9º Seminario Internacional de Investigación
sobre la calidad de la educación
DOCENTES | BOGOTÁ D.C.

GOBIERNO DE COLOMBIA
MINEDUCACIÓN
icfes mejor saber

How to evaluate teachers in a ‘fair’ way?

Kristof De Witte

University of Leuven, Belgium
Maastricht University, the Netherlands

November 2, 2018

Belgium versus Colombia

→ We share taste for good drinks!

Belgium versus Colombia

→ We share losing against England in World Cup 2018

Recall from yesterday

→Attracting and rewarding effective teachers is important

"Monetary incentives are effective" (Paul Glewwe)

"but not for everybody (e.g. double wage in Indonesia)" (Karhik Mualidharan)

→ Question remains: who are the best teachers?

"There is no final test to assess teachers" (Maria Paulina)

Recall from yesterday

→Colombian principles for teacher evaluation (Laura Barragan)

- Multidimensional
- Reflection
- Autonomy
- Transparent
- Depends on class room

Recall from yesterday

→Colombian principles for teacher evaluation (Laura Barragan)

multidimensional

Reflection

Autonomy

Transparent

Depends on class room

This presentation:
Develops a technique to assess the quality of a teacher by using the students' evaluations of a teacher.

The technique meets the Colombian principles

Introduction

Students' evaluations of teaching are increasingly used to evaluate teaching performance

→ e.g. Portugal, Flanders, US, etc.

However, they are still controversial

i.e., they are 'unfair' as they do not control for impact of factors which are outside the teacher's control

- Academic research shows that background characteristics have an effect

- Practical experience of teachers indicates that some environments are more constructive to high quality teaching

Introduction

”Any system of faculty evaluation needs to be concerned about fairness, which often translates into a concern about comparability. Using the same evaluation system fore everyone almost guarantees that it will be unfair to everyone.”

(Emery *et al.*, 2003, p. 44)

How to construct SET (Students’ evaluations of teaching) scores in a fair way?

Introduction

Common construction of SET scores:

→ Step 1: Compute SET scores by the arithmetic mean of the questionnaire items (as such, without accounting for the exogenous environment)

→ Step 2: Determine impact of background characteristics on SET scores

(often by a correlation analysis, regression, multi-level model)

→ Step 3: Adjust SET scores for background characteristics

Introduction

Problem with traditional way of measuring SET:

1. Computation of SET scores in first step:

   Implies often that all teaching aspects are weighted equally

   ↔ Teachers value aspects differently

   ↔ No consensus on how teachings aspects interrelate

   ↔ Using fixed weights is subjective

   ↔ Creates unfairness (and thus disillusioned teachers)

2. Separability assumption in step 2 and 3

   Assumes that there is no direct link between SET scores and teaching environment

Introduction

How do we weight the underlying dimensions?

1. Any predetermined common set of weights will favor some teachers while harming others -> **Unfairness**
2. In the absence of a consensus on how teaching aspects exactly interrelate, any choice of fixed weights will be to some extent **subjective**.
3. The choice of weights may affect the teachers' evaluation score and ranks **undermining** their **credibility**.

- *"There is **no blue print** for being an effective teacher"* (Fraser, 2000 p. 3).
- *"We know what the characteristics of good teaching are, but **we don't know how they relate** to each other"* Weimer (1990, p. 13)

Alternative way

Idea:

