

Athens Institute for Education and Research

ATINER



ATINER's Conference Paper Series

EDU2013-0407

**Student Evaluation of Teaching
(SET) in Flemish Higher
Education. Lessons Learned and
Suggestions for Future Practice**

**Pieter Spooren, PhD
Educational Advisor / Researcher
Faculty of Social and Political Sciences
University of Antwerp
Belgium**

**Francis Van Loon, PhD
Full professor, Honorary Rector
Faculty of Social and Political Sciences
University of Antwerp
Belgium**

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209
Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

Printed in Athens, Greece by the Athens Institute for Education and Research.
All rights reserved. Reproduction is allowed for non-commercial purposes if the
source is fully acknowledged.

ISSN 2241-2891

13/06/2013

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Spooren, P. and Van Loon, F., (2013) "Student Evaluation of Teaching (SET) in Flemish Higher Education. Lessons Learned and Suggestions for Future Practice" Athens: ATINER'S Conference Paper Series, No: EDU2013-0407.

Student Evaluation of Teaching (SET) in Flemish Higher Education. Lessons Learned and Suggestions for Future Practice

Pieter Spooren, PhD
Educational Advisor / Researcher
Faculty of Social and Political Sciences
University of Antwerp
Belgium

Francis Van Loon, PhD
Full professor, Honorary Rector
Faculty of Social and Political Sciences
University of Antwerp
Belgium

Abstract

In contemporary higher education, teachers work within a web of evaluations in which they are appraised, watched and assessed. Student evaluations of teaching (SET) is one method of assessing teaching and course quality. Although SET had a mainly formative character, they also quickly became an important instrument in faculty personnel decisions. Whereas the first standardized student evaluation procedures were introduced at some American and Canadian universities in the 1920's, SET were - much later - almost silently introduced in European universities. On top of that, SET-studies from the 'old' continent are rather rare.

This paper adds a summary of the first Flemish contribution to the SET-literature. We report on the state of the art of the SET-research literature which serves as the theoretical framework. Then, we discuss the construction and validation of the SET37-questionnaire for students' evaluations of teaching at the University of Antwerp. We continue with three topics concerning the influence of possible biasing factors on SET-results, i.e. acquiescence, various student-, teacher- and course characteristics, and the issue of non-response. Using structural equations modeling we found no evidence for acquiescence. At the student level, significant relationships exist between SET-scores and course grade, class attendance and students' age. At the course/teacher level, teachers' rank had a significant relationship with SET-scores. Course grade, program level, and the number of course evaluations a student was asked to complete, are significant predictors for participation. The implications for both further research and practical implications concerning the use of SET are discussed.

Corresponding Author:

Introduction

Whereas SET in the early days had a mainly formative character, in the 1970s they quickly became an important instrument in faculty personnel decisions as well (Galbraith, Merrill, Kline, 2011). And more recently, SET-procedures are included as a key mechanism in internal quality assurance processes to prove an institution's performance in accounting and auditing practices (Johnson, 2000). The main purpose of SET is thus threefold: 1) improving teaching quality, 2) appraisal exercises (tenure/promotion decisions) and 3) institutional accountability (demonstrating adequate procedures for ensuring teaching quality) (Kember, Leung & Kwan, 2002).

This triple use, and the unresolved tension between them, makes the use of SET very delicate (Penny, 2003). On the one hand, most teachers are convinced of the usefulness of SET as an instrument for feedback on their teaching (Balam & Shannon, 2010). On the other hand, it is argued that nowadays the principal purpose of SET lies in its use as a measure for quality monitoring and administrative policy-making and mapping whether or not teachers reach a certain required standard in their teaching practice (Chen & Hoshower, 2003; Penny & Coe, 2004). Evaluation for accountability, rather than improvement, seems to have become the principal goal of SET, which reflects the shift from 'professionalism' to 'performativity' in the managerial approach that currently reigns in almost all institutions for higher education (Arthur, 2009). This sometimes results in teacher's practices aimed at increasing SET-scores rather than improving instruction (Simpson & Siguaw, 2000). The 'tyranny' of the evaluation form may even lead to grading leniency and, as a consequence, grade inflation (Langbein, 2008; Oleinik, 2009). At the same time, many valuable thoughts and suggestions of the students for the improvement of teaching remain untouched.

