SELF Research Centre

Self-concept Enhancement and Learning Facilitation

Assessing Educational Effectiveness

Herbert W. Marsh,

Benjamin Nagengast,

John Fletcher

University of Oxford

Research Methodology Festival 2010

Value-Added (VA) Models, Contextual Value-Added (CVA) Models, and League Tables

Value Added (VA) & Contextual Value Added (CVA)

Raudenbush and Willms (1995) and Willms and Raudenbush (1989)

- **Value-Added (VA) Models** adjust for student intake variables/characteristics.
  - Particularly relevant to school choice.
- **Contextual Value Added (CVA)** Models, in addition, adjust for school context or compositional effects (e.g., school average measures of prior attainment, SES, disadvantage).
  - Designed to assess school practices/processes that explain school differences;
  - used to identify most and least effective schools in relation to aspects under their control.

League Tables

Leckie & Goldstein (2009); Sammons (1996)

- Ranking of institutions within a sector (e.g., education, health, sporting teams) in relation to raw outcomes or outcomes adjusted for input and/or contextual variables.
- Goldstein (1995, 1998): need to consider the confidence intervals; ranks misleading.
- Many years out of date in terms of Parental choice; much decay in predictions over time (~ no reliable differences between schools in long-term predictions of VA effects).
- Distinction between absolute and relative standards.

Education Effects = School Effects + teacher effects + ???

Modest Size of School Effects: ~5% of variance

- Historically, there were positively biased (inflated) estimates.
- multilevel model estimates more modest.
- For current CVA Models (better models & input variables), schools explain ~5%** of variance student achievement in CVA models (50%+ due to prior attainment).
- With further improvements in statistical models, % variance explained values will become even smaller.

** This value would be higher if computed as a % of residual variance (i.e., ~10% rather than ~5% if 50% of total variance was explained)

Education Effects = School Effects + teacher effects + ???

- In UK the main focus is on school effects.
- In US, a major focus of much research is on the effects of individual teachers.
- Educational effects = school + teacher + other sources.
- Some strong & weak teachers in all schools.

Teacher vs. School Effects: Why Teachers Are Important

- Rowe et al. Australian research: school effects very small after controlling class/teacher effects.
- Monk (1992, p. 320) : *how much a student learns depends on the identity of the instructor to which the student is assigned.*
- econometric models of teacher effects assume no school effects
- US teacher merit pay and based on student test scores.
- Interventions to improve teaching effectiveness more effective if aimed at teachers
- Why Has UK Ignored Teacher in CVA Models?

Why Has UK CVAs Largely Ignored Teacher Effects?

- **Do not have appropriate data?**
  - Some countries/states have merit pay based on Teacher CVA models, so it can be done;
  - Could add teacher ID to current system, but requires yearly common tests in all subjects at secondary school
  - might be viable for primary schools
  - In theory, UK inspections are at teacher level
- **Models are not reliable?** Teacher CVA Models more complicated, difficult to interpret; However, apparently little attempt to do so in UK.
- **Undesirable side effects** e.g. teaching to test, narrowing of curriculum, etc.
- **Political Pressure & Lobby Groups?**

How About Effects of:

- **Principal /School Leadership**
- **Classroom** (students sitting in the classroom rather than the teacher standing in front of the class)
- **Department** (Do specific departments in a school consistently outperform other departments?)

Fragility of Causal Inferences

Nature of Causal Inference

- Strongest basis of casual inference are simple experimental studies with random assignment
- Education analogy is to treat each teacher/school as a separate “intervention” and randomly assign students. BUT

Nature of Causal Inference: Discipline Differences

There are alternative quasi-experimental approaches to inferring causality.

- **Econometric models** typically start with an explicit model and establish conditions under which it gives unbiased results and parameter estimates.
- **Psychometric models** tend to be more flexible, but pay less attention to assumptions and validity of causal assumptions.

Econometric Models of Value Added

Teacher VA Econometric Models Harris & McCaffrey, 2009

$$A_{it} = \omega T_{it} + \varphi_1 Z_{it} + \lambda A_{it-1} + \gamma_i + \eta_{it}$$

ωT_{it} = effect of the Teacher at age t. The teacher effect is the mean of the teacher's classroom (but represents combined effect of teacher & classroom). Could also include fixed teacher effects (eg. Age, credentials) but these are part of the teacher effect

$\varphi_1 Z_{it}$ = effect of school other than teachers.

λA_{it-1} = effect of prior ACH with a decay function λ

A_{it} = educational output of student i at time t for school/teacher.

γ_i = the fixed effect contribution of an student i (including family and environmental effects)

However, econometric models typically based on a huge number of problematic assumptions.

Backward Causation:
Bias for selection/sorting effects

VA/CVA models assumed to control for all prior effects, but

- Econometric Teacher VA models in US: Rothstein & others
  - VA estimates for Yr5 students were significantly related to VA estimates for Yr4;
  - Implies backward causation (i.e, Yr5 effects “cause” Yr4 effects) thus violating assumptions of models;
  - bias up to ¾ size of teacher effects
- Psychometric School CVA models in UK: Goldstein & colleagues showed much VA attributed to final high school
  - is due to primary schools;
  - Is due to previous high schools when change schools.

