University of Antwerp

On the (in)validity of students' evaluation of teaching (SET)

A state of the art of the research and suggestions for future practice

Pieter Spooren

University of Antwerp

The classroom is your stage.
Perform.

University of Antwerp

Introduction

"The opinions of those who eat the dinner should be considered if we want to know how it tastes"

(Seldin, 1993,40)

In contemporary higher education, students' evaluation of teaching (SET) is the dominant (and sometimes sole) indicator of teaching quality (Onwuegbuzie et al., 2007; Zabaleta 2007)

SET is of all time

- Ancient History (e.g., the death of Socrates) (Marsh, 1987)
- Medieval universities (Ong, 1958; Knapper, 2001)
- 1920's: the first systematic and standardized SET-procedures in North American and Canadian universities (Kulik, 2001)
- 1970's: SET became a norm at many universities (responding to student activism) (McKeachie, 1996) and became prevalent in personnel decisions (Galbraith, Merrill & Kline, 2012)
- 1980's: introduction of SET in European universities

Introduction

"What's important to recognize is that such judgments will be made even in the absence of good data"

(Knapper, 2001, 3)

Whereas SET in the early days mainly had a formative character, its purpose in contemporary higher education is threefold (Kember, Leung & Kwan, 2002):

1. Improving course and teaching quality

   Formative: using student experience to improve the quality of (teaching in) a course

2. Administrative decision-making

   Summative: tenure / promotion decisions

3. Demonstrating institutional accountability

   Summative: demonstrating adequate procedures for ensuring teaching quality (to prove an institution's performance in accounting and auditing practices)

Introduction

"Making visible the invisible: evaluating teaching"

(Blackmore, 2009, 864)

The double use of SET, and the unresolved tension between them, make SET very delicate (Penny, 2003):

- On the one hand, many teachers are convinced of the usefulness of SET as an instrument to improve their teaching

- On the other hand, it is argued that nowadays the principal focus lies in mapping teaching competence for summative purposes:
  - The 'managerial approach' in higher education (accountability, visibility, transparancy) => 'institutionalization' of SET
  - Teaching performance = the extent to which student expectations are met ('consumer satisfaction')?
  - Students' 'opinions' = 'knowledge'?

Introduction

"Instructors must avoid undermining the rating process by administering the forms with introductions such as 'Here are those silly forms again.'"

(Ory & Ryan, 2001, 41)

As a result, the validity and reliability of SET scores have been questioned for many years:

- Are students capable of providing appropriate teacher evaluations?
- Are there any differences between teachers' views and students' views concerning 'good teaching'?
- SET-surveys are 'happy forms' or 'personality contests'
- SET are influenced by factors that are unrelated to good teaching
- Many SET-questionnaires are poorly designed
- Standardized SET-questionnaires depersonalize the (individual) relationship between a teacher and the students
- Interpreting SET-results is far more complicated as it looks (i.e., risk of inappropriate use by both teachers and SET-administrators)
- Teachers are not aware of the enormous amount of SET-research, which invalidated some persistent myths concerning SET

Introduction

"Opinions about the role of students' evaluations vary from 'reliable, valid and useful' to 'unreliable, invalid and useless'. How can opinions vary so drastically in an area which has been the subject of thousands of studies?"

(Marsh, 1984, 708)

Given these concerns it is not surprising that many teachers fear their next SET-report

- The 'tyranny of the evaluation form' might lead to practices aimed at increasing SET scores rather than improving instruction, grading leniency and grade inflation (Eiszler, 2002; Oleinik, 2009)
- At the same time, many valuable thoughts and suggestions from the student's perspective remain untouched (Simpson & Siguaw, 2000)

Introduction

"One might suppose that the research studies on ratings are similar to many other studies in education: conflicting, confusing and inconclusive"

(Kulik, 2001, 10)

When looking at the literature, it is obvious that most SET-research has been done in Anglo-Saxon educational settings (US, GB, Australia) (due to the long tradition of SET in these contexts)

SET is by far the most studied measure of teaching effectiveness in higher education ('golden age' of SET research in the 70s and 80s)

Many recent reviews are available (Marsh, 2007; Onwuegbuzie et al., 2009; Spooren, Brockx & Mortelmans, 2013)

Overview

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. A state of the art. *Review of Educational Research, 83*(4), 598-642. doi: 10.3102/0034654313496870