Whereas the first systematic and standardized student evaluation procedures were introduced at some North American and Canadian universities in the 1920's (Marsh, 1987), SET were almost silently introduced in European universities. SET-studies from the 'old' continent are rather rare. In Flanders (the Dutch speaking part in Belgium), for instance, standardized SET procedures were implemented as a result of the quality audits, organized by the Flemish Interuniversity Council (VLIR/Vlaamse Interuniversitaire Raad) since the early 1990's. Although, more than 20 years later, only few Flemish studies on the use and validity of SET are available, in almost all Flemish universities and university colleges, SET-reports form part of teachers' personnel files.

When looking at the recent SET-literature, it is clear that SET remain a hot, yet delicate topic in both higher education and educational research. Many stakeholders are not convinced of the usefulness and validity of SET. This has, among others, much to do with the fact that SET research did not succeed in providing clear and unambiguous answers to several critical aspects concerning the validity of SET (Kulik, 2001). The recent literature mainly focuses on four main themes which are crucial for the validity and the relevance of SET, i.e. 1) the psychometric value of the SET-instruments, 2) the validity and reliability of

student opinions, 3) SET-procedures and their context, and 4) SET and its contribution to the improvement of teaching.

Regarding the *SET-instruments*, a common understanding and a conceptual framework concerning the concept of effective teaching, upon which SET-instruments should be grounded, still does not exist (Onwuegbuzie, Daniel & Collins, 2009). Many SET-instruments that are used for administrative decision-making, remain to be *ad hoc* instruments that were not tested at all, although thoroughly validated SET-instruments are available (Richardson, 2005).

Concerning the reliability and validity of *student opinions*, the main topic in the SET-literature still tackles the so called 'bias'-question, i.e. to what extent are SET-scores influenced by student-, teacher-, and course characteristics that are unrelated to good or effective teaching (Centra & Gaubatz, 2000)? Unfortunately, the sometimes contradictory findings concerning the (strength of the) relationships between SET on the one hand, and student, course, and teacher characteristics on the other, do not promote a conclusive idea of possibly biasing factors that affect SET. This has much to do with the number of controlling variables included in these studies, the way these variables are measured, the various research techniques applied, and the characteristics of the samples in these studies. It then is very difficult to make valuable statements concerning the generalizability of the results (Beran & Violato, 2005).

With respect to *SET-procedures and their context*, it seems that students are in general willing to participate in SET-procedures, although they think that teachers and institutions make no or only little use of SET-results (Spencer & Schmelkin, 2002). Teachers generally agree with SET for both demonstrating the quality of education at institutions and personnel decisions (Beran & Rokosh, 2009), but indeed make little use of SET for the improvement of teaching (Nasser & Fresko, 2002). On top of that, responding to SET appears to be harder than many stakeholders assumed (Arthur, 2009; Moore & Kuol, 2005). It is therefore important that SET is done with great caution and that teachers can reckon on peers, colleagues and administrators when interpreting their SET-results. SET-administrators should be trained in both statistics and educational theory, and should be well posted in the SET literature as well (Menges, 2000). A well-skilled and integer administrator can remove much of the teachers' concerns regarding SET. Online SET have become the norm at many institutions for higher education (Nulty, 2008). It is suggested that online SET show similar results compared to paper-based SET, although students take more time to give comments in an online environment (Leung & Kember, 2005; Liu, 2006). A disadvantage of online SET however appears to be the lower response rates (Gamliel & Davidovitz, 2005).