Psychometric Models in Educational Research: Measurement and Sampling Error

Multi-Level VA & CVA Models.

Educational systems are inherently multilevel.

Most school effectiveness research is now based on multilevel models.

- Analyses that ignore these multi-level effects and particularly their standard errors are inherently biased unless very restrictive assumptions are met.
- However, measurement error and sampling error has typically been ignored in multilevel models – including VA and CVA models of school and teacher effects

Multi-level VA & CVA Models

Current CVA and VA models of both school and teacher effects implicitly/explicitly assume that there is:

- no measurement error in L1 student level variables (e.g., student ACH & background variables);
- no measurement error in L2 school or teacher-level constructs—true school level constructs and aggregates student-level variables (e.g., school-average ACH).
- No sampling error in estimating L2 variables from individual student L1 data
  - e.g. school-average ACH is taken to be a true population value—with no sampling error—rather than an estimate based on a sample of students with some uncertainty.

Latent Variable (LV) Models with multiple indicators to control measurement error.

Increasing emphasis LV models with each construct (e.g., student ACH, background) based on multiple indicators.

- In LV models measurement error is estimated as part of the model; estimates corrected for measurement error.
- Possible to fit complex models of measurement error.
- Ignoring measurement error in prior student ACH negatively biases these estimates AND positively biases effects in other parts of the model (including school effects) as in the Phantom Effect.

Phantom Effects: Consequences of Failure to Control for Measurement Error

Harker & Tymms (2004) and others identified **Phantom Effects** ("now you see it, now you don't").

For a very large sample of UK primary schools showed:

- almost no school effects.
- systematically added random error to their pre-test measures to simulate typical value-added estimates; school effects became; schools with initially more able students were seen to be more effective;
- these "effects" were known to be an artefact of the added measurement error—"now you see it, now you don't".

Multi-level Latent Variable Models

Building on work by Goldstein, McDonald, Tymms and others we are evaluating doubly-latent multilevel models in relation to VA and CVA models. The models are doubly latent in relation to measurement error and sampling error:

- Multiple indicators of student level variables (e.g., student ACH), control for L1 measurement error;
- Multiple indicators at the school level, control for L2 measurement error.
- When samples of students used to estimate school-level constructs, control for sampling error

2x2 Taxonomy of Contextual Models

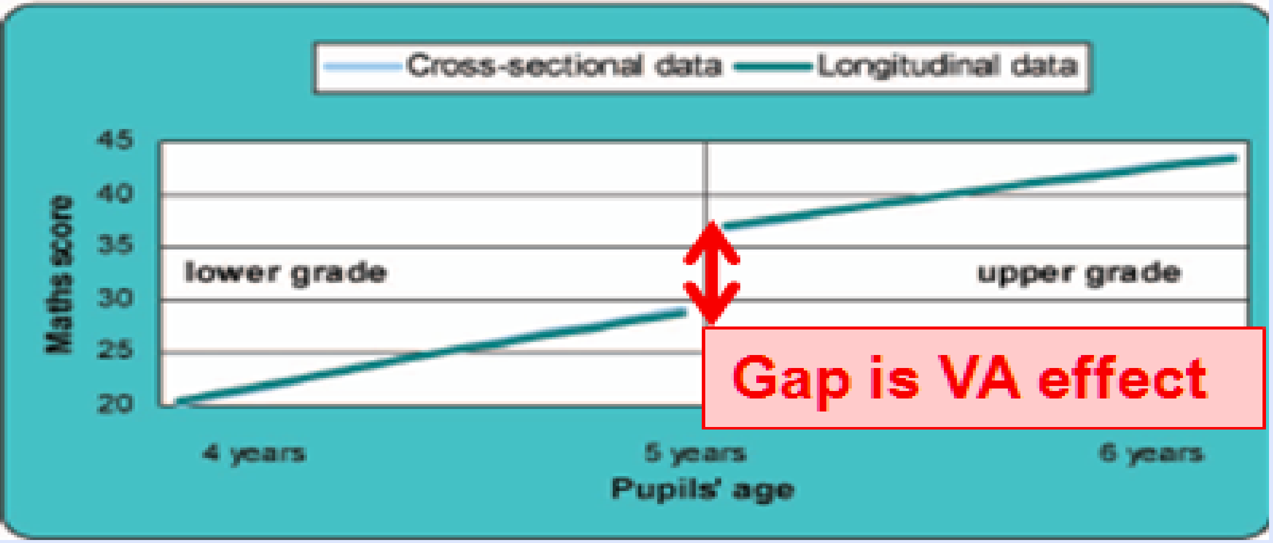
Sampling of Persons (sampling Error)

Sampling of Items (Measurement Error)	Manifest	Latent
Manifest	**Doubly Manifest** • L1-single manifest indicators (one score per factor, manifest L1 constructs) • *manifest aggregation of L1 constructs to form L2 constructs* Lüdtke, Marsh, et al. 2008, Psych Methods	**Manifest-Measurement/ Latent-Aggregation** • L1-single manifest indicators (one score per factor, manifest L1 constructs) • *latent aggregation of L1 constructs to form L2 constructs* Lüdtke, Marsh, et al. 2008, Psych Methods
Latent	**Latent-Measurement/ Manifest-Aggregation** Multiple indicators (L1 & L2 constructs are latent) • Manifest Aggregation of L1 multiple indicators Marsh, Lüdtke, et al. 2009, MultVar Beh Res	**Doubly Latent** • L1-Multiple indicators (L1 constructs are latent) • L2-Latent Aggregation of multiple L1 indicators Marsh, Lüdtke, et al. 2009, MultVar Beh Res

