This item is the archived peer-reviewed author-version of:

Exploring high impact scholarship in research on student's evaluation of teaching (SET)

Reference:

Spooren Pieter, Vandermoere Frédéric, Vanderstraeten Raf, Pepermans Koen.- Exploring high impact scholarship in research on student's evaluation of teaching (SET)

Educational research review - ISSN 1747-938X - Oxford, Elsevier sci ltd, 22(2017), p. 129-141

Full text (Publisher's DOI): <https://doi.org/10.1016/J.EDUREV.2017.09.001>

To cite this reference: <https://hdl.handle.net/10067/1454880151162165141>

Accepted Manuscript

Exploring high impact scholarship in research on student's evaluation of teaching (SET)

Pieter Spooren, Frederic Vandermoere, Raf Vanderstraete, Koen Pepermans

PII: S1747-938X(17)30033-7

DOI: [10.1016/j.edurev.2017.09.001](https://doi.org/10.1016/j.edurev.2017.09.001)

Reference: EDUREV 227

To appear in: *Educational Research Review*

Received Date: 3 November 2015

Revised Date: 29 August 2017

Accepted Date: 8 September 2017

Please cite this article as: Spooren, P., Vandermoere, F., Vanderstraete, R., Pepermans, K., Exploring high impact scholarship in research on student's evaluation of teaching (SET), *Educational Research Review* (2017), doi: 10.1016/j.edurev.2017.09.001.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Exploring high impact scholarship in research on student's evaluation of teaching (SET)

Pieter Spooren*, Frederic Vandermoere*, Raf Vanderstraete# and Koen Pepermans*

* University of Antwerp

Ghent University

Author Note

Pieter Spooren, Faculty of Social Sciences, University of Antwerp, Belgium; Frederic Vandermoere, Department of Sociology, University of Antwerp, Belgium; Raf Vanderstraete, Department of Sociology, Ghent University; Koen Pepermans, Faculty of Social Sciences, University of Antwerp, Belgium.

Correspondence concerning this article should be addressed to Pieter Spooren, Department of Social Sciences, University of Antwerp, Belgium.

E-mail: pieter.spooren@uantwerpen.be

Abstract

Student evaluation of teaching (SET) is the most common way of assessing teaching quality at universities. Since the introduction of SET procedures at the start of the previous century, thousands of research studies on the validity, reliability and utility of SET were written. By means of a citation analysis on journal articles included in Google Scholar, Scopus, and the Social Science Citation Index (Web of Science), this paper aims at mapping the high impact studies, the leading researchers, and the key journals in this research field. The results indicate that, although we find considerable overlap between the three databases, a number of high impact journal papers are not included in all three databases. Furthermore, the analysis reveals three main topics in the SET literature: the use of SET, validity issues concerning SET, and the construction and validation of SET instruments. Also, it is shown that many high impact studies were written by only a few researchers, with Herbert Marsh as the leading author. Although some of the most impactful studies date back to the 1960s, it's coming of age situates in the seventies. Since then SET became increasingly visible. The high proportion (25%) of impactful articles since 2000 indeed suggests a trend of continuous growth in SET research. At the same time, the historical knowledge in the form of classic studies on SET lives on through the many citations in recent studies.

Keywords: student evaluation of teaching, validity, citation analysis, higher education

Exploring high impact scholarship in research on student's evaluation of teaching (SET)

Nowadays, student evaluation of teaching (SET) is used as a measure of teaching performance in almost every institution of higher education throughout the world. This widespread use is largely due to the apparent ease of collecting the data and presenting and interpreting the results (Penny, 2003). Whereas SET in the early days had a mainly formative character, in the 1970s it quickly became an important instrument in faculty personnel decisions as well (Galbraith, Merrill & Kline, 2011). 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: (a) improving teaching quality, (b) appraisal exercises (tenure/promotion decisions), and (c) institutional accountability (demonstrating adequate procedures for ensuring teaching quality) (Kember, Leung and Kwan, 2002).