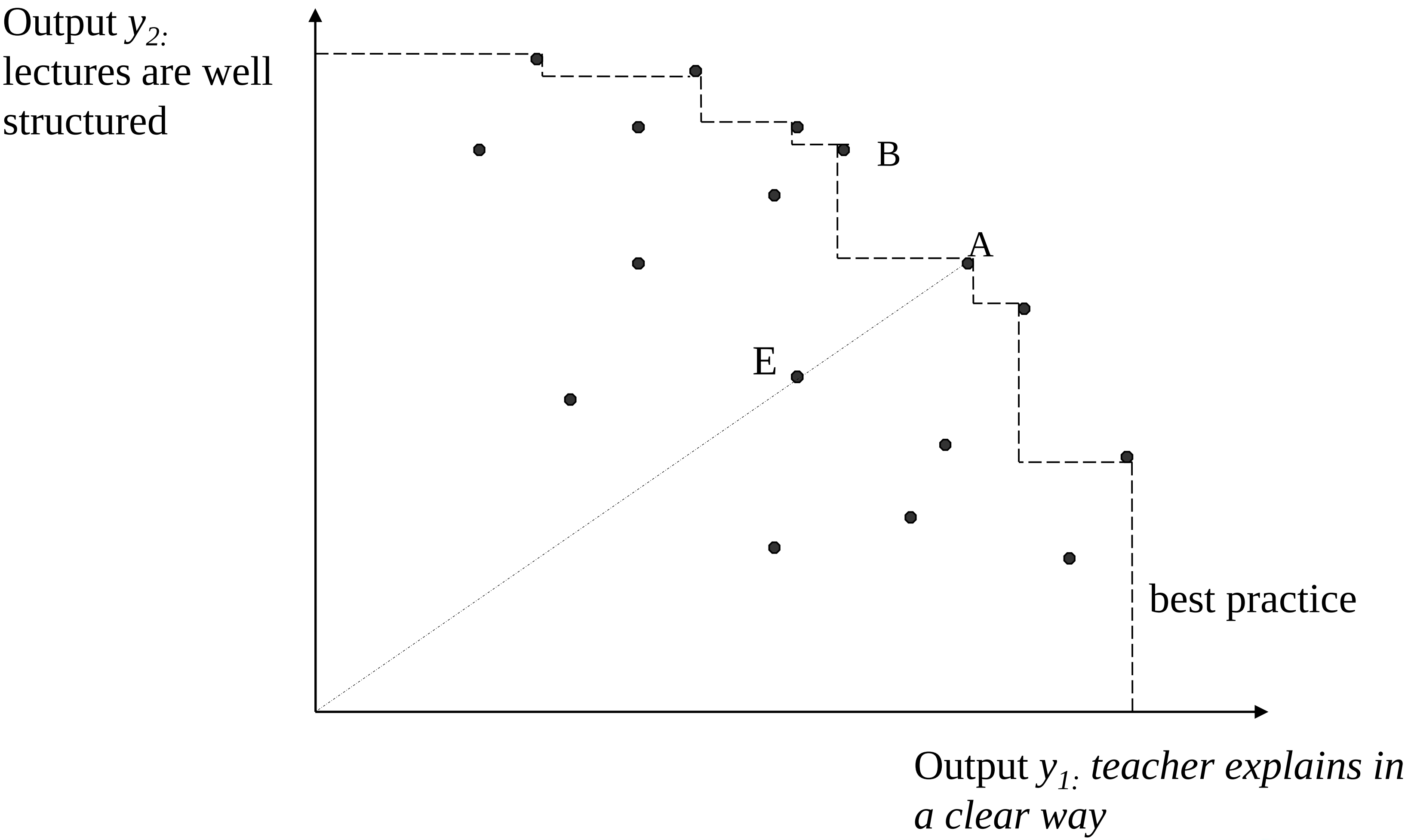
Start from the best performing teachers

and compare the performance to these best teachers

Alternative way

Benefit of doubt model (BoD)

e.g. The model graphically for two dimensions:


Output $y_{2:}$ lectures are well structured
B
A
E
best practice
Output $y_{1:}$ *teacher explains in a clear way*

Alternative way

”Benefit of doubt model” (BoD)

This approach is convenient because the algebraic expression behind this graph determines the weights endogenously

i.e. The ratio of the performance of the evaluated teacher

to the performance of the best teacher

$$SET_c = \max_{w_{c,i}} \frac{\sum_{i=1}^{q} w_{c,i} y_{c,i}}{\max\limits_{y_{j,i}\in\{evaluated\ lectures\}} \sum_{i=1}^{q} w_{c,i} y_{j,i}}$$

Alternative way

⇨ Where are we now?

- Construct SET scores based on single-dimensional performance indicators

- We have no *a priori* understanding of the importance of these indicators

⇩

The model:

- Put for each questionnaire item *i*, the performance of a teacher on his/her course *c* (i.e., $y_{c,i}$) in a relative perspective to the other performances $y_{j,i}$

→ A good relative performance: higher weight for this item

→ A low relative performance: lower weight for this item

Thus: optimal weights which maximise the teacher's SET

Alternative way

Disadvantage of BoD:

✓ It may allow a teacher to appear as a brilliant performer in a manner that is hard to justify (e.g. **zero weights** or **weights contradicting prior views**).