L1 Measurement Error & Sampling Error

Ferrão & Goldstein (2009; Goldstein, Kounali & Robinson, 2008) used a two-step method to control measurement error in traditional value-added models. They considered:

- L1 measurement error at student level
- sampling error in aggregating from student- to school-level constructs based on samples of students.
- Argued for two-stage approach: estimate reliability and then include these estimates into the CVA model
- We are currently comparing our taxonomy and Goldstein approach with simulated and UK PLASC data



Multilevel Regression-Discontinuity (RD) VA Models

Luyten, Tymms & Jones, 2009

- produces "absolute" VA estimates from one year to the next and school-level variation ("relative" VA).
- Gain due to one school year is "gap" between regression functions relating age to ACH for each year group at point of discontinuity (oldest students in lower grade & youngest in higher grade).
- UK is well-suited because rigid about starting age birthday
- Applied longitudinally (same students in two consecutive years) or cross-sectionally (two consecutive year groups, so do not need pretests); Here both analyses give ~same results.

Use of Multiple Outcome Measures: Alternatives to VA/CVA Models

Alternative Approaches Coe Bell Little (2008, ETS)

Measures of school/teacher effectiveness are not valid in of themselves; it is interpretations/uses that must be validated.

Need to distinguish between:

- inputs (what teacher brings), processes (what the teacher does), and outputs (student results, but other outcomes as well).
- Formative and summative evaluation

Alternative Outcomes Include:

- School/Teacher/Classroom observations;
- Instructional artifacts (lesson plans, student work, marking, etc)
- Teacher portfolios;
- Teacher self-reports.

Summary of VA/CVA Models

- Useful for assessing school/teacher effectiveness and research tool;
- Modestly correlated with other measures of teaching;
- Better than teacher credential and experience

However:

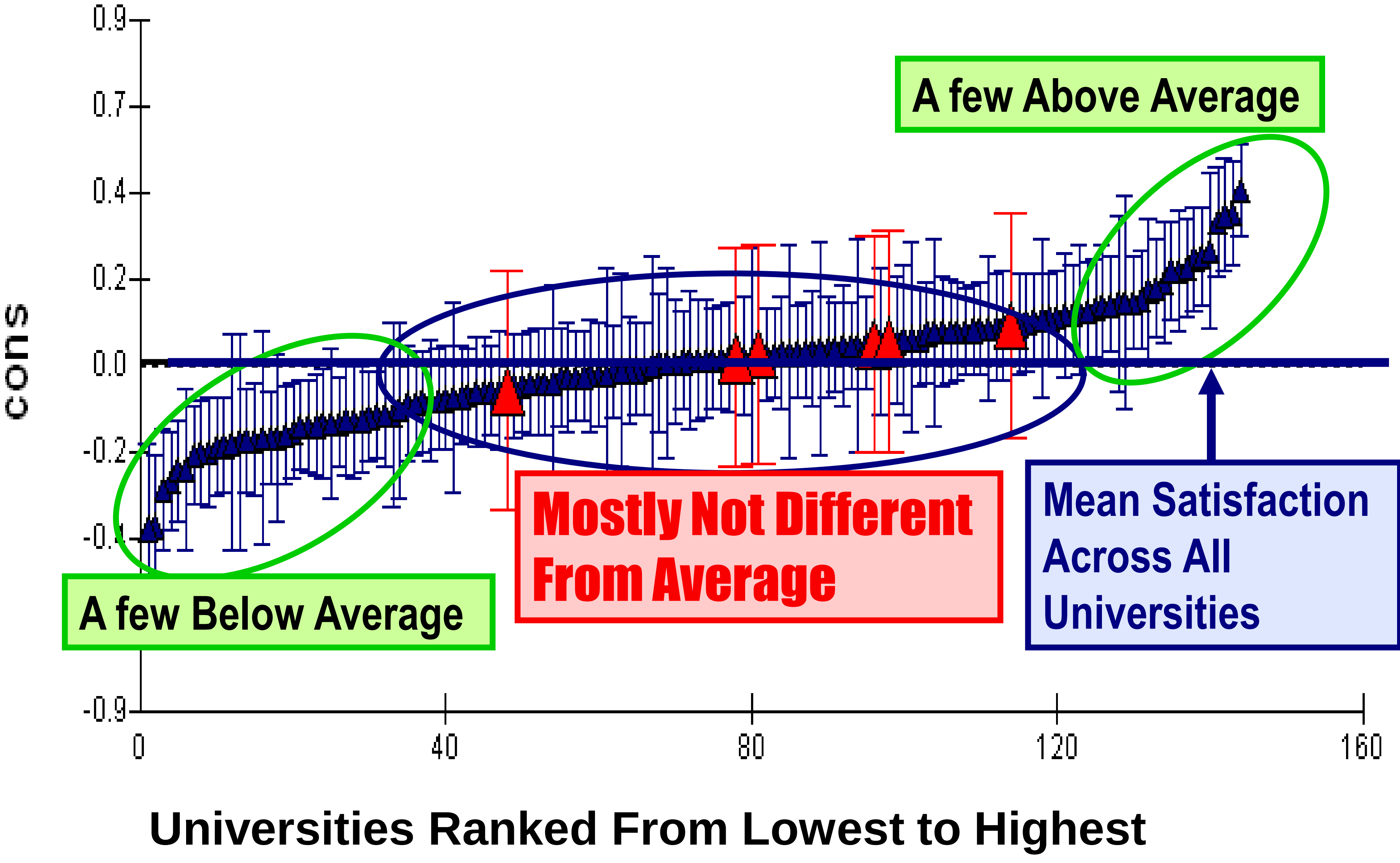
- some dubious assumptions which are not supported
- Large SEs suggest estimates not very precise
- Too little research on relations with other teaching measures
- Difficult to distinguish teacher, classroom, department & school effects; surprisingly few studies with 3- and 4-level models.
- Not useful as formative feedback to improve effectiveness as do not tell teachers what they need to do to improve.

