, physics, engineering)?

f. Teaching an easier course?

g. Using Active Learning in class?

h. Providing a friendly syllabus?

i. Being physically attractive?

j. Being a non-native english speaker?

Do SETs get worse with:

a. Being female?

b. Being non-white?

c. Teaching with higher rank (e.g. full vs. associate professor)?

d. Teaching a larger class?

e. Teaching in Quantitative Fields (e.g. math, physics, engineering)?

f. Teaching an easier course?

g. Using Active Learning in class?

h. Providing a friendly syllabus?

i. Being physically attractive?

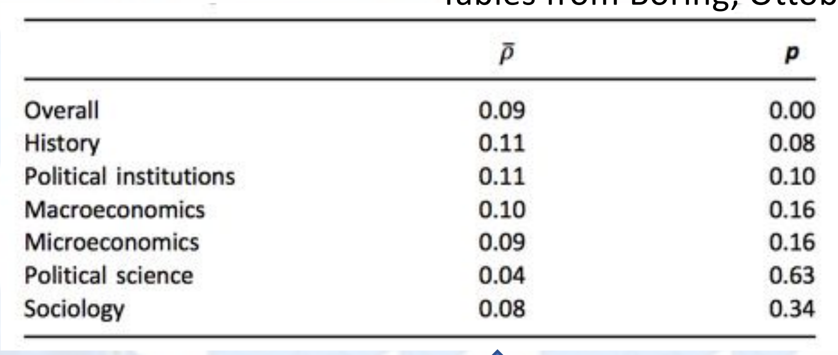
j. Being a non-native English speaker?

A Deeper Dive into Gender Bias in SETs

Boring, Ottoboni, and Stark (2016) studied over 23,000 SETs from 379 instructors and found that:

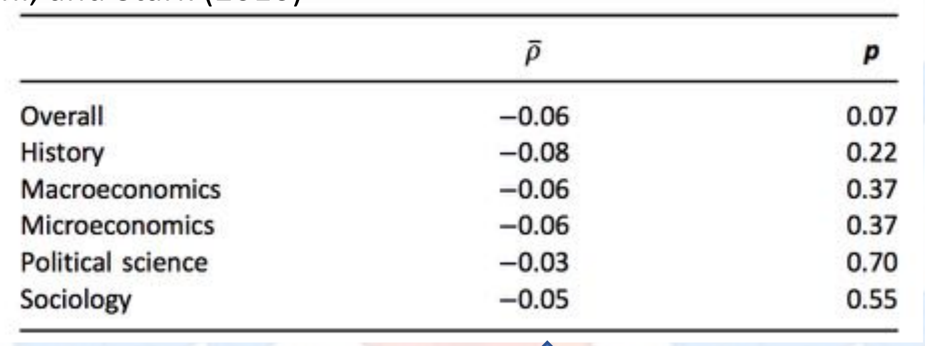
- Male instructors get significantly higher SETs than female instructors in a wide range of disciplines.
- Students may perform better on final exams with female instructors than with male instructors.

Tables from Boring, Ottoboni, and Stark (2016)



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

Correlation between male instruction and SET ratings



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

Correlation between male instruction and final exam scores

A Deeper Dive into Gender Bias in SETs

MacNeil, Driscoll, and Hunt (2015) compared SETs from four different sections of the same class run by two TAs in an on-line setting:

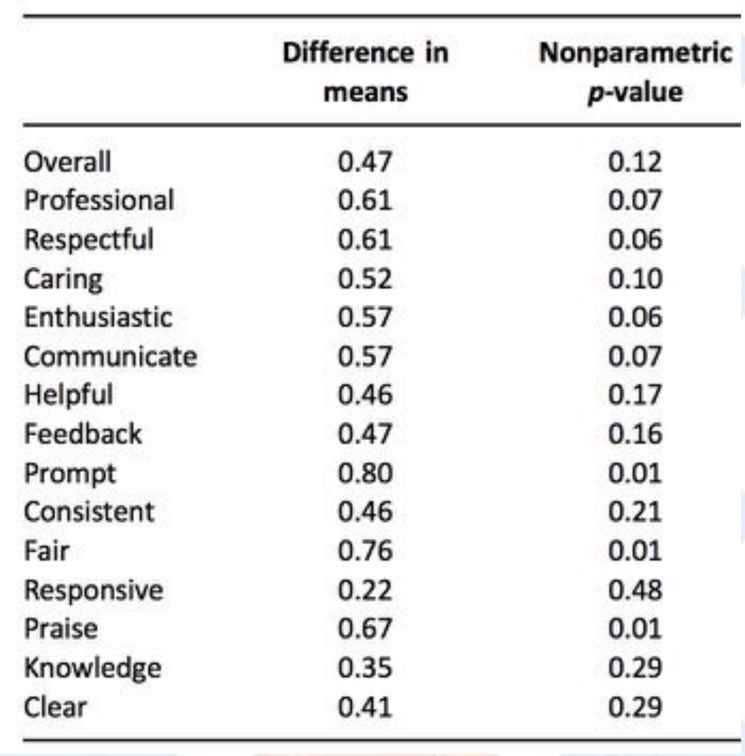
- Section #1: TA #1 (female) adopting true (female) identity
- Section #2: TA #1 (female) adopting false (TA#2, male) identity
- Section #3: TA #2 (male) adopting true (male) identity
- Section #4: TA #2 (male) adopting false (TA#1, female) identity

If no gender bias were present, SETs from Section #1 and Section #2 would demonstrate no statistically significant differences and SETs from Section #3 and Section #4 would demonstrate no statistically significant differences.

A Deeper Dive into Gender Bias in SETs

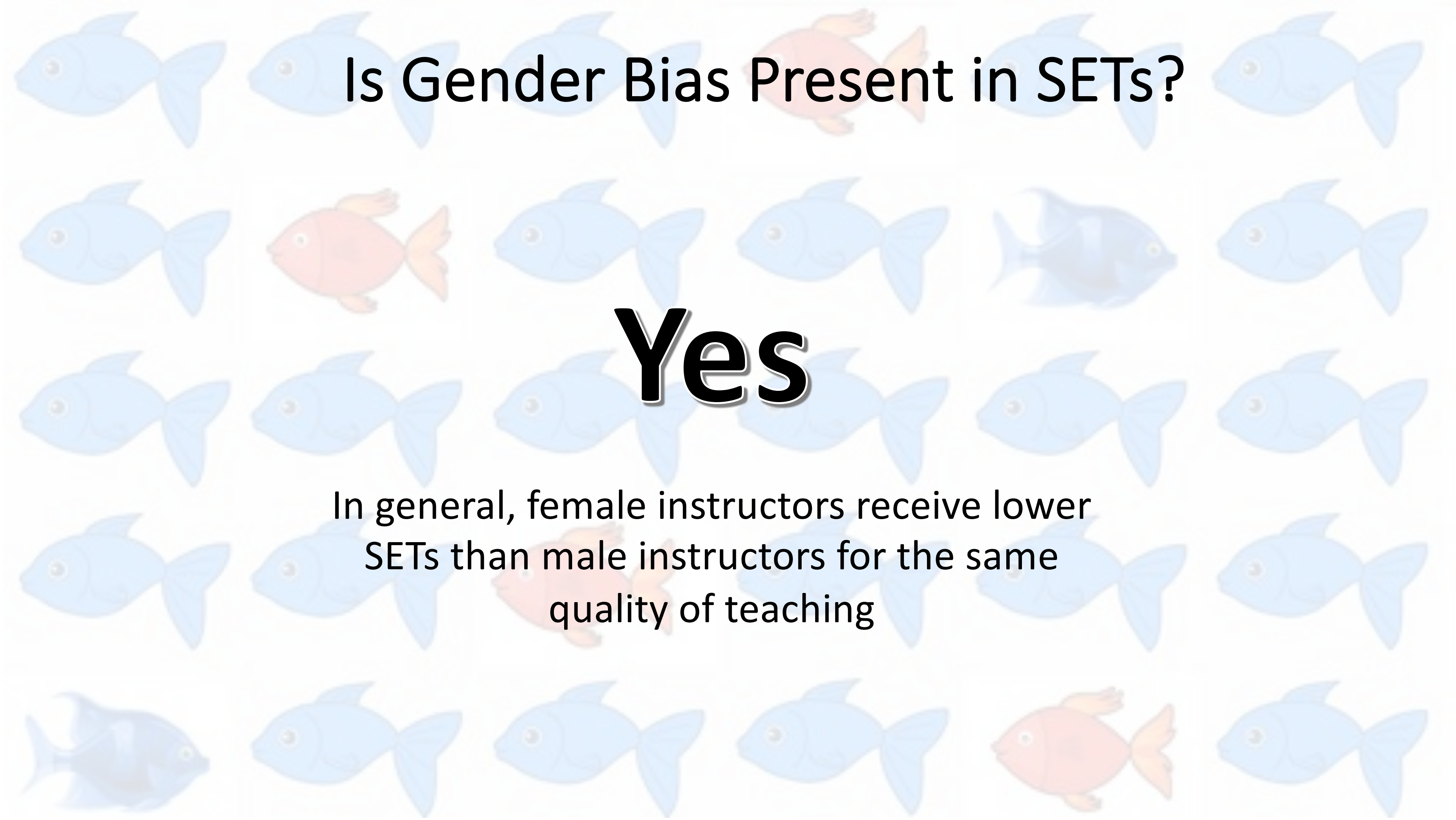
From the MacNeil, Driscoll, and Hunt (2015) study, SET ratings of Fairness, Praise, and Promptness are significantly higher for male instructors than for ”identical” female instructors (*p*<0.05).