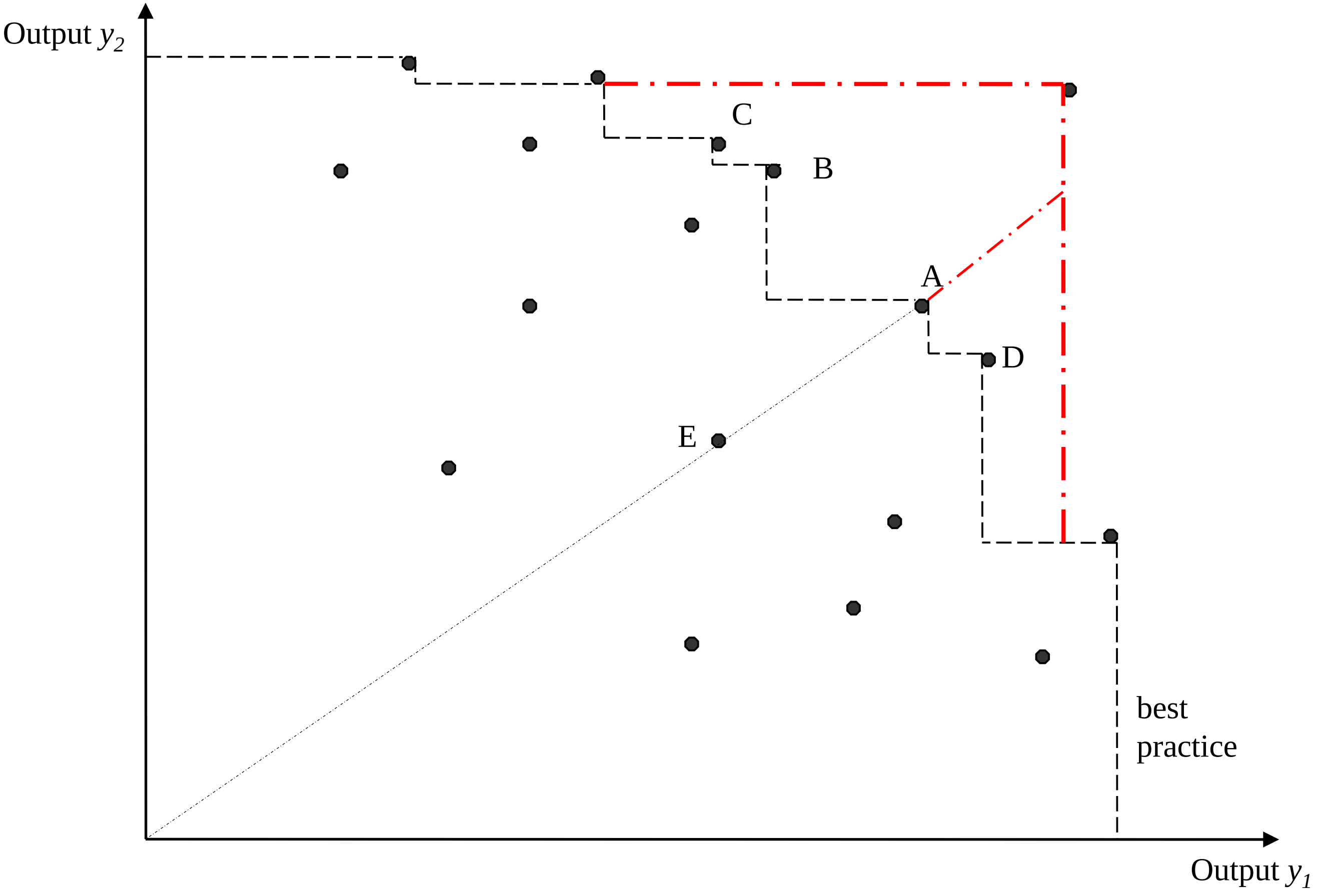
Solution:

Take into account **expert/stakeholder (e.g. students, lecturers, etc.) opinion**, while recognizing that agreement in a unique and fixed weighting scheme is the exception rather than the rule.

Alternative way

Benefit of doubt model (BoD)

e.g. The model graphically for two dimensions:



Alternative way

➡ The robust Benefit of doubt model (robust BoD)

➡ Reasons:

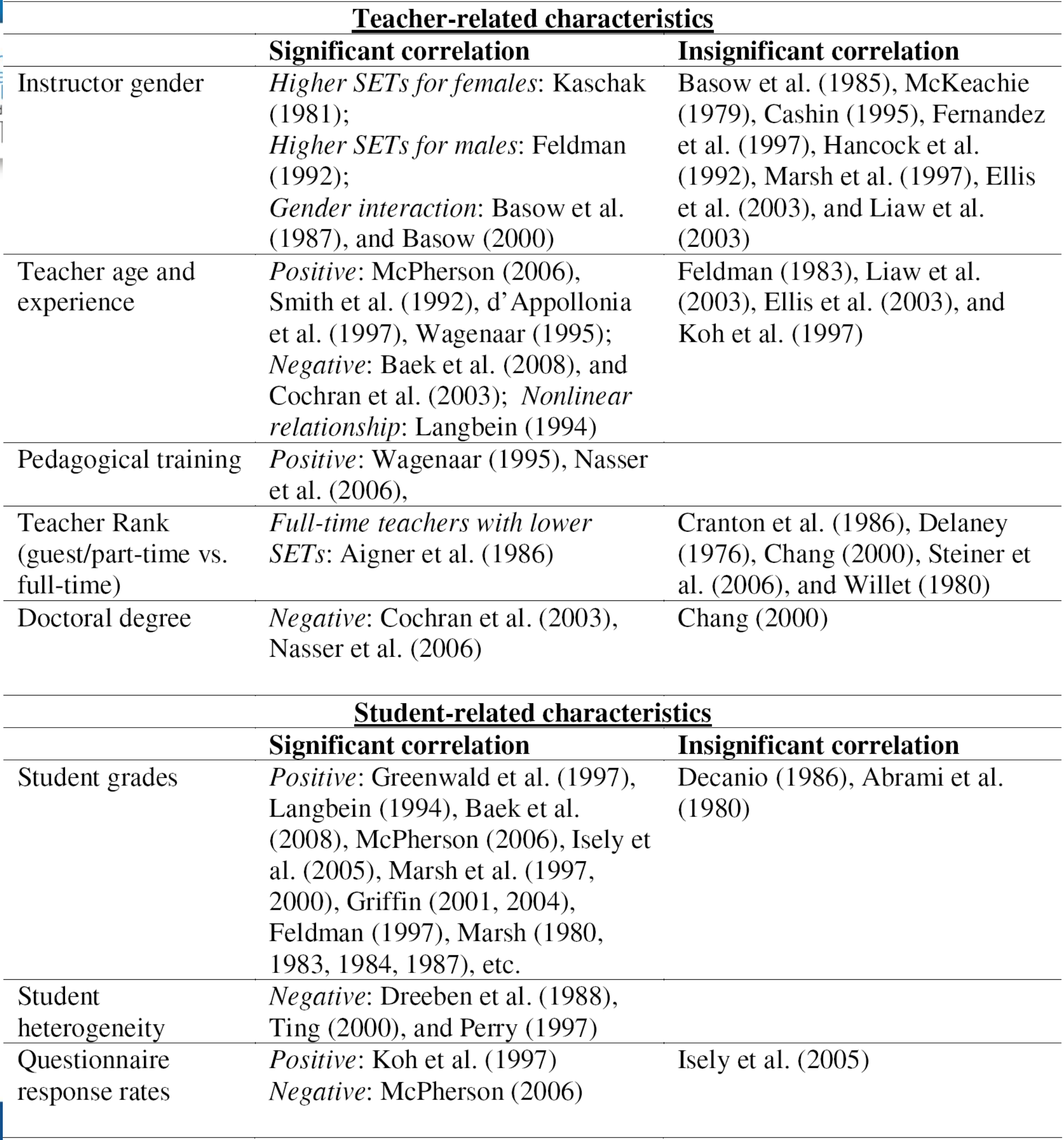
1. Allow for outlying observations (e.g., from measurement error)
2. Statistical inference