- Overview of the SET research published in internationally peer-reviewed journals since 2000
- Using Onwuegbuzie's meta-validation model for assessing the score-validity of SET(Onwuegbuzie, Daniel & Collins, 2009)
- General conclusions:
  - SET remain a hot, yet delicate topic in higher education
  - Many stakeholders are not convinced of the validity of SET
  - SET research did not succeed in providing clear and unambiguous answers to several critical aspects concerning SET
  - Although an important and valuable source of data, SET alone is not enough for the assessment of teaching

Overview

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. A state of the art. *Review of Educational Research, 83*(4), 598-642. doi: 10.3102/0034654313496870

This lecture

6 topics which are crucial for the use and validity of SET

1. What is 'effective teaching'?
2. SET instruments
3. The relationship between SET scores and factors unrelated to good teaching
4. SET procedures and administration
5. SET and the internet
6. SET and the improvement of teaching

1. What is 'effective teaching'?

"Student ratings can be no more valid than the instrument to collect the information"

(Penny, 2003, 401)

- A clear understanding of effective teaching is a prerequisite for the construction of SET instruments
- There is no universally accepted definition of effective university teaching (Devlin & Samarawickrema, 2010)
- Although there exists some consensus between educational scientists concerning the characteristics of the 'effective teacher' (i.e. subject knowledge, course organization, helpfulness, enthousiasm, feedback, interaction with students, etc.), there exists *great variety* in the dimensions that are captured in existing SET instruments

1. What is ‘effective teaching’?

Instrument	Author	N° dimensions	Dimensions
SEEQ	Marsh	9	Learning/value, Instructor enthousiasm, Organization/clarity, Group interaction, Individual rapport, Breadth, Exam/graded materials, Readings/assignments, Workload difficulty
ECTQ	Kember & Leung	9	Understanding fundamental content, Relevance, Challenging beliefs, Active learning, Teacher-student relationships, Motivation, Organization, Flexibility, Assessment
SPTE	Burdsal & Bardo; Jackson et al.	6 (2)	General Quality of Teaching: Rapport with students, Course value, Course organization & design, Fairness of grading Course Demands: Course difficulty, Course Workload
TBC	Keeley, Smith & Buskist	2	Caring and supportive Professional competency and Communicational skills
SETERS	Toland & De Ayala	3	Delivery of Course Information Facilitating Instructor/Student Interactions Regulating Students’ Learning

1. What is ‘effective teaching’?

“It is fair to say that many of the forms used today have been developed from other existing forms without much thought to theory or construct domains” (Ory & Ryan, 2001, 32)

- However, many SET instruments are developed without a clear theory on effective teaching (and were, in many cases, not tested for their reliability/validity)

- Result: a panoply of SET-instruments that show great variation in both content and construction due to the characteristics and wants of each particular institution

- These instruments might lack any evidence of **content-related validity**

1. What is ‘effective teaching’?

- Penny (2003) therefore argues in favor of an interinstitutional task force to set a list of standards or characteristics in a common framework of effective teaching which can be used as a basis for the development of SET-instruments

- Two conditions:
  - institutions can choose which aspects seem, according to their educational vision and policy, less or more important and that they can develop a SET instrument consistent with their own preferences
  - all stakeholders (i.e., teachers and students) are involved when defining these characteristics

1. What is ‘effective teaching’?

- The latter originates from the growing body of research showing that SET instruments do not always reflect the students’ perspective concerning effective teaching (Bosshardt & Watts, 2001; Onwuegbuzie *et al.*, 2007; Pan *et al.*, 2009)

- This might bias SET results, since students tend to answer to items in accordance with their own conceptions of good teaching

- SET scores are higher when students and teachers agree on the characteristics of excellent lecturers (Goldstein & Benassi, 2006)

2. SET instruments

"Until the validity of rating forms improves, the credibility of ratings feedback will continue to be compromised"

(Penny, 2003, 402)

1. The dimensionality debate

2. Questionnaire design & completing SET forms

2. SET instruments

1. The dimensionality debate

- SET = multidimensional, although there is no consensus on the nature and the number of dimensions:
  - we lack a theoretical framework concerning effective teaching (Penny, 2003)
  - views on effective teaching differ both across and within institutions (Ghedin & Aquario, 2008)
  - the measurement of dimensions continues to be relatively data-driven (with a few exceptions) (D' Apollonia & Abrami, 1997)
- Global items for summative purposes?
  - there is a need for a unidimensional and global SET score that provides a clear and precise measure of overall teaching quality
  - several authors support the multidimensionality of teaching, and furnish proof of higher-order factors that reflect, in some way, general teaching competency (Apodaca & Grad, 2005; Burdsal & Harrison, 2008; Cheung, 2000; Harrison, Douglas & Burdsal, 2004; Mortelmans & Spooren, 2009; Spooren, 2010)