The findings concerning *SET and the improvement of teaching* in the long run are somewhat surprising, as it is suggested that SET as such do not lead to better teaching (Kember, Leung & Kwan, 2002; Marsh, 2007). As a consequence, 1) SET should be embedded in a more holistic approach on the evaluation of teaching wherein teachers take the matter up and actually reflect

on the improvement of their teaching in a course, and 2) that, presumed that teachers actually become more mature and better teachers as they build up experience, SET should not be the only means to map a teacher's (progress in) teaching.

Objectives

The SET-literature is dominated by research in Anglo-Saxon educational settings. The studies presented in this paper add a first Flemish contribution to the SET-literature. These studies originate from a research project that aimed at validating a SET-instrument at the University of Antwerp, a Flemish medium-sized multidisciplinary university (15,000 students). To provide both decent student feedback to the teachers and input concerning teaching quality for both administrative decision-making and institutional accountability purposes, the Educational Board of the University of Antwerp searched for a well-designed and thoroughly validated instrument to collect student experiences in courses. Nevertheless, it appeared that most existing questionnaires were ad-hoc instruments that used a single-item approach and lacked a theoretical foundation and/or evidence for its validity and reliability. After succeeding in designing the SET37-instrument that meets all of these requirements, we decided to replicate major findings in the existing SET-literature in the Flemish context (i.e., the relationship between student-, course-, and teacher characteristics and SET-scores) and to explore some rather undeveloped themes concerning the validity of SET (i.e., acquiescence in SET and non-response). In this paper, the main results of each study are summarized and discussed. The final sections of this paper provide some suggestions for both further research and practical implications concerning the use of SET.

Flemish studies on SET

The construction and validation of the SET37-questionnaire

Consistent with the argument that SET-instruments should be based on both educational theory and thorough validation procedures, an instrument for students' evaluation of teaching at the University of Antwerp was constructed (Spooren, Mortelmans & Denekens, 2007). In a three-step procedure, we defined 22 dimensions of teaching based on both the educational theory and the institution's educational policy plan. These dimensions were cut back to 10 dimensions (10 scales, 31 items) during an empirical validation procedure that contained (exploratory and confirmatory) factor analysis and various reliability tests. In the meantime, two new scales were added and the instrument was put through a re-validation procedure (Mortelmans & Spooren, 2009) (Table 1).

The results of these studies underline the value of the use of scaling techniques in SET-instruments. In comparison with the single-item approach (which is used in most SET-instruments), scale-type evaluations measure instructional

skills (which must be seen as latent constructs) better since they are less sensitive to social desirability, ambiguous interpretations and accidental fluctuations of the answers given. On top of that, the internal consistency of each scale is easily tested by calculating the Cronbach's alpha statistic. It may be clear that the construction of valid and reliable SET-scales requires systematic research, in which both the literature and empirical data should play an important role. Still, these validation procedures do not seem to take place in all institutions for higher education yet. This strengthens the arguments of those who are concerned about the reliability, validity and the usefulness of SET.

A higher-order factor?

Although there is consensus on the multidimensionality of effective teaching (i.e., teaching consists of various aspects) and the recommendation that SET-instruments should capture as much dimensions as possible (especially for providing more detailed feedback concerning specific aspects of a course), some authors argued in favor of a one dimensional way of collecting SET by means of a single, global score (especially for summative purposes). Apodaca & Grad (2005) showed that SET indeed have a multidimensional structure, but are compatible with a very strong general underlying factor (that represents a 'general instructional skill').

Confirmatory factor analysis on the scales and of the SET37-questionnaire provides evidence for the existence of such a higher-order factor which explains much of the variance in seven dimensions (Spooren & Mortelmans, 2006). We called this factor the 'teacher professionalism'-factor, as it captures these scales that measure the way a teacher built up, organised and executed his/her course. If he/she managed to do this professionally, this will be rewarded by the students as 'good teaching' and thus with higher ratings on the 7 scales in the instrument.