Further, this study had a small sample size (*N* = 43) suggesting that marginally significant *p* values between 0.05 and 0.1 merit further study – students may also perceive professionalism, respect, communication, enthusiasm, and caring to be higher from male instructors than from female instructors.



	Difference in means	**Nonparametric *p*-value**
Overall	0.47	0.12
Professional	0.61	0.07
Respectful	0.61	0.06
Caring	0.52	0.10
Enthusiastic	0.57	0.06
Communicate	0.57	0.07
Helpful	0.46	0.17
Feedback	0.47	0.16
Prompt	0.80	0.01
Consistent	0.46	0.21
Fair	0.76	0.01
Responsive	0.22	0.48
Praise	0.67	0.01
Knowledge	0.35	0.29
Clear	0.41	0.29

From Boring, Ottoboni, and Stark (2016)



Is Gender Bias Present in SETs?

Yes

In general, female instructors receive lower SETs than male instructors for the same quality of teaching

Other Biases in SETs

Quantitative Fields (e.g. math, physics, engineering): class subject is strongly correlated to SET ratings with professors in quantitative fields more likely to be labelled unsatisfactory or non-excellent and receiving lower ratings overall. Professors teaching quantitative courses are more likely not to receive tenure, promotion, or merit pay when their teaching is evaluated against institution-wide standards (Uttl and Smibert 2017).

Conclusion:

Professors and instructors at all ranks in quantitative fields are at a disadvantage when compared to professors and instructors in non-quantitative fields.

Other Biases in SETs

Class size: the larger the class, the lower the SETs. This relationship is also non-linear – the decline in SETs as class size increases gets worse with increasingly larger class sizes (Spooren, Brockx, and Mortelmans 2013).

Conclusion:

Teaching large classes is bad for any professor's teaching ratings.

Other Biases in SETs

Course difficulty: the more difficult a course is or the more elementary the course, the worse the SETs. A sweet spot exists in the level of difficulty for which students will give high SETs (Spooren, Brockx, and Mortelmans 2013).

Conclusion:

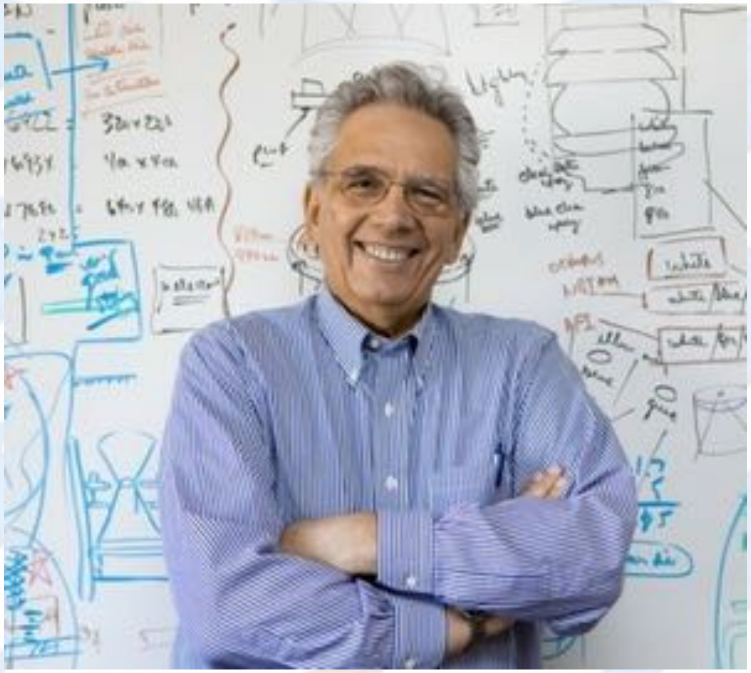
Good luck in finding the optimal difficulty for a course!

