Using data to confuse and deceive

Matthew Brett

How to lie with statistics

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

(Huff 1954)

"Research Excellence" Framework

> Over thirty years the RAE / REF has supported a sustained improvement in the quality and productivity of the UK research base. It is used by universities to attract students, staff and external funding.

Research Excellence Framework (REF) review: Building on success and learning from experience by Lord Nicholas Stern.

What is improvement?

	1947–66	1967–86	1987–2006
US	50	88	126
UK	20	25	9

Number of science Nobel prizes by country and time period. REA / REF started in 1986.

(Charlton 2007)

What is improvement?

> In contrast to the picture of long term decline in Nobel-prize-winning revolutionary science; UK and European scientific production (also

> that of Chinese science) is probably catching up with the USA in terms of scientometric measures such as numbers of publications and citations.

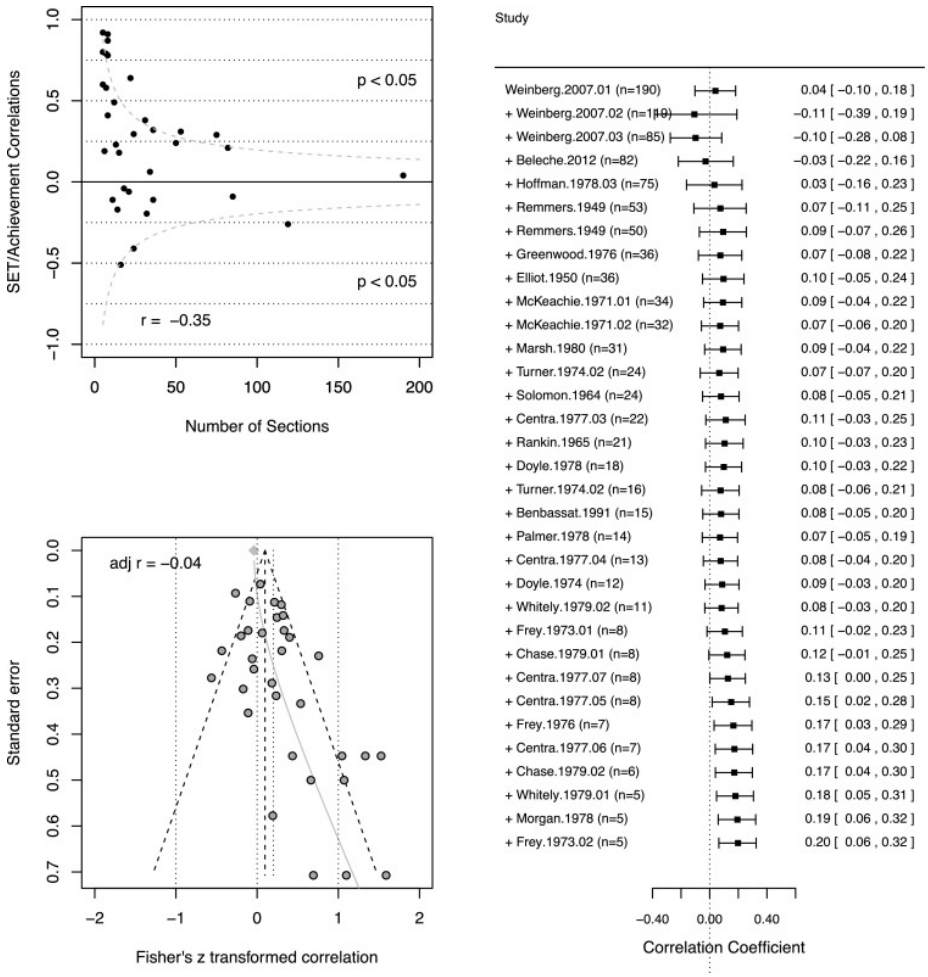
(Charlton 2007)

"Teaching Excellence" Framework

- student satisfaction using the teaching on course, assessment and feedback and academic support scales from the National Student Survey;
- retention using [Higher Education Statistics Agency] UK Performance Indicators;
- proportion in employment in further study using 6 month [Destination of Leavers from Higher Education Survey].

UK government Higher education: success as a knowledge economy - white paper

Student Evaluations do not measure teaching effectiveness



(Uttl, White, and Gonzalez 2017)

Summary of meta analysis

> The reported correlations between [Student Evaluations of Teaching] ratings and learning are completely consistent with randomly generating correlations from the population correlation with rho = 0 and applying publication selection bias.