This offers a promising perspective towards the use of SET37-questionnaire in both a formative and a summative way. When using the SET37 as an instrument for feedback, one could use the results on one or more particular dimensions when working on the improvement of (teaching) a course. On the other hand, an overall score, derived from the (weighted if necessary) SET37-scores on dimensions of which it is known that they belong to an underlying general factor, could be used for the evaluation of teaching staff.

Acquiescence in SET?

The use of Likert scales in the SET37-questionnaire allows to test for acquiescence (or 'yeah-saying') as a response style which might lead to incorrect SET-results. After all, if students are very attentive when filling in course evaluation forms, no substantial effect of acquiescence will be found. Acquiescence might appear as a consequence of personal considerations (such as social desirability), motivational aspects (i.e., students who are confronted with the umpteenth SET-questionnaire), or psychometric shortcomings of SET-instruments (i.e., vague or too difficult item wordings).

Using structural equation modeling in two different datasets, we found no evidence for acquiescence (Spooren, Mortelmans & Thijssen, 2012), since the correlation between a higher-order factor in the instrument and a ‘sum of agreements’-variable is rather weak (.10). Because only few studies studied acquiescence in SET, it remains necessary to control and to correct (if necessary) for acquiescence when interpreting and using SET-results for both summative and formative evaluation of teachers and their courses. Designers of SET-instruments should be well posted in the subject and take into account the potential risk of acquiescence as a response style in student populations.

Are SET influenced by possibly biasing student, teacher, and course characteristics?

A preliminary analysis in the aforementioned ‘higher-factor’ study revealed the existence of a significant and positive association between a students’ course grade and his/her evaluation of the course (and the teacher). Although there are many possible explanations for this relationship (Brockx, Spooren & Mortelmans, 2011; Gump, 2077; Marsh, 1987), one could advocate in favor of the so called ‘grading leniency’ hypothesis, which suggest that high SET-scores can be ‘bought’ by giving high grades (and low workload). In this case, student’s course grade should be considered as a ‘biasing’ variable, as it has nothing to do with effective teaching. Similar variables are students’ gender, teachers’ gender, teachers’ sexual orientation, class size *et cetera*. To find out which variables at the student level, and the course/teacher level influenced SET-scores at the University of Antwerp, we conducted a cross-classified multilevel study with the score on the ‘teacher professionalism’-factor as the dependent variable (Spooren, 2010). The results show that at the student level, significant relationships exist between SET-scores and course grade, examination period, class attendance, and students’ age. The associations between SET-scores and students’ gender and students’ overall grade were not significant. At the course/teacher level, it was found that teachers’ rank had a significant relationship with SET-scores, whereas teachers’ department, teachers’ age, teacher’s gender, and class size were not correlated with SET-scores. Still, the models suggest that these variables explain only little variance in SET-scores (maximum PRV is 6%). Thus, one can hardly speak from serious bias, since these ‘popular’ and much studied predictors (although statistically significant) appear to be not important.

A non-response analysis on online SET-data (using multilevel logistic models) showed that at the student level, course grade, program level, and the number of course evaluations a student was asked to complete are significant predictors for participation. Students gender and study domain are not significant (Spooren & Van Loon, 2012).

Limitations

When we stem from the main research topics in the recent SET-literature, i.e.,

1) the psychometric value of the SET-instrument, 2) the validity and reliability of student opinions, 3) SET-procedures and their context, and 4) SET and its contribution to the improvement of teaching, it is clear that the studies presented in this paper contribute to the topics 1 and 2. On the one hand, we proved the validity of the SET37-questionnaire for students' evaluation of teaching. On the other hand, it was shown that students' opinions in SET do not suffer that much from possibly biasing factors such as response styles (i.e., acquiescence), student characteristics (e.g., course grade, examination period, student's age), course characteristics (i.e., class size), and teacher characteristics (i.e., teacher's rank).