Other Biases in SETs

Image: the closer a professor looks to the ideal instructor for a particular discipline, the higher the SETs (Spooren, Brockx, and Mortelmans 2013).

Conclusion:

If you don't look like this guy in engineering, you can expect lower teaching ratings.



Other Biases in SETs

Leniency in grading: courses that are graded more leniently get higher SET ratings from students. Students rate instructors more highly if they expect a higher grade for a course regardless of actual grade, level of the course, or discipline (Boring, Ottoboni, and Stark 2016; Spooren, Brockx, and Mortelmans 2013).

Conclusion:

If students expect good things (grades), they will offer good things (SETs).

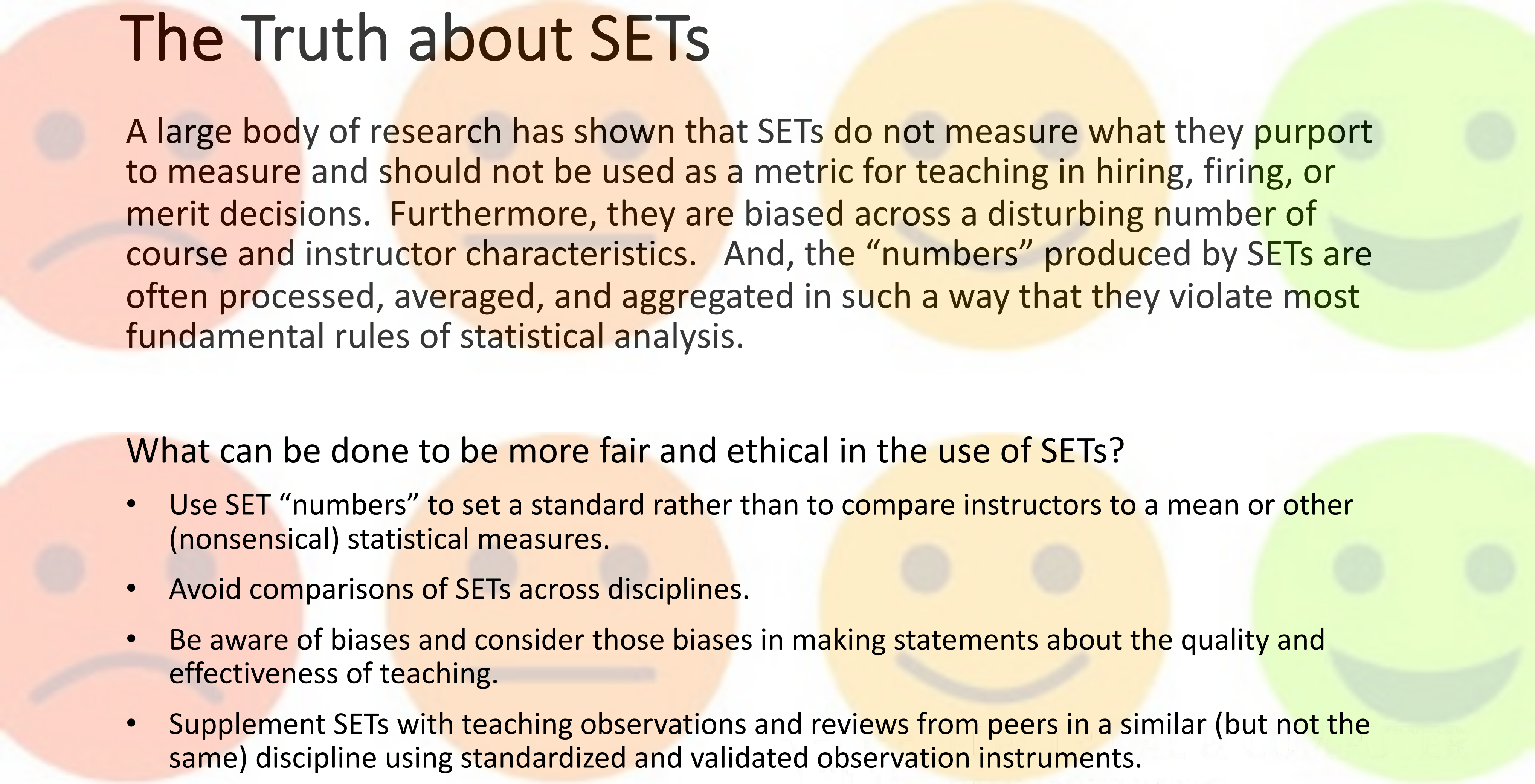
Other Biases in SETs

- **Physical Characteristics:** good looking male professors receive higher SETs than less physically attractive male professors (Spooren, Brockx, and Mortelmans 2013).
- **Race:** in upper level courses, white professors receive higher SETs than non-white faculty (Spooren, Brockx, and Mortelmans 2013).
- **English as a second language:** non-native speaking instructors receive lower SETs than native speakers (Spooren, Brockx, and Mortelmans 2013).

Other Biases in SETs

- **Active Learning:** students new to active learning tend to rate an active learning course poorly even if they end up learning better (Hill 2015).
- **Rank:** adjunct professors receive higher SETs than tenure track faculty, and full professors receive higher SETs than assistant and associate faculty (Spooren, Brockx, and Mortelmans 2013).
- **Friendliness:** the more friendly the syllabus sounds, the higher the SETs will be (Spooren, Brockx, and Mortelmans 2013).

The Truth about SETs



A large body of research has shown that SETs do not measure what they purport to measure and should not be used as a metric for teaching in hiring, firing, or merit decisions. Furthermore, they are biased across a disturbing number of course and instructor characteristics.   And, the "numbers" produced by SETs are often processed, averaged, and aggregated in such a way that they violate most fundamental rules of statistical analysis.

What can be done to be more fair and ethical in the use of SETs?

- Use SET "numbers" to set a standard rather than to compare instructors to a mean or other (nonsensical) statistical measures.
- Avoid comparisons of SETs across disciplines.