(Uttl, White, and Gonzalez 2017)

What do student ratings measure?

- the grade the student expects to get;
- the subject being taught;
- whether the instructor is white and male
- biscuits

Expected grade

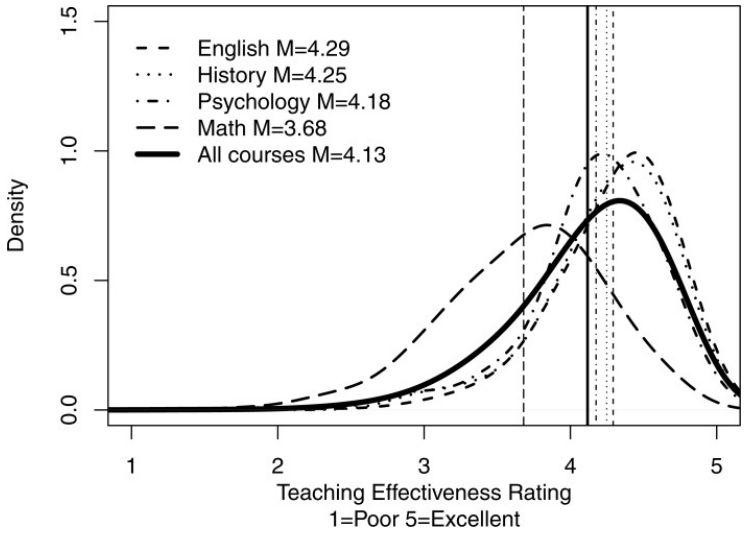
Table 7. Average correlation between SET and interim grades

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

p-values are one-sided.

(Boring, Ottoboni, and Stark 2016)

Subject being taught



14872 class summary evaluations from New York University

(Uttl and Smibert 2017)

Whether the instructor is male

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

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

p-values are two-sided.

(Boring, Ottoboni, and Stark 2016)

Biscuits

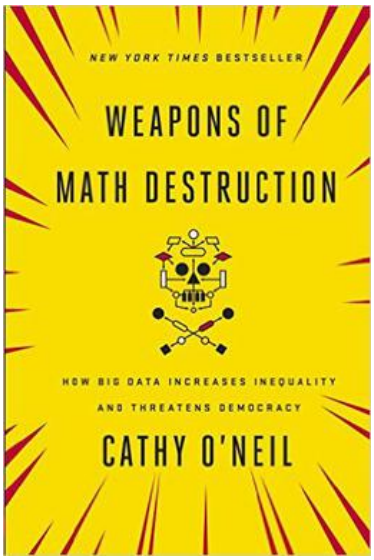
> The cookie group evaluated teachers significantly better than the control group (113.4 *pm* 4.9 versus 109.2 *pm* 7.3; p = 0.001, effect size 0.68). Course material was considered better (10.1 *pm* 2.3 versus 8.4 *pm* 2.8; p = 0.001, effect size 0.66) and summation scores evaluating the course overall were significantly higher (224.5 *pm* 12.5 versus 217.2 *pm* 16.1; p = 0.008, effect size 0.51) in the cookie group.

(Hessler et al. 2018)

For discussion

- Given the data here, why did the government want to use student ratings to evaluate teaching?
- Imagine I propose an alternative rating of teaching excellence, which is a random number between 1 and 10. Make arguments for preferring a metric based on student evaluations.
- If you could decide how to evaluate teaching, what do you propose?

Algorithms and public policy



(O'Neil 2016)

Risk scores for re-offending

Full analysis at https://github.com/propublica/compas-analysis

COMPAS

"Correctional Offender Management Profiling for Alternative Sanctions"

Proprietary algorithm

Marketed by Northpointe

Risk score officer questions

8. How many prior misdemeanor assault offense arrests (not sex, or domestic violence) as an adult?
   ○ 0 ○ 1 ○ 2 ○ 3+
9. How many prior family violence offense arrests as an adult?
   ○ 0 ○ 1 ○ 2 ○ 3+
10. How many prior sex offense arrests (with force) as an adult?
    ○ 0 ○ 1 ○ 2 ○ 3+
11. How many prior weapons offense arrests as an adult?
    ○ 0 ○ 1 ○ 2 ○ 3+
12. How many prior drug trafficking/sales offense arrests?
    ○ 0 ○ 1 ○ 2 ○ 3+