University Student Ratings of Educational Experience: University, Course or Department as Unit of Analysis

Differences Between UK Universities: Caterpillar Plots

(170,000 Students, 141 Universities, 1500 Departments)

Cheng & Marsh (in press). National Student Survey: Are differences between universities and course reliable and meaningful. Oxford Rev Educ



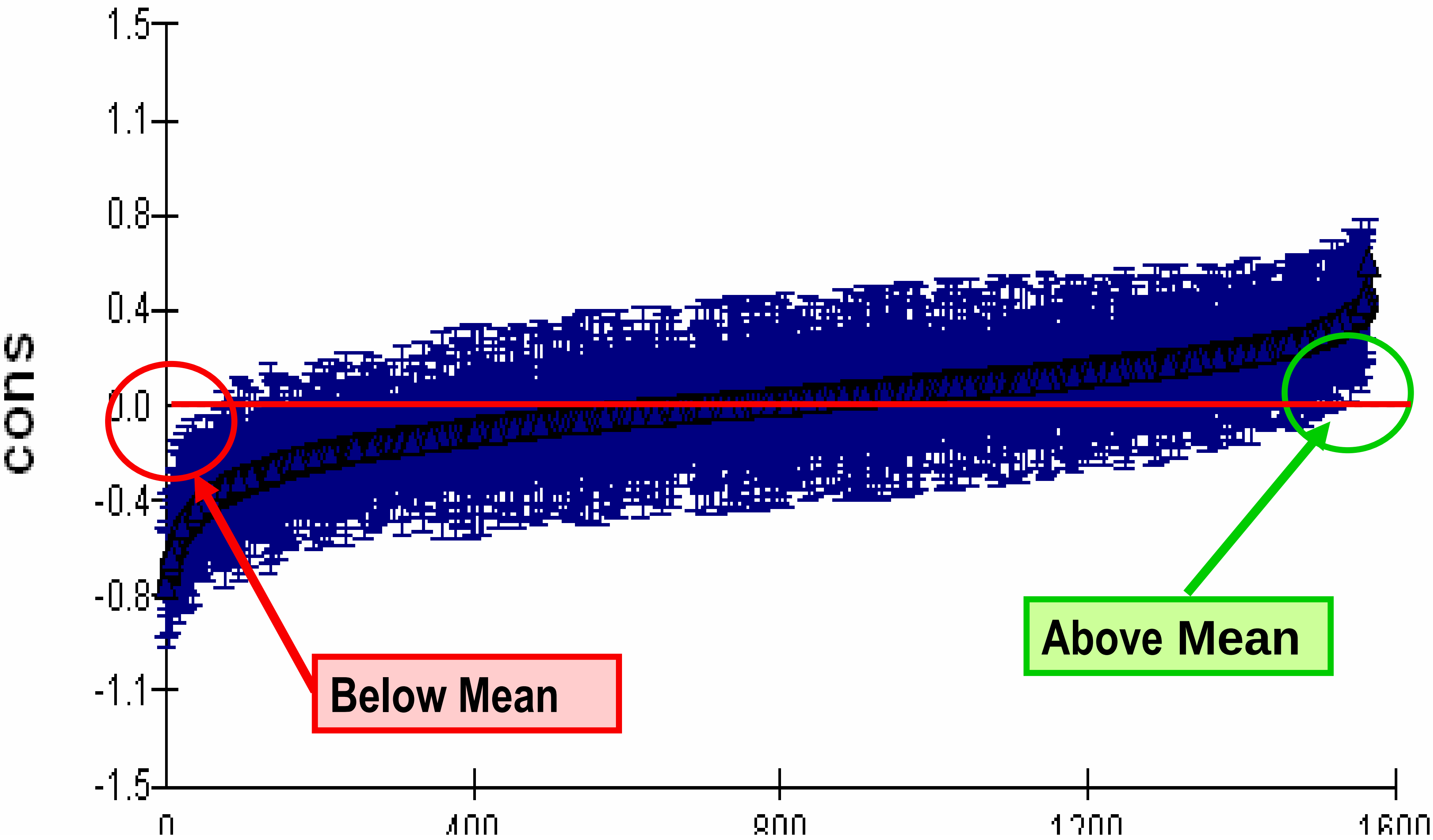
A few Above Average

A few Below Average