➡ Implementation

Robust efficiency scores of Cazals *et al.*, 2002

➡ Idea:

- Draw repeatedly and with replacement *m* observations from the original sample of *n* observations
- Estimate relative to this smaller reference set of size *m* the BoD model
- Take the arthemitic average of the *B* SET scores:



Teacher-related characteristics

	Significant correlation	Insignificant correlation

Instructor gender	*Higher SETs for females*: Kaschak (1981);
	Higher SETs for males: Feldman (1992);
	Gender interaction: Basow et al. (1987), and Basow (2000)
		Basow et al. (1985), McKeachie (1979), Cashin (1995), Fernandez et al. (1997), Hancock et al. (1992), Marsh et al. (1997), Ellis et al. (2003), and Liaw et al. (2003)

Teacher age and experience	*Positive*: McPherson (2006), Smith et al. (1992), d'Appollonia et al. (1997), Wagenaar (1995); *Negative*: Baek et al. (2008), and Cochran et al. (2003); *Nonlinear relationship*: Langbein (1994)	Feldman (1983), Liaw et al. (2003), Ellis et al. (2003), and Koh et al. (1997)
Pedagogical training	*Positive*: Wagenaar (1995), Nasser et al. (2006),	
Teacher Rank (guest/part-time vs. full-time)	*Full-time teachers with lower SETs*: Aigner et al. (1986)	Cranton et al. (1986), Delaney (1976), Chang (2000), Steiner et al. (2006), and Willet (1980)
Doctoral degree	*Negative*: Cochran et al. (2003), Nasser et al. (2006)	Chang (2000)

Student-related characteristics

	Significant correlation	Insignificant correlation
Student grades	*Positive*: Greenwald et al. (1997), Langbein (1994), Baek et al. (2008), McPherson (2006), Isely et al. (2005), Marsh et al. (1997, 2000), Griffin (2001, 2004), Feldman (1997), Marsh (1980, 1983, 1984, 1987), etc.	Decanio (1986), Abrami et al. (1980)
Student heterogeneity	*Negative*: Dreeben et al. (1988), Ting (2000), and Perry (1997)	
Questionnaire response rates	*Positive*: Koh et al. (1997) *Negative*: McPherson (2006)	Isely et al. (2005)

Alternative way

➡ The robust and conditional Benefit of doubt model

➡ Reasons:

1. Incorporate background characteristics in the BoD model
2. Compare ’like with likes’
3. Does not assume a separability assumption
4. Statistical inference on impact of characteristics

➡ Implementation

Conditional efficiency estimates for mixed (i.e., continuous and discrete) exogenous variables of De Witte and Kortelainen (2008)

➡ Idea:

- Draw repeatedly and with replacement *m* observations from the original sample of *n* observations, and draw with a probability that $z_{c,r} \approx Z$

Data

⇨ Questionnaire setup:

16 questionnaire statements were asked to 5,513 students

→ 112 college courses by 69 teachers

→ Commercial Sciences at University College Brussels (Belgium)

→ Year: 2006-2007

⇨

Questionnaire dimensions:

→ Questions are rated on a Likert scale from 1 (disagree) to 5 (agree).