Still, these studies have two important limitations which confine their generalizability since, in general, it can be said that these studies were executed in *a particular setting* using *a particular instrument*. First, we lack a common theoretical framework concerning effective teaching upon which SET-instruments can be built. This argument undermines the construct validity of the SET37-questionnaire itself, since it was 'only' based on educational theory and the educational policy plan of the University of Antwerp. Although the SET37 went in for various validation procedures, it stems not from a general (i.e., inter-institutional or international) agreement on the concept of effective teaching in higher education. Which makes the SET37 nothing more or less than a well-designed *institutional* SET-questionnaire. Cross-validation procedures in other institutions are needed to prove the generalizability of the instrument in other settings. Second, the results of the bias-studies might be influenced by the SET-practice at the University of Antwerp. Contrary to some other institutions, SET are not compulsory for students. Although they are invited several times to take part in a SET-procedure, they still can decide not to participate. *Self-selection of the respondents* can bias the results in SET-studies such as these presented in this paper. The results are applicable for this particular *institution*, and only serve as a useful indicator for SET-research at other institutions. Moreover, this statement counts for most studies reported in the SET-literature, which usually do not provide much information on the SET-procedures and response rates upon which these studies were done.

Suggestions for further research

The results and limitations of the studies presented in this paper provoke some suggestions for further research.

First, we resume the call of other authors (Onwuegbuzie, Daniel & Collins, 2009; Penny, 2003) with respect to the installation of a common conceptual framework concerning effective teaching in higher education, which *inter alia* supports the construction of SET-instruments. Especially concerning the SET-matters, it seems very important to draw all stakeholders into this process.

Second, the results of the acquiescence study need to be repeated in other institutions with other SET-procedures and other SET-instruments. For instance, does acquiescence appear, due to motivational reasons, in SET-results

from institutions where SET are compulsory?

Third, although much attention was paid to the relationship between SET and (quite immediate observable) student, teacher, and course characteristics, the role of contextual factors (i.e., an institution's SET-practice) should be brought into play as well.

Fourth, future research should replicate previous findings concerning the validity of SET-instruments and student opinions (which originally are based on paper-and-pencil procedures) in online environments, since many institutions now use electronic media when collecting SET.

Fifth, research is needed concerning the feelings, competences and actual practices of administrators in SET-procedures. Although they should be considered key figures in SET-practice (designing and using instruments, gathering SET-data, reporting and interpreting SET-results, providing educational advice or personnel advice based on SET-scores), only little is known about them. For instance, several authors claim that the validity of SET is challenged by administrators' psychometric knowledge concerning their interpretation of SET (Franklin, 2001; Sproule, 2000).

Sixth, with respect to the (possible) self-selection bias question, it would be interesting to profile those students which do participate in SET, and compare them with those who decided not to co-operate. In this way, the representativeness of the respondents for the whole population can be defined and more insight can be gained in students' motives to (not) take part in SET.

Practical implications

The studies presented in this paper have some implications for current and future SET-practice as well.

First, it is shown that SET-questionnaires by prefer consist of multiple dimensions, wherein each of these dimensions is measured by means of multiple Likert-type items. This grants a straightforward quality check of the reliability of SET-scores for each dimension every time a single course is measured, and is very usable when controlling for response styles, –especially when using balanced or quasi-balanced scales.

Second, SET-questionnaires should be based on both educational theory and thorough validation procedures, rather than being *ad hoc* instruments that were not tested at all.

Third, gathering SET and interpreting SET-results should be done with great caution, since bias can occur at many levels and at each step in the *SET-procedure*. It is therefore important to regularly search for response styles, and spurious relationships between SET-scores and variables that having nothing to do with the quality of (teaching in) a course. Above all, SET-administrators should be trained in both statistics and educational theory, and should be well posted in the SET-literature which addressed already many pitfalls concerning SET-practice. Besides, a well-skilled, well-informed and integer administrator can remove much of the teachers' concerns regarding SET.