Mostly Not Different From Average

Mean Satisfaction Across All Universities

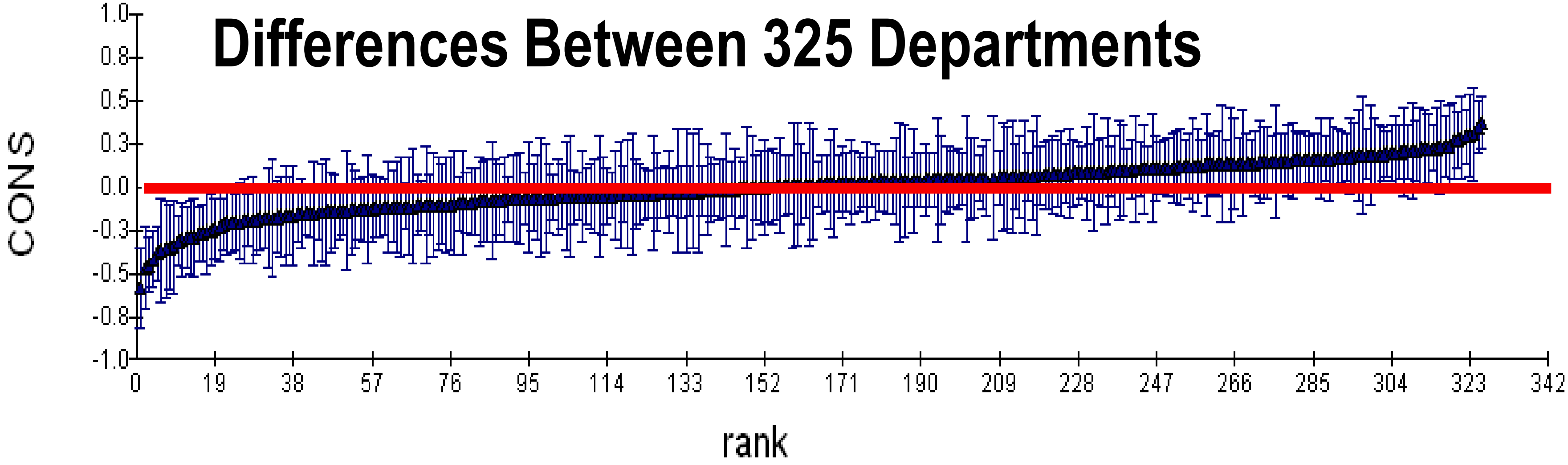
Departments/courses within UK Universities.

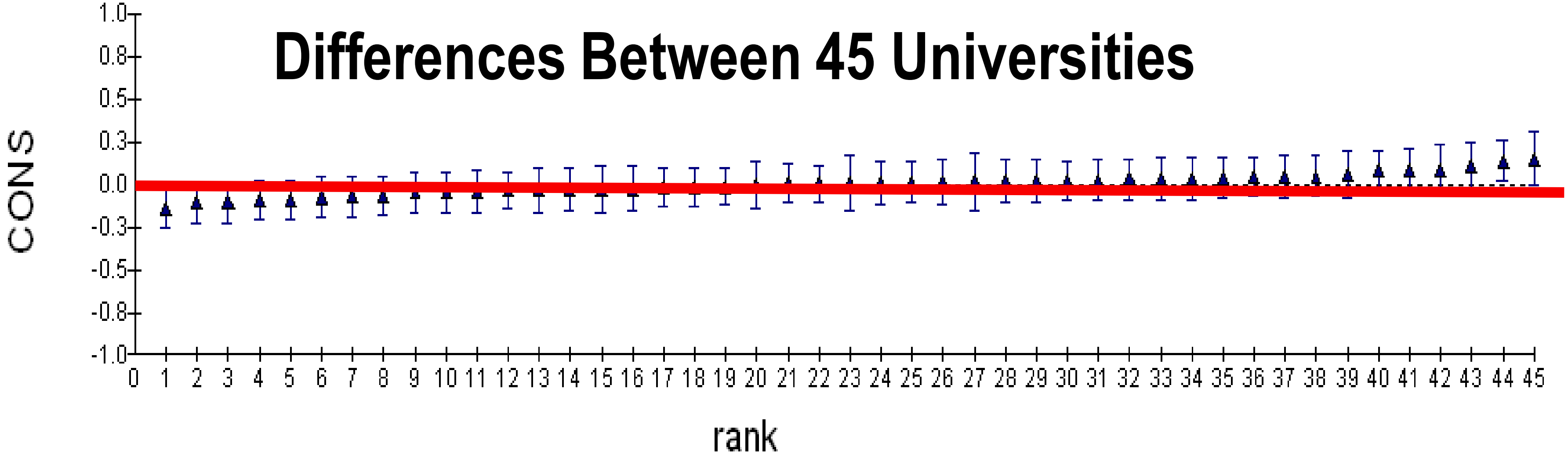


Discipline-Within-University Groups Ranked From Lowest to Highest

Cheng, J. H. S. & Marsh, H. W. (in press). National Student Survey: Are differences between universities and course reliable and meaningful. Oxford Rev Educ