→ The questions are grouped by the university coordination in 4 dimensions:

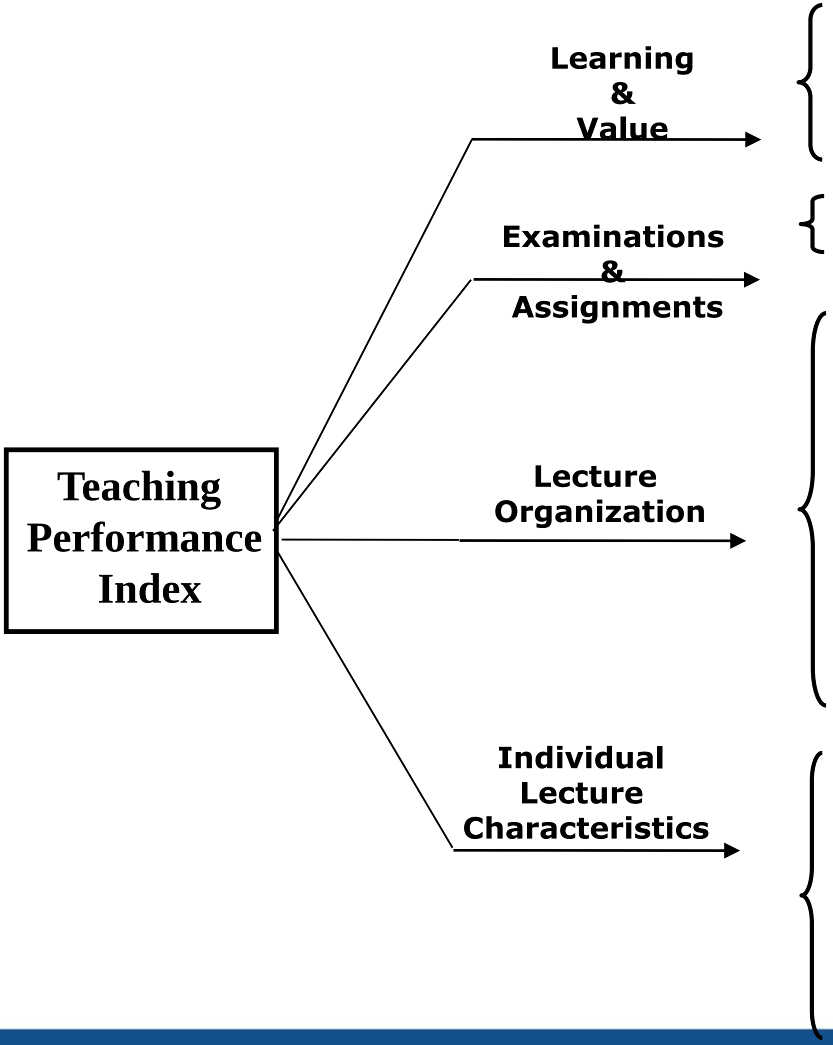
1. Learning and Value
2. Examinations and Assignments
3. Lecture Organisation
4. Individual Lecturer report

→ Relate to background characteristics

Data

16 QUESTIONNAIRE ITEMS

4 KEY DIMENSIONS



Teaching Performance Index

Learning & Value

i. The lecturer justifies this part of the schooling in function of our cultivation/formation.

ii. In this part of the schooling I have learned a lot.

iii. In general, I have a good impression of these lectures.

Examinations & Assignments

iv. The requirements and agreements concerning the exam evaluation are clear.

Lecture Organization

v. The lectures takes into account my knowledge and skills

vi. The employed lecture material (syllabus, hand book, texts, electronic documentation) is conveniently arranged and understandable.

vii. During the lectures didactical equipment is functionally used (black board, tranparents, video, computer, language practicum, laboratory,…).

viii. The lectures encourage reflecting and actively digesting the course material.

ix. The lectures are well-structured.

x. The pace of the lecture.

Individual Lecture Characteristics

xi. The lecturer reacts to questions, suggestions and critical remarks in a serene and constructive manner.

xii. The lecturer has good contacts with the students.

xiii. During the lectures one speaks sufficiently load and clear.

xiv. The lecturer treats each student with respect.

xv. The lecturer gives useful examples, applications or exercises.

xvi. The lecturer explains the course material in a good way.