Fourth, the literature suggests that SET as such do not guarantee the *improvement of teaching*. It is therefore important to use SET as only one (and not the only one) means when mapping teaching quality and to challenge teachers to reflect on the improvement of their teaching by providing them with information based on multiple indicators of teaching quality. Another (and even more important) means for the improvement of teaching is an ongoing ‘thinking together’-culture concerning teaching (a particular course), in which the teacher, fellow teachers, educational advisors and the students are involved. This can be done in focus groups, informal meetings, or in the corridor.

Fifth, when using SET for both formative and summative purposes, the code word is ‘trust’. Teachers should be well-informed about the use and the importance of the SET-results and be surrounded by peers, colleagues and administrators when interpreting (and dealing with) their SET-results. One or two poor SET-reports should not be decisive when judging a teacher’s performance. On the contrary, teachers should be encouraged to experiment with teaching methods and course contents, no matter what the SET-consequences are. Only then SET-results can and certainly will be a very valuable means for the improvement of teaching.

References

- Apodaca, P. & Grad, H. (2005). ‘The dimensionality of student ratings of teaching: integration of uni- and multidimensional models’. *Studies in Higher Education* 30: 723-748.
- Arthur, L. (2009). ‘From performativity to professionalism: lecturer’s responses to student feedback’. *Teaching in Higher Education* 14: 441-454.
- Balam, E., & Shannon, D. (2010). ‘Student ratings of college teaching: a comparison of faculty and their students’. *Assessment and Evaluation in Higher Education* 35: 209-221.
- Beran, T. & Rokosh, J.L. (2009). ‘Instructor’s perspectives on the utility of student ratings of instruction’. *Instructional Science* 37: 171-184.
- Beran, T., & Violato, C. (2005). ‘Ratings of university teacher instruction : how much do student and course characteristics really matter ?’ *Assessment and Evaluation in Higher Education* 30: 593-601.
- Brockx, B., Spooren, P., & Mortelmans, D. (2011). ‘Taking the ‘grading leniency’ story to the edge. The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education’. *Educational Assessment, Evaluation and Accountability* 23: 289-306.
- Centra, J.A., & Gaubatz, N.B. (2000). ‘Is there gender bias in student evaluations of teaching?’ *The Journal of Higher Education* 71: 17-33.
- Chen, Y., & Hoshower, L. (2003). ‘Student evaluation of teaching effectiveness: an assessment of student perception and motivation.’ *Assessment and Evaluation in Higher Education* 28: 71-88.
- Franklin, J. (2001). ‘Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately.’ *New Directions for Teaching and Learning* 87: 85-100.
- Galbraith, C., Merrill, G., & Kline, D. (2011). ‘Are student evaluations of teaching

- effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses.' *Research in Higher Education* 53: 353-374.
- Gamliel, E. & Davidovitz, L. (2005). 'Online versus traditional teaching evaluations: mode can matter.' *Assessment & Evaluation in Higher Education* 30: 581-592.
- Gump, S.E. (2007). 'Student evaluation of teaching effectiveness and the leniency hypothesis: a literature review' *Educational Research Quarterly* 30: 55-68.
- Johnson, R. (2000). 'The authority of the student evaluation questionnaire.' *Teaching in Higher Education* 5: 419-434.
- Kember, D., Leung, D., & Kwan, K. (2002). 'Does the use of student feedback questionnaires improve the overall quality of teaching?' *Assessment and Evaluation in Higher Education* 27: 411-425.
- Kulik, J.A. (2001). 'Student ratings: validity, utility and controversy.' *New Directions for Institutional Research* 27: 9-25.
- Langbein, L. (2008). 'Management by results: student evaluation of faculty teaching and the mis-measurement of performance.' *Economics of Education Review* 27: 417-428.
- Leung, D.Y.P. & Kember, D. (2005). 'Comparability of data gathered from evaluation questionnaires on paper and through the internet.' *Research in Higher Education* 46: 571-591.
- Liu, Y. (2006). 'A Comparison Study of Online versus Traditional Student Evaluation of Instruction.' *International Journal of Instructional Technology and Distance Learning* 4: 15-29.
- Marsh, H.W. (1987). 'Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research.' *International Journal of Educational Research* 11: 253-388.
- Marsh, H.W. (2007). 'Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years.' *Journal of Educational Psychology* 99: 775-790.
- Menges, R.J. (2000). 'Shortcomings of research on evaluating and improving teaching in higher education.' *New Directions for Teaching and Learning* 83: 5-11.
- Moore, S. & Kuol, N. (2005). 'Students evaluating teachers: exploring the importance of faculty reaction to feedback on teaching.' *Teaching in Higher Education* 10: 57-73.
- Mortelmans, D. & Spooren, P. (2009). 'A revalidation of the SET37-questionnaire for student evaluations of teaching.' *Educational Studies* 35: 547-552.
- Nasser, F. & Fresko, B. (2002). 'Faculty views of student evaluation of college teaching.' *Assessment & Evaluation in Higher Education* 27: 187-198.
- Nulty, D.D. (2008). 'The adequacy of response rates to online and paper surveys: what can be done?' *Assessment & Evaluation in Higher Education* 33: 301-314.
- Oleinik, A. (2009). 'Does education corrupt? Theories of grade inflation.' *Educational Research Review* 4: 156-164.
- Onwuegbuzie, A.J., Daniel, L.G., & Collins, K.M.T. (2009). 'A meta-validation model for assessing the score-validity of student teaching evaluations.' *Quality & Quantity* 43: 197-209.
- Penny, A.R. (2003). 'Changing the agenda for research into student's views about university teaching: four shortcomings of SRT research.' *Teaching in Higher Education* 8: 399-411.
- Penny, A.R., & Coe, R. (2004). 'Effectiveness of consultation on student ratings feedback: a meta-analysis.' *Review of Educational Research* 74: 215-253.

- Richardson, J.T.E. (2005). 'Instruments for obtaining student feedback: a review of the literature.' *Assessment and Evaluation in Higher Education* 30: 387-415.
- Simpson, P., & Siguaw, J. (2000). 'Student evaluations of teaching: an exploratory study of the faculty response.' *Journal of Marketing Education* 22: 199-213.
- Spencer, K.J. & Schmelkin, L.P. (2002). 'Student perspectives on teaching and its evaluation.' *Assessment & Evaluation in Higher Education* 27: 397-409.
- Spooren, P. (2010). 'On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching.' *Studies in Educational Evaluation* 36: 121-131.
- Spooren, P., Mortelmans, D., & Denekens, J. (2007). 'Student evaluation of teaching quality in higher education. Development of an instrument based on 10 Likert scales.' *Assessment and Evaluation in Higher Education* 32: 667-679.
- Spooren, P. & Mortelmans, D. (2006). 'Teacher professionalism and student evaluation of teaching: will better teachers receive higher ratings and will better students give higher ratings?' *Educational Studies* 32: 201-214.
- Spooren, P., Mortelmans, D., & Thijssen, P. (2012). 'Content vs. style. Acquiescence in student evaluations of teaching?' *British Educational Research Journal* 38: 3-21.
- Spooren, P. & Van Loon, F. (2012). 'Who participates (not)? A non-response analysis on students' evaluations of teaching.' Paper presented at the International Conference on Education and Educational Psychology, Istanbul, October 10-13th, 2012.
- Sproule, R. (2000). 'Student Evaluation of Teaching: Methodological Critique.' *Education Policy Analysis Archive*, 8, 50. Retrieved from <http://epaa.asu.edu/ojs/article/view/441>