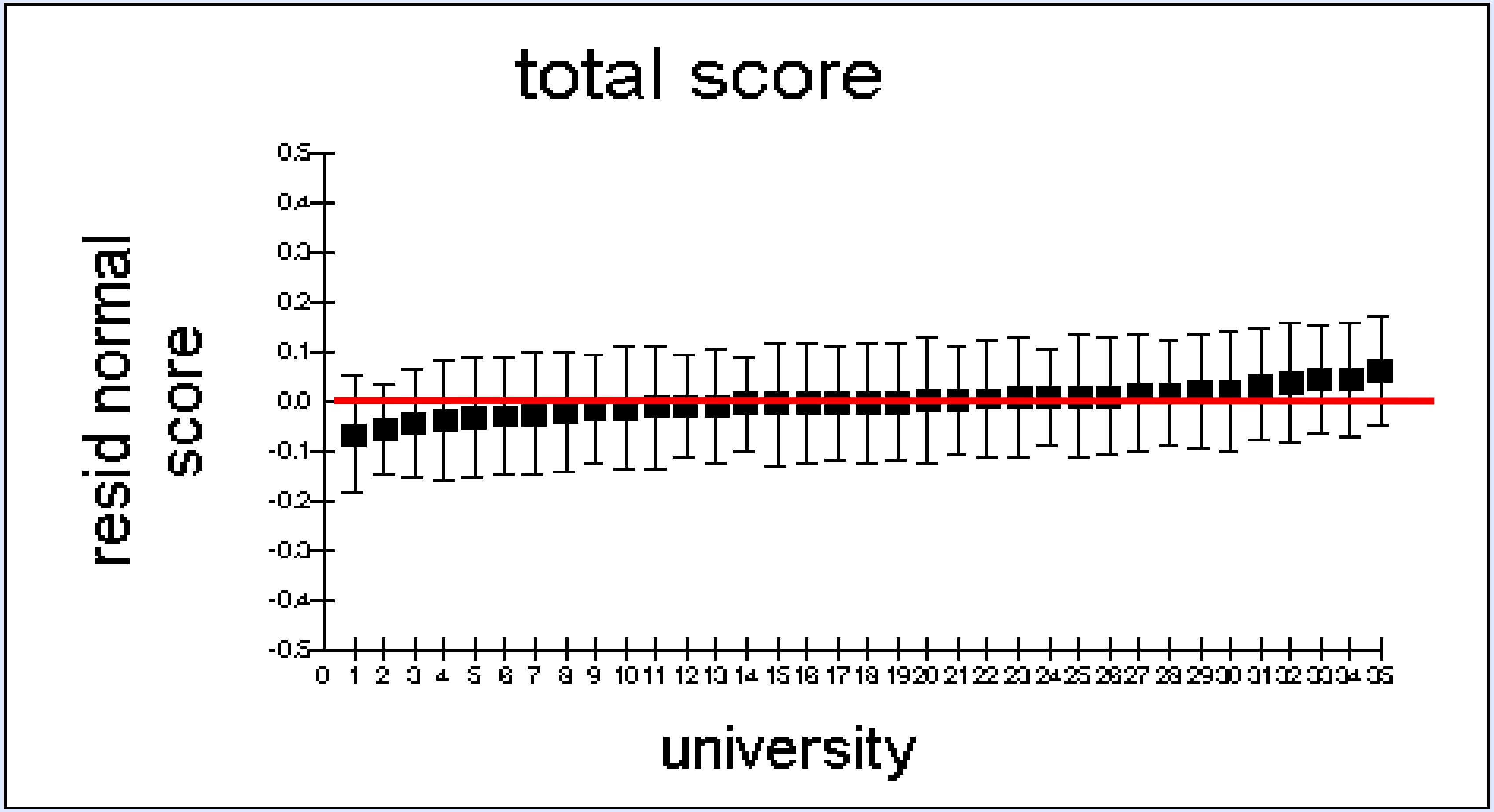
Australian CEQ Responses (44,000 students, 45 universities, 325 departments)


Differences Between 325 Departments


Differences Between 45 Universities

Marsh, h. W., Ginns, p., Morin, a. J. S., Nagengast, b., Martin, a. J. (In review). The course evaluation questionnaire (ceq): USE OF STUDENT RATINGS TO

Differences Among 35 Australian/NS Universities: Research Student Ratings of Postgraduate Research Experience (0.4% of Var Explained)



Marsh, H. W., Rowe, K., Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities. *Journal of Higher Education, 73 (3)*, 313-348.

University Student Ratings of Individual Teachers (SETS): Individual Teacher as the Unit of Analysis

SET Research Shows:

- Multidimensional well-defined, replicable factor structure;
- Reliable and stable;
- Primarily a function of the instructor who teaches a course rather than the course that is taught;
- Valid in relation to many indicators of effective teaching, including objective measures of learning;
- Relatively unaffected by a variety of variables hypothesized as potential biases;
- Seen to be useful by students for use in course selection, by administrators for use in personnel decisions, by faculty as feedback about teaching
- SET feedback + consultation Improves teaching

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp.319-384). New York: Springer.