Evaluating teachers

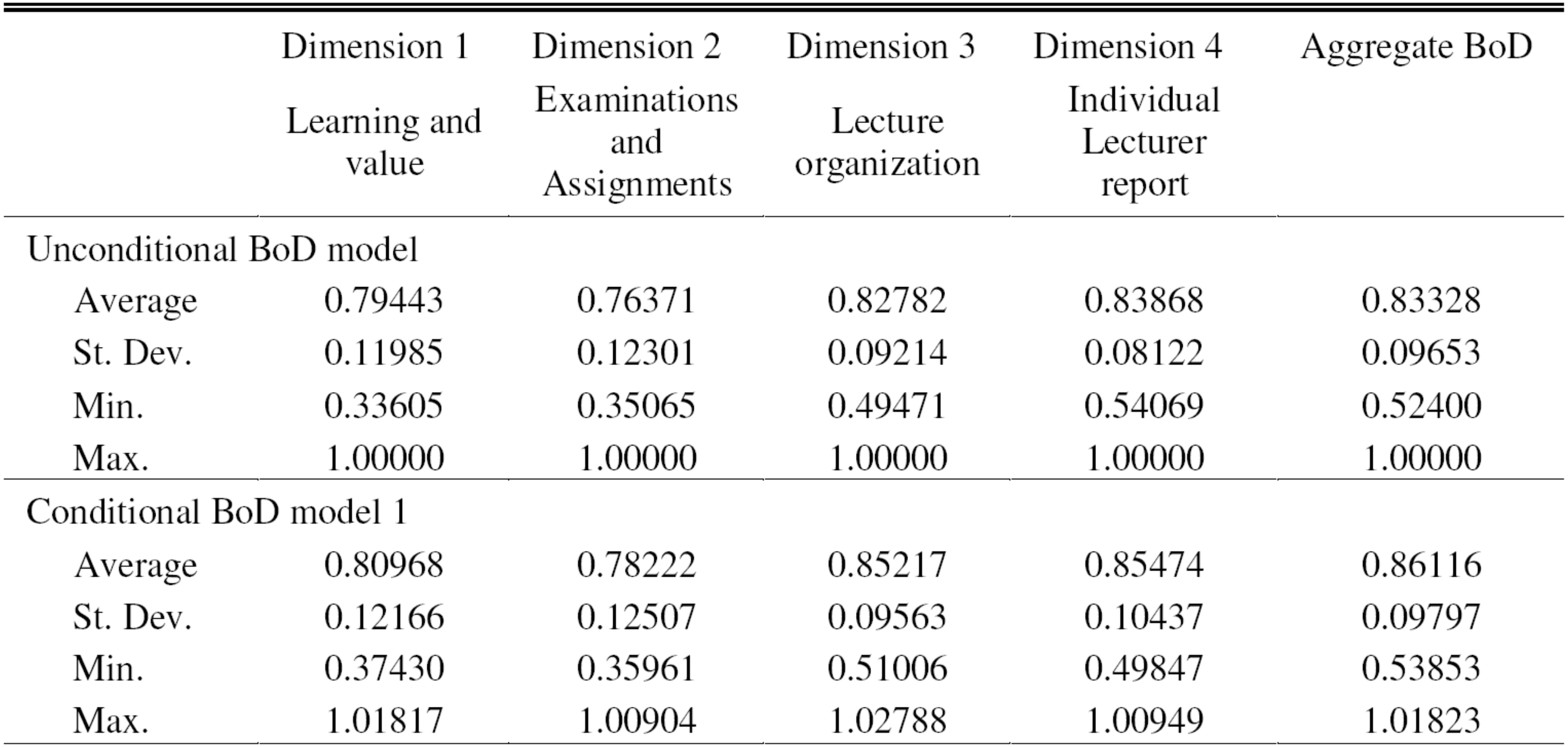
Results

Nr.	Teacher	Course	Class	Contact	EW	BoD	BoD_R	Order-m BoD_R
…	…	…	…	…	…	…	…	…
8673	Professor B	Micro Economics A	1BW [1]	45	3.650	85.50%	79.25%	82.84%
8674	Professor B	Micro Economics B	1BW [2]	30	3.697	86.19%	80.10%	83.92%
9487	Professor B	Micro Economics B	1DW [2]	30	4.101	94.81%	88.14%	92.36%
66607	**Professor C**	**Banks & Stock B**	**2JU** [1]	**16**	**3.582**	**83.31%**	**83.05%**	**86.28%**
1421	**Professor C**	**Corporate finance**	**1EW** [2]	**30**	**3.981**	**94.31%**	**73.58%**	**76.81%**
8522	**Professor C**	**Banks & Stock A**	**1BE** [1]	**30**	**3.677**	**85.09%**	**75.72%**	**78.26%**
8636	**Professor C**	**Banks & Stock A**	**1BW** [1]	**30**	**3.750**	**89.77%**	**78.02%**	**81.08%**
8911	**Professor C**	**Corporate finance**	**1EW** [1]	**30**	**3.801**	**91.79%**	**78.84%**	**82.29%**
9029	**Professor C**	**Banks & Stock B**	**1LC** [1]	**16**	**3.250**	**77.16%**	**65.99%**	**68.95%**
9157	**Professor C**	**Banks & Stock B**	**1SB** [1]	**16**	**2.944**	**76.61%**	**64.14%**	**66.96%**
8927	Professor D	Quantitative Methods	1EW [1]	30	3.508	87.60%	75.51%	74.46%
9583	Professor D	Quantitative Methods	2LB [2]	30	3.400	83.60%	75.22%	78.38%
…	…	…	…	…	…	…	…	…

[1]: academic year 2005/2006, [2]: academic year 2006/2007, EW = Equal Weighting, BoD = full flexibility Benefit of the Doubt weighting, BoD_R = Restricted Benefit of the Doubt weighting, and Order-m BoD_R = restricted and robust order-m Benefit of the Doubt weighting

Evaluating teachers

Conditional and unconditional Benefit of the Doubt model (BoD)