2. SET instruments

2. Questionnaire design

- SET dimensions should be seen as latent constructs
  - not immediately observable using a single-item approach
  - Likert scales (incl. check internal consistency)
- Sources of bias due to both the content and the structure of scales
  - midpoint or neutral categories in SET scales (Onwuegbuzie & Weems, 2004)
  - endpoint numbering and different ranges in scales (Sedlmeier, 2006)
  - the number of response options (Landrum and Braitman, 2008)
- Response bias
  - acquiescence (Richardson, 2012; Spooren, Mortelmans &Thijssen, 2012)
  - extreme responding (Richardson, 2012)
  - responding at the favorable end of evaluation scales (Darby, 2008)
  - understanding educational terms (Billings-Gagliardi, Barrett, and Mazor, 2004; Robertson, 2004)

2. SET instruments

2. Questionnaire design

- Open comments (Burdsal & Harrison, 2009; Nasser & Fresko, 2009)
    - open comments are more often positive
    - correlate with answers to the closed-ended questions in the questionnaire (correlations ranged between .23 and .79)

3. The relationship between SET scores and factors unrelated to good teaching

"I liked your course because you taught me well"

(Remedios & Lieberman, 2008, 91)

- Many recent SET studies continue to address the question of bias, or the effect of factors that are not necessarily related to teaching quality on SET scores (e.g. teacher's gender, race, grading leniency)

- This involves the discriminant validity and divergent validity of SET, which has received considerable attention from researchers, administrators, and teachers

- Although most leading SET researchers are convinced of this type of validity of SET, 'bias'-studies continue to play a central role in the recent literature

3. The relationship between SET scores and factors unrelated to good teaching

- Not all so-called 'biasing' factors are biasing factors…
  - Student's (expected) grade and self-reported learning as proxies for student learning (convergent validity of SET)
  - Teacher's rank, teaching experience and research productivity are valuable indicators of a teacher's educational skills and his/her knowledge of the subject matter
  - Class attendance and student's effort in class are indicators of student's interest and motivation in a particular course and are at least partly dependent upon the organization of and the teaching in that course

- Limited overview of some recent studies that included possibly biasing factors and their relationship with SET
  - Mixed results, due to the great variety in measures, SET instruments, educational settings, and methods

    => generalizability?

3. The relationship between SET scores and factors unrelated to good teaching

- Student-related characteristics
  - Student's gender:
    - Female students give higher SET than male students (Centra & Gaubatz, 2000; Kohn & Hartfield, 2006; Santhanam & Hicks, 2001; Smith et al., 2007)
    - Female students give higher SET to *female* teachers (Basow et al., 2006)
    - No significant effect (Spooren, 2010)
  - Pre-course motivation/interest:
    - The stronger the desire to take/ the interest in the course, the higher SET (Griffin, 2004; Olivares, 2001)
  - Grade discrepancy:
    - Students give lower SET when expected grades are lower than they believed to deserve (Griffin, 2004)
  - External factors:
    - Students who were offered chocolate before completing SET, gave higher SET… (Youmans & Jee, 2007)

3. The relationship between SET scores and factors unrelated to good teaching

- Teacher-related characteristics
  - Teacher's gender:
    - Female teachers receive higher SET (Basow & Montgomery, 2005; Smith et al., 2007)
    - Male teachers receive higher SET (McPherson et al., 2009)
    - No significant effect (McPherson & Todd Jewell, 2007)
  - Teacher's race:
    - White teachers receive higher SET (in upper level courses) (McPherson et al. 2009)
  - Personal traits:
    - Charisma (Shevlin et al. 2006), personality (Big Five) (Clayson & Sheffet, 2006), Physical attractiveness (Patrick, 2011; Feeley, 2002; Gurung & Vespia, 2007; Hamermesch & Parker, 2005; Riniolo et al., 2006), attitude (Kim et al., 2000); likability (Delucchi, 2000), initial impression of a teacher (Tom et al., 2010)