Dimensionality: The SEEQ Factors

Learning/Value: You found course intellectually challenging/stimulating;

Instructor Enthusiasm: Instructor dynamic/energetic in conducting course;

Organisation: Course materials were well prepared/carefully explained;

Individual Rapport: Instructor was friendly towards individual students;

Group Interaction: Students encouraged to participate in class discussions;

Breadth of Coverage: Presented background/origin of ideas/concepts;

Examinations/Grading: Feedback valuable from exams/graded materials;

Assignments/Readings: Readings, homework, etc. contributed to appreciation and understanding of subject;

Workload/Difficulty: Relative course difficulty (very easy...medium…very hard).

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. Structural Equation Modeling,16, 439-476.

Relative Importance of the Teacher vs. Course Effects

How highly correlated are SETs in:

- two different courses taught by the same instructor
- same course taught by different teachers on different occasions?

For Overall Instructor Ratings of:

- same instructor teaching same course on two occasions (*r* = .72) [*teacher & course effect*],
- same instructor teaching two different courses (*r* = .61) *[teacher effect],*
- same course taught by two different instructors (*r* = -.05) [*course effect*].

SETs primarily reflect the teacher who is doing the teaching, not the course that is being taught.

Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6,* 47-59.

In Support of the Validity of SETs

SETs are positively related to many criteria of teaching effectiveness, including:

- the ratings of former students;
- student achievement in multisection validity studies;
- teacher self-evaluations of their own teaching effectiveness; and
- observations of trained observers on specific processes (e.g., teacher clarity).

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp.319-384). New York: Springer.

Multisection Validity Paradigm: Validating SETs in Relation to Student Learning

- Many sections of the same course;
- Same materials in each section (e.g., course outline, textbooks, objectives, final exam);
- Random assignment (and pre-test measures);
- SETs collected prior to final exam/course grade;
- Common final exam;

Research Question: Are SETs valid in relation to objective measures of student learning (when plausible counter explanations are not viable)?

Meta-Analysis

Cohen conducted a classic meta-analysis of multisection validity studies. Student achievement was consistently correlated with SETs:

For a subset of 41 "well-designed" studies, correlations between achievement and SETs were more substantial:

Structure (.55), Interaction (.52), Skill (.50), Overall Course (.49), Overall Instructor (.45), Learning (.39), Rapport (.32), Evaluation (.30), Feedback (.28), Interest (.15), and Difficulty (-.04).

SETs are valid in relation to student learning. Note that the multisection validity study is a value-added study with random assignment to classes.

Improving Teaching Effectiveness

Many SET Feedback studies in which:

- Teachers randomly assigned to experimental (feedback) and control (no feedback) groups;
- SETs collected; Experimental Teachers get SETs feedback;
- Groups compared subsequent SETS (and other variables).

In a meta-analysis of these studies:

- Feedback teachers .33 SD higher than control teachers
- Feedback+consultation produced much larger effects.

I developed & Tested a new Prototype Feedback/Consultation Based on My SEEQ Instrument

- Teachers randomly assigned to Feedback & control Groups;
- Using SEEQ, all teachers evaluated themselves and were evaluated by their students in two consecutive terms
- Intervention: Feedback Teachers selected target SEEQ factors that were the focus of their intervention.
- Teachers were given a book of strategies for their selected factor
- Teacher (with consultant) selected a few strategies for implementation as their intervention.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30,* 217-251.

Marsh, H. W., & Roche, L. A (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52,* 1187-1197.

Results/Discussion

SEEQ feedback and the feedback/consultation provided an effective means of improving university teaching;

- ●Feedback Teachers rated .5 SD higher than control teachers;
- ●Differences much larger for targeted SEEQ factors;
- ● Effects stronger for the initially less effective teachers;
- ●Teaching Books important: f teachers need concrete strategies to facilitate teaching improvement efforts.

However, few universities implement teaching improvement programmes as part of the collection of SETs even though clear evidence that they work. We are planning large-scale trial of this intervention with UK HEA with all UK universities.

Longitudinal Stability over 13 Years
What Happens with no Intervention?

Cross-sectional studies at different levels of education suggest that teaching effectiveness declines with experience/age.