	Dimension 1 Learning and value	Dimension 2 Examinations and Assignments	Dimension 3 Lecture organization	Dimension 4 Individual Lecturer report	Aggregate BoD
Unconditional BoD model					
Average	0.79443	0.76371	0.82782	0.83868	0.83328
St. Dev.	0.11985	0.12301	0.09214	0.08122	0.09653
Min.	0.33605	0.35065	0.49471	0.54069	0.52400
Max.	1.00000	1.00000	1.00000	1.00000	1.00000
Conditional BoD model 1					
Average	0.80968	0.78222	0.85217	0.85474	0.86116
St. Dev.	0.12166	0.12507	0.09563	0.10437	0.09797
Min.	0.37430	0.35961	0.51006	0.49847	0.53853
Max.	1.01817	1.00904	1.02788	1.00949	1.01823

What correlates to SET?

Favorable influence

- Pedagogical training
- Class size (cfr. Selection effects – Andrea Canales)

Unfavorable influence

- Guest lecturer
- Mean grade of students
- Evening course

No significant influence

- Age
- Spread in students’ scores

Applications

➡ Potential applications in education:

- Evaluation of teaching of **university professors**
    - at HUB university (Belgium)
- Evaluation of **research** of university professors

- Evaluation of **secondary schooling teachers**
    - cf. Portugal; see OECD, 2009

      ”The teacher evaluation model involves the use of a wide array of instruments, including self-evaluation, classroom observation, interviews, student results and standardised forms to record teacher performance - this is an ambitious model, as it attempts to tap all areas of the functioning of a teacher.”

    - Pilot project in Flanders (Klasse, 2001)

    - Large literature in US: evaluation as a tool for instructional improvement → follows from the ’No Child Left Behind’ Act.

Applications

- Reward teachers according to their evaluation

- Reward institutions (e.g. schools or universities) according to their performance

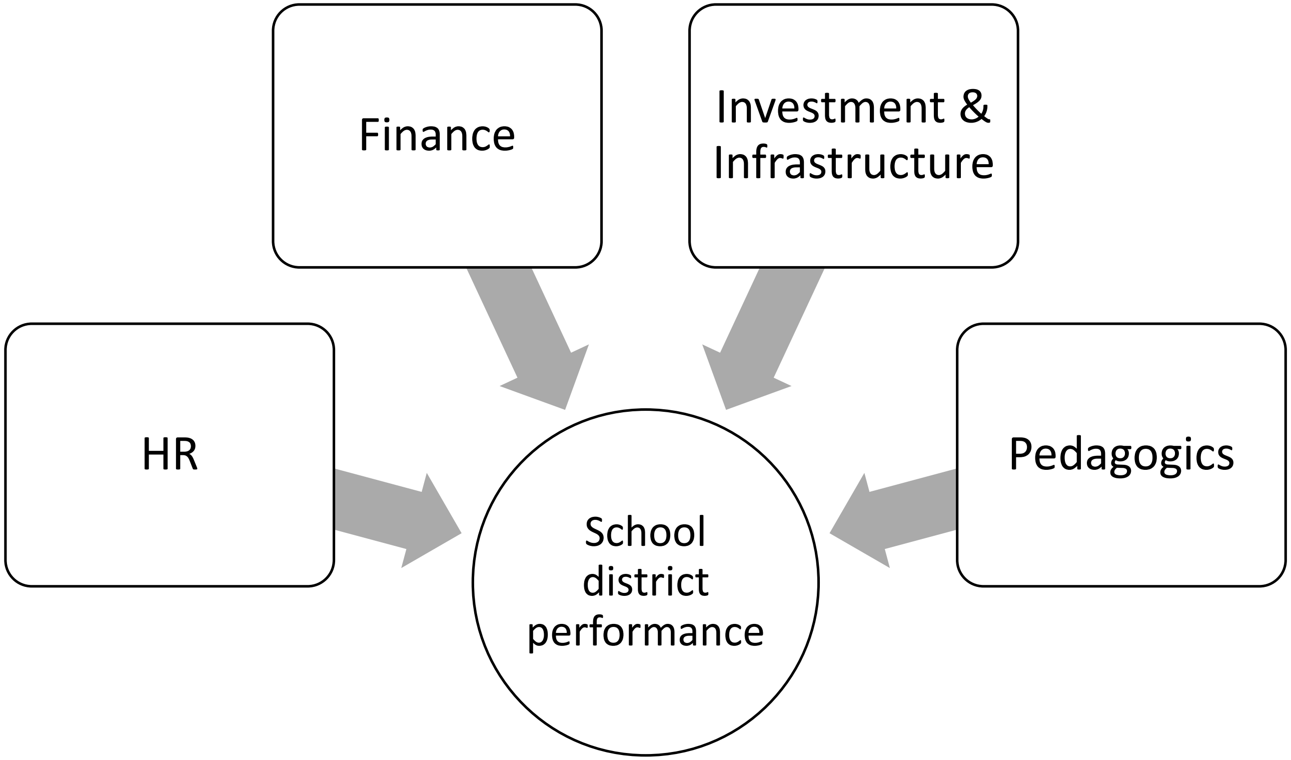
↓

Only possible if the evaluation is considered by all parties as ’fair’