Since the introduction of the first 'teacher rating scale' in 1915 (Marsh, 1987; Wachtel, 1998), SET instruments and procedures have however provoked heated discussions among the various stakeholders. In general, their concerns include (a) the differences between the ways in which students and teachers perceive effective teaching, (b) the relationships between SET scores and factors that are unrelated to "good teaching" (Centra, 2003; Marsh, 2007), (c) SET procedures and practices (i.e. the contents of SET reports, the depersonalization of the individual relationship between teachers and their students due to the standardized questionnaires and respondents' anonymity, the competency of SET administrators, the low response rates, etc.), and (d) the psychometric properties of SET instruments. In the wake of the implementation of SET procedures, a new research theme has emerged in the field of educational psychology and educational sciences. The 'golden age of research' on SET however must be situated in the 1970's, when a lot of research dealt with issues concerning the utility and validity of SET (Centra, 1993).

This has probably much to do with what the American sociologist Harold L. Wilensky (1964) famously described as the professionalization of everyone. With the advance of experts, he argued, the profession became a pervasive occupational model that was spreading well beyond traditional professional areas, such as medicine and law (Wilensky 1964). However, the distinction between professionals and their public was also redefined. The public became more influential. In this sense, the growing interest in SET signals an evolution towards more inclusive forms of (higher) education. The interest in professionalization is complemented by an ‘upgrading’ of the public. SET signals an active concern with inclusion, with the expansion of an upgraded membership of students in higher education. The forms of ‘acceptance’ of students in the relevant setting of higher education changed: students have become entitled to evaluate the professionalism of their teachers, they are given a ‘voice’ in decision-making processes. The relationship between professors and students can no longer remain an asymmetrical, hierarchical one. SET procedures are an indication of the ‘upgraded’ roles for the “laity”, for the public (similar to, for example, the demands for internal democracy in the churches, for giving the laity a ‘voice’ in the ecclesiastical decision-making bodies).

At the moment, many comprehensive reviews on this subject are available (see a.o. Costin, Greenoug, & Menges, 1971; Marsh, 1984, 1987; McKeachie, 1979; Onwuegbuzie et al., 2009; Spooren, Brockx & Mortelmans, 2013; Wachtel, 1997). These studies often present a state of the art of the research on SET, thereby offering a structured overview of the (both classic and recent) research themes and study results in the field. But these reviews are often written from a particular point of view. The inclusion of a particular study in such a review mainly depends on its authors as they decide on several grounds (with respect to content and/or more technical aspects) whether or not they will mention a certain study in their work (i.e., research theme, methods, suited for the review theme, publication date, publication type,

etc.). Still, this is only one way to evaluate the impact of a certain study on its research area. Another possibility, which is of a more quantitative nature, is performing a citation analysis to map out the most important studies in the field (Moed, 2005). The scholarly impact of a certain study is then measured by means of calculating the number of times this study article is cited by other studies. In this view, highly cited studies can be considered as studies that receive(d) much attention and, as a consequence, are in the center of the debate.

The quantification of scholarly impact has been significantly affected by the development of academic search databases such as Web of Science (WoS). Over the past decades WoS evolved from an early idea by Eugene Garfield to one of the most widely used citation-indexing services. Today it counts several thousand of institutional subscriptions and contains more than a billion citations (Clarivate Analytics, 2017). Around the turn of the century also other databases entered the market of bibliometric analysis. In 2003 Elsevier, one of the world's largest publishers of academic journals, launched another subscription-based tool named Scopus. One year later (November 2004) Google introduced Google Scholar (GS), a freely accessible software system that searches for scholarly literature on the Internet. A number of studies soon appeared on the relative pros and cons of each of these interdisciplinary, and other more specialized, citation indexes such as PubMed (in the sphere of biomedical research and the life sciences) or SciFinder (in the domain of chemistry). In addition, there has been some debate in the literature on the need of using more than one database when assessing the impact of research output (see e.g. Harzing & Alakangas 2016). Previous research suggests that the use of GS and Scopus, in addition to WoS, may provide us with more accurate results (see e.g. Meho & Yang 2007). This may be especially true for the field of education where the correlation between citations in WoS and other databases such as GS seems to be lower than in other academic fields (Kousha & Thelwall 2009). Moreover, although WoS is an extensive database some journals slip through the cracks. This