3. The relationship between SET scores and factors unrelated to good teaching

- Course-related characteristics
  - Class size:
    - Nonlinear negative relationship between class size and SET (Bedard & Kuhn, 2008; McPherson, 2006)
    - No significant effect (Ting, 2006; McPherson et al., 2009)
  - Course discipline:
    - Natural sciences courses receive lower SET (Basow & Montgomery, 2005; Beran & Violato, 2005)
  - Required vs elective courses:
    - Elective courses receive higher SET (Ting, 2000)
  - Course type:
    - Lab-type courses receive higher SET compared to lectures (Beran & Violato, 2005)
  - Syllabus tone:
    - Teachers with a 'friendly written syllabus' receive higher SET

      (Harnish & Bridges, 2011)

3. The relationship between SET scores and factors unrelated to good teaching

"I liked your course because you taught me well"

(Remedios & Lieberman, 2008, 91)

- This high degree of variation calls the generalizability of these results into question and makes it almost impossible to make statements concerning the strength of the relationship of various possibly biasing effects on SET scores

- However, several researchers have found that the effect of the possibly biasing factors on SET is relatively small (Beran & Violato, 2005; Centra, 2003; Marsh & Roche, 2000; Smith et al, 2007; Spooren, 2010)

- These findings suggest that SET outcomes depend primarily upon teaching behavior (Barth, 2008; Greimel-Fuhrmann & Geyer, 2003)

4. SET procedures and administration

"To improve the validity of our student ratings, we need to both improve our practices and conduct research on their use and consequences"

(Ory & Ryan, 2001, 40)

- Even if all of these biasing challenges are under control and even if SET provides valid information concerning the quality of teaching, it is still possible for such evaluations to be administered and used in inappropriate ways

- Use affects the outcome validity of SET

- The ways in which administrators engage with SET constitute one of the greatest threats to the validity of SET (Penny, 2003)

4. SET procedures and administration

- Guidelines for the collection and interpretation of SET data are available, but many SET users are not sufficiently trained to handle these data (and they may even be unaware of their own ignorance)

- Moreover, they lack knowledge about the existing research literature on SET

- Although the misuse and miscollection of data might have consequences for both the improvement of teaching and the careers of the teachers involved, little research is available concerning this topic

4. SET procedures and administration

- Administrators have challenged the validity of SET based on limited psychometric knowledge (Franklin, 2001; Sproule, 2000; Wolfer & Johnson, 2003)
- They prefer aggregated and overall measures of student satisfaction, often failing to consider both basic statistical and methodological matters (e.g., response rate, score distribution, sample size) when interpreting SET (Gray & Bergmann, 2003; Menges, 2000) and making spurious inferences based on these data
- For example, Franklin (2001) reported that about half of the SET administrators involved in the study were unable to provide sound answers to several basic statistical questions
- The proper collection and interpretation of SET data depend upon administrators having sound methodological training and regular briefing on the major findings and trends in the research field

5. SET and the internet

"(BEST TEACHER EVER)^Infinity. Hands down. I hope she teaches Calc 3. She makes even the hardest sections understandable. I made my schedual [sic] around this professor and this class. I wish she taught all of my classes"

(Anonymous, on ratemyprofessors.com)

- SET-administrators nowadays use electronic evaluation procedures to collect SET data instead of the more classic paper-and-pencil instruments
  - greater accessibility to students
  - quick and accurate feedback
  - no disruption of class time
  - more accurate analysis of the data
  - better written comments
  - guaranteed student anonymity (e.g., decreased risk of recognition due to hand-writing)
  - decreased vulnerability to faculty influence
  - lower costs
  - reduced time demands for administrators

5. SET and the internet

- Some parties nevertheless fear that SET results obtained in this way are easier to trace and can be consulted by almost everyone (Gamliel & Davidovitz, 2005)

- Response rates in such evaluation procedures are lower than is the case with paper-and-pencil questionnaires. Dommeyer et al. (2004) reported average response rates of 70% for in-class surveys and 29% for online surveys

5. SET and the internet

- In recent years, the territory of SET has also expanded beyond the exclusive domain of institutions to the World Wide Web through such faculty-rating sites as RateMyProfessors.com, PassCollege.com, ProfessorPerformance.com, Ratingsonline.com, and Reviewum.com

- Recent SET research focused on:

  1) the validity of SET results that are obtained from institutional electronic procedures to ascertain if these procedures provide SET scores that are comparable to those obtained from the more classic paper-and-pencil procedures

  2) the rise of online SET platforms (such as RateMyProfessors) and their relationship with SET scores obtained from institutional procedures