i.e.: - favaroble performance score
- account for background characteristics

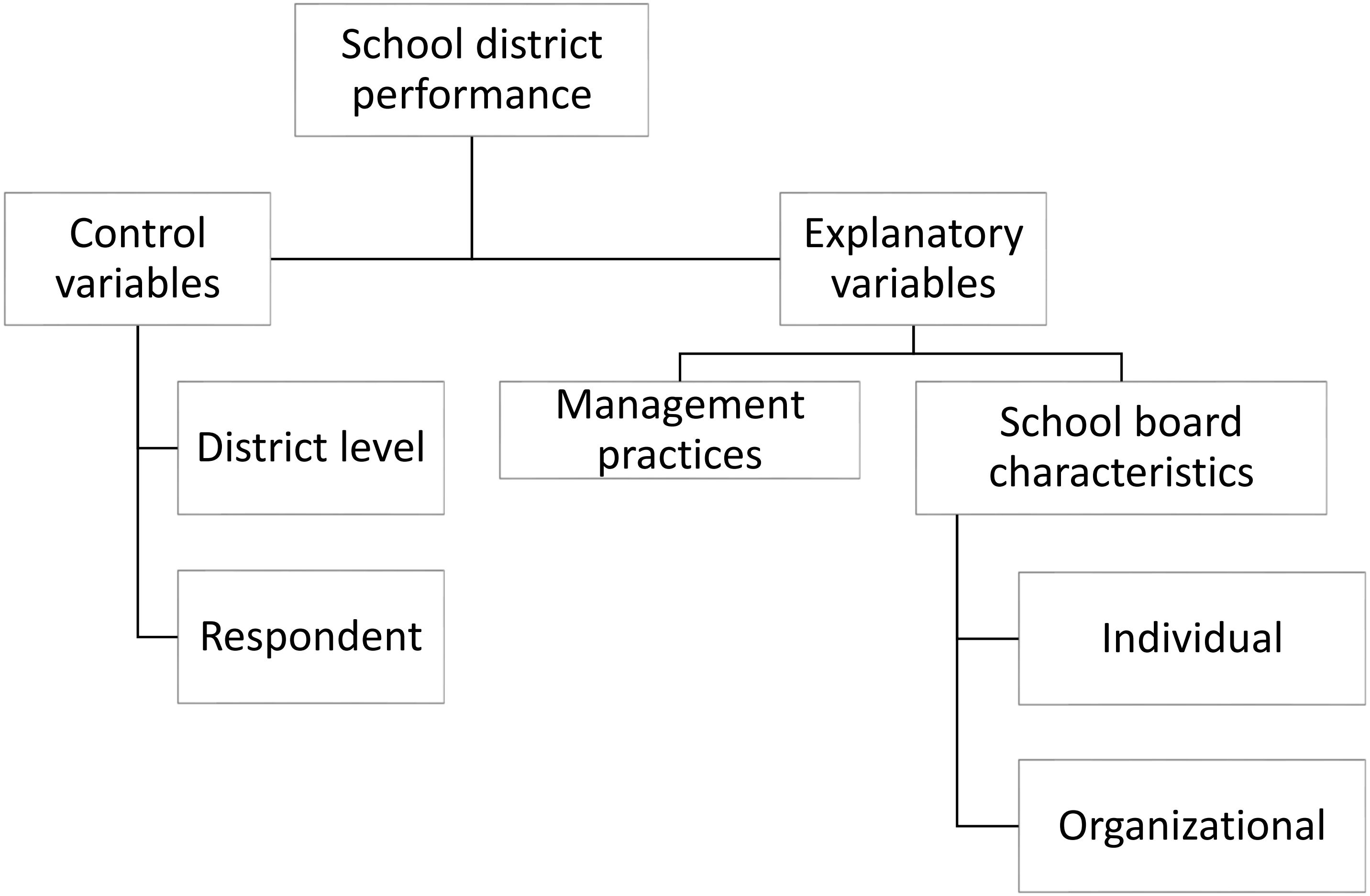
Applications

- Evaluation of school boards / school districts


Finance
Investment & Infrastructure
HR
Pedagogics
School district performance

Applications

- Evaluation of school boards / school districts



Applications

- Evaluation of school boards / school districts

What correlates to school district performance (evidence for Belgium)?

1. Higher performance in non-governmental districts (private school boards):
2. Participative management style is favorable for performance
3. Consolidation is better than cooperation among school boards
4. Expertise of the board members
5. Size doesn’t matter <> Cost efficiencies can be obtained (Schiltz & De Witte, 2016)

9º Seminario Internacional de Investigación sobre la calidad de la educación
DOCENTES | BOGOTÁ D.C.

GOBIERNO DE COLOMBIA
MINEDUCACIÓN
icfes mejor saber

How to evaluate teachers in a ‘fair’ way?

Kristof De Witte

University of Leuven, Belgium
Maastricht University, the Netherlands

November 2, 2018