certainly applies to (national) journals in which the articles are not written in English (a). Related and other criticisms directed at the databases of the Institute for Scientific Information (ISI) concern the arguments that: (b) only a small percentage of educational journals is indexed in ISI (Corby, 2001; Fairbairn et al., 2008); (c) ISI citation measures and journal impact factors can be manipulated (for instance, by a high number of self-cites); (d) citation numbers are based on cites in other ISI-journals only; and (e) particular types of studies have higher chances to be published in ISI journals (for instance, quantitative studies).

Some of these criticisms might be familiar to educational researchers. For example, while the first issue of *Assessment & Evaluation in Higher Education*, a peer-reviewed journal that publishes many studies on SET, appeared in 1975 it was only recently (2008) that it became included in Web of Science (WoS). Another example is the extensive review on SET findings by Marsh (1987) that was published in *International Journal of Educational Research* and still serves as a strong introduction into the SET literature (+1100 citations in Google Scholar yet not included in WoS).

Against this background, in this study the exploration of high impact scholarship in research on SET will be based on three databases: GS, Scopus, and WoS. Analyzing SET research from this comparative and quantitative point of view allows us to reflect on the way in which the research field has been constituted over the past decades. It allows us to map the development of the research field in a more 'distanced' way. It may also lead to a renewed reflection on the relationship between educational research and educational practice.

Objectives

The present study aims at deepening our insights into the history of SET research by selecting and exploring those 'classic' studies that have (or had) a large impact on the work of past, present and (perhaps) future SET researchers. In addition, a concise thematic analysis of the

selected studies will reveal the dominant research streams in SET research thus far. This would help both policy makers and researchers to summarize the important conclusions in SET research and provide a basis for developing ideas for future research. It might also enhance the development of a more reflective orientation among SET researchers. The research questions for this review were:

- (a) Which are the most frequently cited articles in the field of SET?
- (b) Who are the authors of these articles?
- (c) In which journals were these articles published?
- (d) Which SET topics are discussed in these articles?
- (e) To what extent are the results on high impact studies in the field of SET converging between Google Scholar, Scopus, and Web of Science?

In the following sections, we discuss the methods we used to identify the studies with the highest impact and we provide a brief overview of both the content and the most important conclusions from these studies. Finally, we discuss the implications of our study for future research in the field of SET.

Method

Overview

In June 2016, we searched for journal articles in Google Scholar, Scopus, and the Web of Science database (Social Science Citation Index) to identify the most cited articles in the SET research field. This resulted in a list of the top 50 articles for each of these three databases. The final database consisted of 75 pieces, including several types of articles (review articles, empirical studies, etc.).

We are aware of the limitations of our quantitative approach. For example, it is sometimes argued that citation analyses exclude certain type of documents such as books or so-called grey literature like workings papers, conference proceedings and reports. However,

we believe it is correct to assume that a significant amount - if not to say a majority - of the influential SET publications concern journal articles. Different from some specialties in the humanities and the social sciences, in the SET-literature, and (educational) psychology by extension, publishing in international journals is the group norm. In order to take into account different types of journal contributions, we included review articles and the more primary empirical analyses, theoretical driven studies as well as work focusing on the statistical and methodological aspects of SET.