COMPAS questionnaire

Risk score offender questions

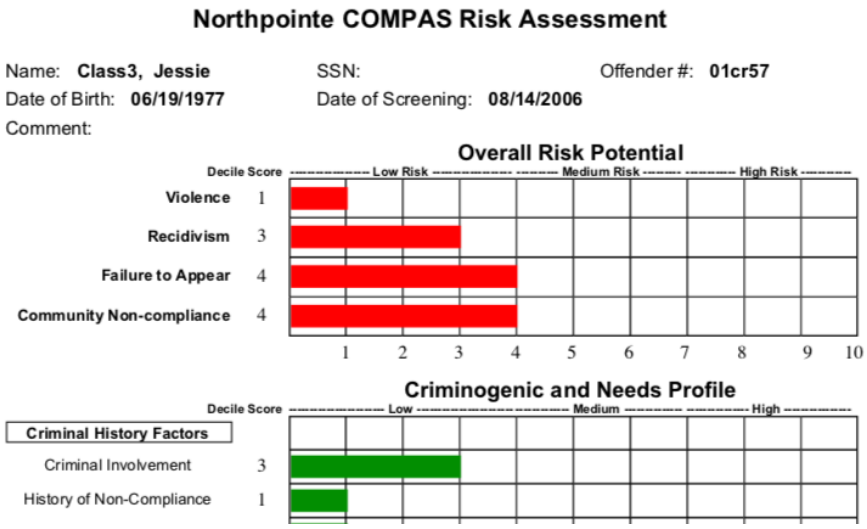
24. In the last 12 months before this incarceration, how often did you have contact with your family (may be in person, phone, mail)?

    ○ No family ○ Never ○ Less than once/month ○ Once per week ○ Daily
25. In the last 12 months before this incarceration, how often did you move?
    ○ Never ○ 1 ○ 2 ○ 3 ○ 4 ○ 5+
26. Did you have a regular living situation prior to your current incarceration (an address where you routinely stayed and could be reached)?

    ○ No ○ Yes
27. How long had you been living at your last address prior to this incarceration?
    ○ 0-5 mo. ○ 6-11 mo. ○ 1-3 yrs. ○ 4-5 yrs. ○ 6+ yrs.
28. Was there a telephone at this residence (a cell phone is an appropriate alternative)?
    ○ No ○ Yes

COMPAS questionnaire

Risk score results



COMPAS results

Background - a crash course on machine learning

See https://www.inferentialthinking.com/chapters/17/Classification

Prediction errors by race

Black Defendants			White Defendants		
	Low	High		Low	High
Survived	990	805	Survived	1139	349
Recidivated	532	1369	Recidivated	461	505
FP rate: 44.85			FP rate: 23.45		
FN rate: 27.99			FN rate: 47.72		
PPV: 0.63			PPV: 0.59		
NPV: 0.65			NPV: 0.71		
LR+: 1.61			LR+: 2.23		
LR-: 0.51			LR-: 0.62		

See this description of the analysis.

Garbage in, Gospel out

From Garbage in, garbage out

Assessing risk scores

> Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice.

Eric Holder, US Attorney General, 2014 (quoted in Propublica).

For discussion

- Given the importance of the risk scores, and their potential for bias, why was there so little study of their performance?
- What would be your recommendation for the use of the Northpointe system studied here?
- What recommendations would you make, to a state that was considering using a system like this? What resources would you provide to help them?

Is this the end?

Yes, it's the end.

References

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

Charlton, Bruce G. 2007. "Scientometric Identification of Elite 'Revolutionary Science'research Institutions by Analysis of Trends in Nobel Prizes 1947–2006." *Medical Hypotheses* 5 (68): 931–34.

Hessler, Michael, Daniel M Pöpping, Hanna Hollstein, Hendrik Ohlenburg, Philip H Arnemann, Christina Massoth, Laura M Seidel, Alexander Zarbock, and Manuel Wenk. 2018. "Availability of Cookies During an Academic Course Session Affects Evaluation of Teaching." *Medical Education* 52 (10): 1064–72. https://doi.org/10.1111/medu.13627.

Huff, Darrell. 1954. "How to Lie with Statistics."

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. "New York": Crown Publishing Group.

Uttl, Bob, and Dylan Smibert. 2017. "Student Evaluations of Teaching: Teaching Quantitative Courses Can Be Hazardous to One's Career." *PeerJ* 5: e3299.

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