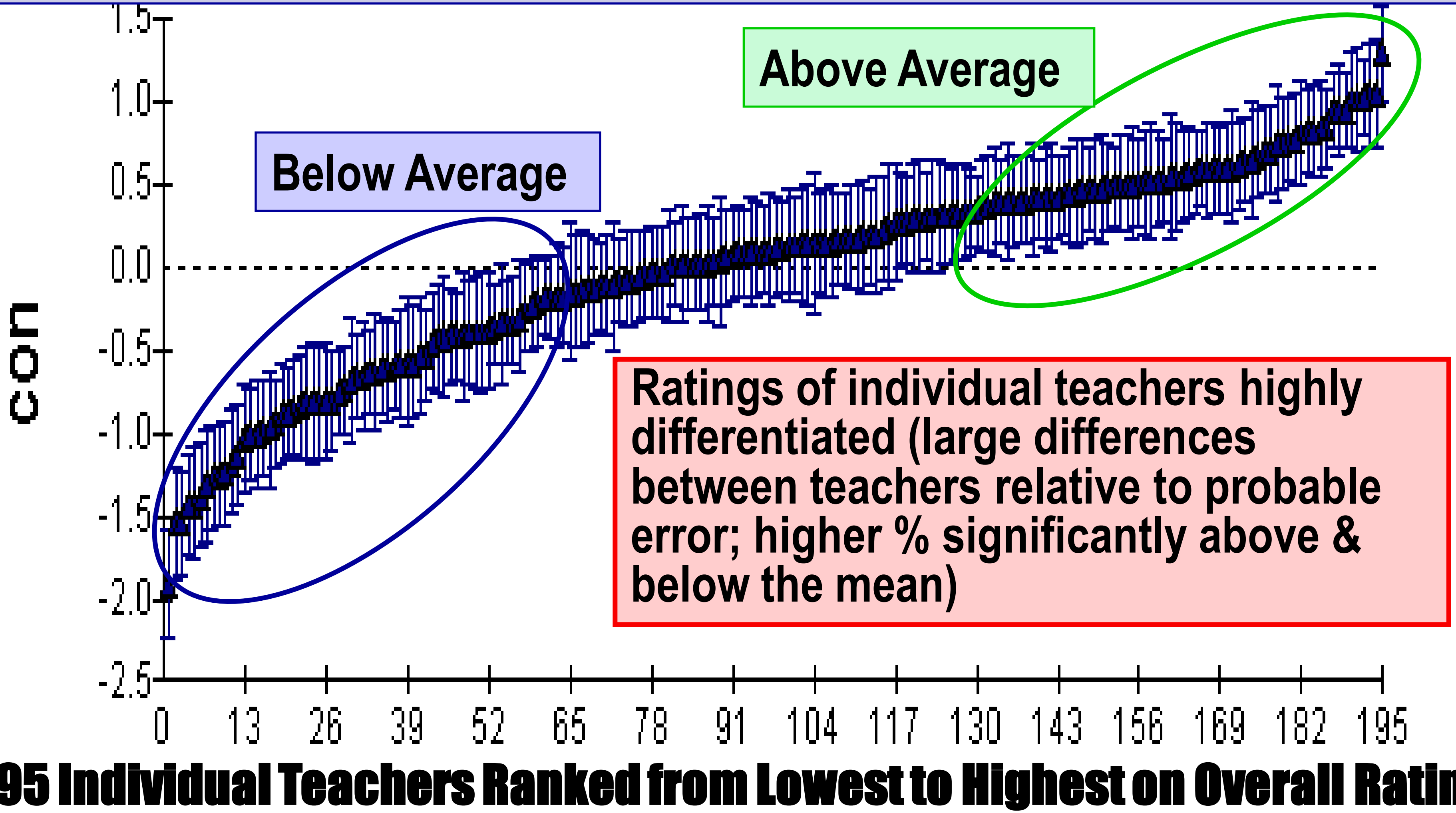
In a true longitudinal study I considered 195 teachers evaluated continuously over 13 years (average of 30.9 classes/ teacher).

- I evaluated the linear and nonlinear effects of year, course level (graduate vs. undergraduate), and their interaction.
- Changes in ratings over time were all close to zero for the 9 SEEQ factors and the two overall rating items.

mean ratings of same teachers are VERY stable over 13 years.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology,* 99, 775-790.

For Purposes of Comparison consider data from earlier SEEQ Longitudinal study of 195 Different Teachers (Consistency across an average of 31 Classes per teacher over 13 Years)



Ratings of individual teachers highly differentiated (large differences between teachers relative to probable error; higher % significantly above & below the mean)

Marsh, h. W., Ginns, p., Morin, a. J. S., Nagengast, b., Martin, a. J. (In review). The course evaluation questionnaire (ceq): USE OF STUDENT RATINGS TO BENCHMAR AUSTRALIAN UNIVERSITIES.

Summary of University SETs

- SETs weak in differentiating universities and departments
- SETs are strong in differentiating between teachers.
- Much research shows SETs based on the teacher as the unit of analysis are:
  - Multidimensional;
  - Reliable & Stable over time;
  - a function of the teacher, not the class or course;
  - valid in relation to a variety of criteria;
  - relatively unaffected by potential biases;
  - seen to be useful by stakeholders;
  - lead to improved teaching when coupled with appropriate consultation intervention

Summary

Juxtaposition Between School & University Studies

- Both show that there is not much variance at the institutional level (school or university), but much more at the teacher level.
- Particularly the CVA school research is narrowly focused on a single (unidemsional) outcome with little focus on construct validation (other than, perhaps, potential biases).
- SET university research incorporates a broad perspective to construct validation: multidimensional factor structure, reliability/stability, relations with multiple criteria of effect teaching, interventions to improve teaching.
- CVA school estimates are not very stable over 5-7 years, but SET ratings of university teachers stable over 13 years.

Both areas of research could learn from the other

To improve educational effectiveness

- reinforce good teachers through recognition – promotion & monetary rewards, social recognition. This requires a good assessment procedure so you know who to reward.
- Provide assistance to weak teachers to improve their teaching and reinforce involvement in interventions with a combination of rewards and disincentives
- To improve teaching effectiveness, teachers need evidence about their teaching effectiveness and, particularly weak teachers, need interventions, consultation, and clear strategies to improve their teaching.

To reinforce this, all applications for promotions, tenure, etc. should have some reliable and valid assessments of teaching effectiveness as part of the application.

Thank You