Literature Search and Selection Criteria

Given the inconsistent use of terminology concerning SET, the literature search for this study was based on a variety of terms that refer to the concept of SET. The following keywords were used: *student ratings*, *student ratings of instruction*, *student course evaluations*, *student evaluation of teaching*. These key terms were then entered in three major bibliometric databases currently available Google Scholar, Scopus, and the Web of Science. Other selection criteria were that the studies should (a) concern the context of higher education, (b) focus on student evaluation of teaching in a course or a unit (and not in a whole educational program or an institution), and (c) have the use and/or the validity of SET as (one of the) the main subject(s) of the study. The abstracts were screened by the first author and publications that did not fit into these criteria were discarded. In case of doubt, the other authors were consulted.

Results

In each database, the remaining articles were ranked by their number of citations and the 50 articles with the highest number of citations were selected¹. Appendix A provides an overview of all selected articles, ranked by their number of citations in Google Scholar, Scopus, and Web of Science. In the reference list, all of these studies are indicated with an asterix.

In our case the final sample consisted of 75 top ranked articles as measured by total citations in Google Scholar (GS), Scopus (SC), and Web of Science (WoS) (see figure 1). Given a potential range from 50 (complete overlap among the three databases) to 150 articles (no overlap at all) there thus seemed to be a moderate overlap among the three databases. More specifically, from the pool of 75 top ranked articles, more than one third (29/75) appeared in each of the three top fifty hit lists. Some of the recurring journals in which these core articles regularly appear include *Journal of Educational Psychology* and *Research in Higher Education*. Also noteworthy is the inclusion of five publications on student ratings that appeared in the November 1997-issue of *American Psychologist*, the official journal of the American Psychological Association.

¹ As it is not possible to rank papers based on their number of citations in Google Scholar itself, we used the *Publish or Perish 4.0* software (Harzing, 2010) to draw up the Google Scholar list.

Figure 1. Core articles in SET-research: A comparison between Scopus, Google Scholar, and Web of Science.

Two thirds (49/75) of the most frequently cited articles appeared in at least two of the three databases under study (i.e. GS and SC; GS and WoS; SC and WoS; or: GS, SC and WoS). For reasons of convenience we can describe the sum of these subsets as the overlapping core (n=49; see shaded areas in light and dark gray in Figure 1). This also means that one third (26/75) of the top ranked articles are included in just one database (i.e. GS or WoS or SC). The majority of these single hits appeared in Web of Science (11) and Scopus (10). From the top 50 articles in Google Scholar, only 5 did not appear in the top 50 of WoS and SC. In other words, no less than 90% of the articles in GS' top fifty also appeared in either WoS and/or SC. This is slightly higher than the percentages for both SC (81%) and WoS (78%).

The inclusiveness of the three databases becomes further apparent if we look at the number of double hits (i.e. articles that are included in the top 50 of each of the two other databases; see Figure 1, shaded areas in light gray) that would be missing when the analysis

was limited to the top ranked articles in just one database. In the case that we would have based the analysis on GS only, 4 double hits would be missing. This is a little less than in case of SC (6). However, in case of WoS more than twice as many double hits (10) would be missing. Given an overlapping core of 49 articles, the proportional margin of error thus raises from 8,2% (GS), to 12,2% (SC) to 20,4% in case of WoS. Furthermore, it seems that each of the 75 top ranked articles are indexed in GS, if not in its top 50 then at least in its index in general. This is different from both SC and WoS. Whereas 10 out of 75 articles are not included in Scopus' index (13%), the number of absolute missing values raises to 17 articles (23%) in case of WoS.

Some journals thus slip through the cracks and are simply not indexed in Scopus and Web of Science. Other journals, especially in WoS, only get indexed several years or even decades after the publication of its first issues (see e.g. *Research in Higher Education* (first issue in 1973, indexed by WoS in 2000), or *Assessment and Evaluation in Higher Education* (first issue in 1975, indexed by WoS in 2010). On the other hand, it seems that the research output in some of the more institutionalized subdomains are more visible in WoS (especially in comparison with GS). This holds particularly true for articles that were published in journals on medical education such as *Journal of Medical Education* and *Academic Medicine*.