5. SET and the internet

1. Electronic institutional procedures

- No significant differences between SET scores obtained from paper-and-pencil evaluations and those obtained through electronic evaluations (Leung and Kember, 2005; Liu, 2006)
- At the aggregate level, electronic SET scores are lower than are those obtained with paper-and-pencil surveys. These differences disappear, however, when controlling for course and instructor (Barkhi & Williams, 2010)
- Electronic SET instruments generate more extreme negative responses to Likert-type items than do paper-based surveys (Barkhi & Williams, 2010)
- Student comments in electronic evaluations are more detailed than are those in paper-and-pencil questionnaires (Venette, Sellnow, & McIntyre, 2010)

- Although electronic surveys obviously offer considerable advantages, their greatest challenge continues to involve increasing the response rate (Johnson, 2003)

5. SET and the internet

2. Online SET platforms

- Most popular site: RateMyProfessors
- 15 million ratings, 1.4 million professors and over 7,000 (Anglo-Saxon) schools

  - The rating form consists of five single-item type questions concerning the easiness, clarity, and helpfulness of the teacher, as well as the student's level of interest prior to attending class and the use of the textbook during the course
  - Opportunity to add additional detailed comments about the course or the professor
  - Finally, students are asked to rate the appearance of the teacher involved as "hot" or "not" (although the website suggests that this rating is "just for fun")

5. SET and the internet

2. Online SET platforms : RateMyProfessors

- Subject to a noncontrolled self-selection bias => representativeness, validity, and reliability of the results? (Davison & Price, 2009)
- Nevertheless, many students use these ratings as a source of information about their teachers (Otto et al., 2008)
  - Ratings on the RateMyProfessors website show statistically significant positive correlations (that exceed .60) with institutionally based SET (Sonntag, Bassett, & Snyder, 2009; Timmerman, 2008)
  - In general, more lenient instructors receive higher overall quality ratings: Instructor's Easiness predicted 50% of the variance in the scores on the Overall Quality measure (Stuber et al., 2009)
  - There is a positive correlation between overall ratings and the leniency and sexiness of instructors (correlations were .61 and .30, respectively) (Felton et al., 2004). The "hotness" variable accounted for almost 9% of the variance in SET scores (Freng & Webber, 2009)

6. SET and the improvement of teaching

"Student ratings are the start of the instructor's journey toward improvement, not the end." (Cashin, 1994, 1)

- An important outcome of SET would be to provide student feedback for the improvement of teaching in particular courses
- One important question addressed in the recent SET literature, therefore, involves the relationship between SET and the improvement of teaching

  - Davidovitch & Soen (2006) showed that SET improves over time (with the age and seniority of teachers as particularly important predictors)
  - Kember et al. (2002) found no evidence that such evaluations contribute to the improvement of teaching, as SET scores did not increase over the years
  - Marsh (2007) demonstrated that SET reports are highly stable over time, including with regard to the individual differences between teachers

6. SET and the improvement of teaching

- Possible explanation: student feedback obtained from the SET questionnaire is not used effectively
- Teachers should have the opportunity to consult with colleagues or educational experts about their SET reports:
    - Consulting with faculty about their SET has a moderate to large positive effect (.68) on teaching quality, even when controlling for variables reflecting bias and unfairness (Dresel & Rindermann, 2011)
    - Providing feedback by SET reports alone (without consultation) is far less effective than many assume in the long run (a strong increase in SET results the next semester, which was followed by declines over the next three semesters) (Lang & Kersting, 2007)

6. SET and the improvement of teaching

- Penny & Coe (2004) listed eight strategies that are important when providing consultative feedback:
  - active involvement of teachers in the learning process
  - use of multiple sources of information
  - interaction with peers
  - sufficient time for dialogue and interaction
  - use of teacher self-ratings
  - use of high-quality feedback information
  - examination of conceptions of teaching
  - setting of improvement

Conclusion

The validity of SET remains at stake

Concept of effective teaching in SET instruments

Poorly designed SET instruments in many institutions

The influence of factors that are unrelated to good teaching

Competence of SET administrators

SET in online environments: self selection bias?

SET feedback alone is not enough for the improvement of teaching

Generalizability of SET studies

Conclusion

Suggestions for future SET practice

Use well-designed and thoroughly validated SET-instruments

Bias can occur at many levels and in each step of a SET-procedure: collecting and interpreting SET should be done with great caution

Check for response sets and spurious relationships between SET and variables that are unrelated to good teaching

The competent SET administrator is well posted in educational theory, the SET-literature and statistics

SET is only one indicator of teaching competence

SET for both formative and summative purposes: code word = 'TRUST'

pieter.spooren@uantwerpen.be

University of Antwerp

University of Antwerp