- Be aware of biases and consider those biases in making statements about the quality and effectiveness of teaching.
- Supplement SETs with teaching observations and reviews from peers in a similar (but not the same) discipline using standardized and validated observation instruments.

References

Boring, Anne. 2017. “Gender Biases in Student Evaluations of Teaching.” *Journal of Public Economics* 145 (January): 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006.

Boring, Anne, Kellie Ottoboni, and Philip Stark. 2016. “Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness.” *ScienceOpen Research*.

Clayson, Dennis E. 2009. “Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature.” *Journal of Marketing Education* 31 (1): 16–30.

Cohen, Peter A. 1981. “Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies.” *Review of Educational Research* 51 (3): 281–309.

Feldman, Kenneth A. 1989. “The Association between Student Ratings of Specific Instructional Dimensions and Student Achievement: Refining and Extending the Synthesis of Data from Multisection Validity Studies.” *Research in Higher Education* 30 (6): 583–645.

Hill, Phil. 2015. “Student Course Evaluations and Impact on Active Learning.” E-Literate. 2015. https://eliterate.us/student-course-evaluations/.

Kornell, Nate, and Hannah Hausman. 2016. “Do the Best Teachers Get the Best Ratings?” *Frontiers in Psychology* 7. https://doi.org/10.3389/fpsyg.2016.00570.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. “What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching.” *Innovative Higher Education* 40 (4): 291–303.

Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. 2013. “On the Validity of Student Evaluation of Teaching: The State of the Art.” *Review of Educational Research* 83 (4): 598–642.

Stark, P., & R. Freishtat. 2014. “An evaluation of course evaluations.” ScienceOpen. *Center for Teaching and Learning, University of California, Berkley.* https://www.scienceopen.com/document/read?vid=42e6aae5-246b-4900-8015-dc99b467b6e4.

Uttl, Bob, and Dylan Smibert. 2017. “Student Evaluations of Teaching: Teaching Quantitative Courses Can Be Hazardous to One’s Career.” *PeerJ* 5 (May): e3299. https://doi.org/10.7717/peerj.3299.

Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. 2017. “Meta-Analysis of Faculty’s Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related.” *Studies in Educational Evaluation* 54: 22–42.