The studies (articles)

Tables 1-3 show the top ten articles in each of the databases (GS, SC, WoS) we used for the present study. We observe considerable overlap as six papers appear in all three lists, five papers are included in two lists, and only two papers occur in only one list.

The most cited paper on SET in Google Scholar (Table 1) is the extensive review study by Herbert Marsh (1987) in the *International Journal of Educational Research*. This paper (containing 133 pages) still serves as a comprehensive introduction to the SET

literature and provides both an overview of the history of SET-research and a review of all (in those days) available studies concerning a number of important themes, including the dimensionality of SET, validity and reliability issues, the so-called ‘witch hunt’ for bias in SET-scores, and the utility of SET. Ramsden’s (1991) piece on the construction and the psychometric properties of the CEQ-instrument is the first empirical study in the list. This instrument was derived from the Course Perceptions Questionnaire (Ramsden, 1979) and is used as an instrument for mapping perceived teaching quality in many institutions to this very day. The third most cited article is another state of the art article by Herbert Marsh (1984), who provided a detailed overview of all SET studies published before 1984 and made a number of suggestions for future SET research. He concluded that (class-average) SET are “(a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases; and (f) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions” (p.707). The top five of the GS-list further consists of Cohen’s (1981) meta-analysis of studies on the relationships between SET-scores and student achievement, a relationship that continues to be at the center of the debate (Brockx, Spooren & Mortelmans, 2011), and a renewed review study by Marsh & Overall (1997) on issues concerning the validity and utility of SET. This paper is one of the five highly cited papers that were published in the special issue on SET in *American Psychologist* (November 1997).

Table 1 about here

With 498 citations, Marsh's 1987-review is the most cited SET-paper in Scopus (Table 2). His two other reviews (1984, 1997) in Google Scholar's top five list are also in the top of the Scopus list, accompanied by Ramsden (1991). Still, it is remarkable to see that a relatively recent publication by Nulty (2008) already received 287 citations in the Scopus database (average cites/year = 35.9). This paper tackles a very important issue in actual SET-practice as it describes several ways to increase response rates in online SET-surveys. After all, in recent years, electronic SET-procedures have become the norm in many institutions (Arnold, 2009). Several studies indicate that electronic surveys yield similar results compared to the more traditional paper-and-pencil questionnaires, although their greatest challenge exists in boosting the response rates (Spooren, Brockx & Mortelmans, 2013). Another highly cited recent paper in the Scopus database is Marsh's article (co-authored with Muthén, et. al) on the construct validity of the SEEQ instrument using the Exploratory Structural Equation Modeling (ESEM) technique (2009) (10th in the list with 214 citations). This study is also frequently cited by authors who are active in other research fields than SET and refer to this study for its contribution to the ESEM technique as a means to explore the validity of survey instruments.

Table 2 about here

Costin, Greenough and Menges' article tops the list with more than 300 citations in the Web of Science (Table3). They wrote a review that included a high number of studies on the reliability and validity of SET and added a supplement by means of an empirical study that measured student's opinions on the value of SET at the University of Illinois. They concluded that SET cannot be used as a sole indicator to evaluate a teacher's teaching contribution and suggested that several other measures of teaching effectiveness should be

taken into account. Although this study was published in 1971, it is still cited on a regular basis in the last decade (i.e., 12 cites in WoS-articles that were published since 2010). One example of the above mentioned hypothesis that research output in the more institutionalized subdomains such as Medicine (i.e., medical education) is more visible in WoS is Irby's paper on the characteristics of (best and worst) clinical teachers in medicine that was published in 1978 in *Academic Medicine*. This paper ranked tenth in the WoS-list with 164 citations, but received relatively much fewer citations in Google Scholar (283 citations, 40th place) and Scopus (154 citations, 17th place).

Table 3 about here

Finally, it is worth mentioning yet not surprising to see that the top ten lists in all databases are dominated by review studies. This is a good example of the scientific method at work in which researchers graft their work upon the shoulders of giants to contribute to the cumulative evidence, thus building a better estimate of what is 'true'. After all, these review studies can be seen as introductions into the field of SET and are used by many researchers to embed their studies within the various research streams in this field. The dominance of review studies in this list also transmits a particular understanding of (the development of) knowledge about SET. It illustrates the additive qualities of knowledge in the field of SET. The impression that is conveyed is that knowledge can be understood as cumulative through the progressive development of knowledge. In addition, new research can be presented as research that adds to the 'knowledge base'.

In sum, our final sample of 75 top ranked papers contains 25 review articles, 49 empirical studies (with all studies using quantitative research techniques), and 2 papers that

should be classified as ‘other’ (i.e., one introduction to a special issue (Greenwald, 1997) and one discussion of the papers in a special issue (McKeachie, 1997))².

Figure 2. The core articles on SET in WoS, GS and Scopus: Years of publication (N=75). 5-yearly moving averages.

Figure 2 displays the years of publication of the core articles in our sample (N=75). Whereas the birth of impactful SET research dates back to the 1960s (Isaacson et al., 1964), the further development and growth can be situated from the 1970s onwards. Particularly, 15 core articles appeared in the 1970s, 18 articles in the 1980s, and 22 articles in the 1990s. Further, we see that 19 publications ever since 2000 are already included in the initial list. This finding is rather remarkable, given the fact that articles of previous decades obviously had more time to gather citations. This has most likely to do with the increasing importance of journal publications (also in the educational sciences) on the one hand, and the increasing importance of SET as a worldwide institutionalized practice in higher education on the other. Overall, however, the data presented in Figure 2 seem proof of the orderly historical growth

² Costin, Greenough and Menges' article (1971) was counted twice as it provides both an extensive review and an empirical study.

of knowledge in the field of SET. We mentioned before that reviews of available research are highly quoted. Older publications are also still receiving much attention. In this sense, these citation practices inscribe a particular line of reasoning in SET research. They ‘conceal’ some regulatory mechanisms and shed light on some of the latent ways in which problems of education and their solutions are re-constructed over time. They show the kind of expectations that exist regarding the ways in which research needs to be conducted and presented, and they illuminate some mechanisms that delimit the range of ‘viable’ possibilities open to SET research and SET researchers.

The authors

Less than half of the publications in our sample are single-authored (34/75, 46,7%). About one third of the publications are co-authored (23/75, 30,7%) and 24% (18/75) of the articles are multiple authored. The articles in our sample were written by 149 authors, implying that there are 2 authors per article on average (149/75). Of these 149 authors only 104 are unique, signifying that several authors are involved in the publication of more than one core publication on SET. Specifically, only 17 authors published more than one core paper on SET. Moreover, our analyses reveal that 8% of the authors (8/104) are responsible for almost half of the core output (34/75). Educational psychologist Herbert Marsh is at the top of the list on SET research. This holds true both in terms of number of (co-)authored core articles (11/75, 15%) as well as in terms of number of citations. Particularly, from a total of 27388 references to the core articles on SET in GS, Marsh is involved in no less than 20% of the cases (5648/27388). Concerning SC and WoS, Marsh is involved in 21,4% and 21,3% respectively. The average cites per year outlined previously reaffirm the central position of his contributions to SET research (see also Appendix A). Table 1 further reveals some of the other central authors in SET research. Next to Marsh, other leading researchers are Feldman,

McKeachie, Greenwald, Centra, Cohen, Abrami, Ramsden, Irby, Roche (as an important co-author with Marsh), and Gillmore. They (co-)authored at least three core papers and received hundreds of citations for these core papers alone.

All in all, the picture this table provides is that of a research field dominated by a few leading scholars. While the most-cited articles are review articles, as mentioned before, it cannot be said that the 'idiosyncratic' interests of these leading scholars dominate the research field. But it is remarkable that but a few scholars dominate the way in which the research field is presented and summarized in GS-, SC-, and WoS-articles – as the cited SET research has been conducted in a period of about half a century.

Table 4

Top authors in the field of SET

Author	N° core articles	First author of core articles	Total times cited (core articles only)			Total times cited/total number of citations to the core articles		
			GS	SC	WoS	GS	SC	WoS
Marsh, H.W.	11	11	5648	1833	1444	20,6	21,4	21,3
Feldman, K.A.	9	9	2903	897	411	10,6	10,5	6,1
McKeachie, W.J.	5	3	1695	308	489	6,2	3,6	7,2
Greenwald, A.G.	3	3	1346	476	353	4,9	5,5	5,2
Centra, J.A.	3	3	706	151	96	2,6	1,8	1,4
Cohen, P.A.	3	2	1819	229	199	6,6	2,7	2,9
Abrami, P.C.	3	2	1319	122	185	4,8	1,4	2,7
Ramsden, P.	3	1	1763	283	198	6,4	3,3	2,9
Irby, D.M.	3	2	646	339	287	2,4	4,0	4,2
Roche, L.A.	3	0	1647	130	187	6,0	1,5	2,8
Gillmore, G.M.	3	0	1027	385	237	3,7	4,9	3,5

Note. total number of citations to the core articles in GS=27388, SC=8578, WoS=6767.

The journals

Looking at the sources in which the most impactful studies on SET have been published, 27 different journals can be identified (Table 5). The journal titles reveal that a majority of these publications take residence in the field of educational psychology, followed by the subfield of medical education. Furthermore, three international journals seem to stand out: 15 core articles appeared in the *Journal of Educational Psychology*, 11 papers in *Research in Higher Education* (from which 9 papers were authored by Kenneth Feldman) and 7 papers in

Assessment and Evaluation in Higher Education. In other words, almost half of the high impact articles in our sample (33/75 or 44%) are published in three journals only. This is followed by 5 publications in the *American Psychologist*, which is due to the highly cited special issue on the validity of SET in 1997, and 5 publications in *Academic Medicine*.

Table 5

Sources of high impact articles (N=75)

Journal	# core articles
Journal of Educational Psychology	15
Research in Higher Education	11
Assessment and Evaluation in Higher Education	7
American Psychologist	5
Academic Medicine	5
American Educational Research Journal	4
Review of Educational Research	3
Higher Education	2
Journal of Higher Education	2
Journal of Medical Education	2
Science	2
Studies in Higher Education	2
Academe - Bulletin of the AAUP	1
American Journal of Distance Education	1
British Journal of Educational Psychology	1
Communication Education	1
Computers & Education	1
Economics of Education Review	1
International Journal of Educational Research	1
International Journal on E-Learning	1
Journal of General Internal Medicine	1
Journal of Personnel Evaluation in Education	1
New Directions for Institutional Research	1
Organizational Research Methods	1
Review of Research in Education	1
Structural Equation Modeling – A Multidisciplinary Journal	1
Teaching and Learning in Medicine	1

Research topics

We read all top-cited articles in order to determine the most important research topics in the SET-literature thus far. A coding scheme was developed based on an initial screening of all articles by the first author, and related to article type (review, empirical study, other) and research topic (validity issues, SET instruments, use of SET in higher education practice).

The coding scheme thus consisted of nine (3 x 3) categories. Next three authors

independently coded these articles using this coding scheme. Inter-rater reliability was calculated by means of Krippendorff's alpha using the KALPHA-macro in SPSS (Hayes & Krippendorff, 2007). This test provided an alpha-value of .88 (LL95%CI=.82; UL95%CI=.93) suggesting an acceptable level of inter-rater agreement (Krippendorff, 2004). When an article received different codings, all three coders discussed this paper and jointly assigned it to a category. Some papers tackled more than one of the above mentioned topics and were therefore assigned to two or more categories.

Although it is not our intention to discuss each of these studies in detail, we provide a concise review of the main topics which include (a) validity issues concerning SET, (b) the construction and validation of SET-instruments (including the dimensionality debate), and (c) the use of SET when evaluating teaching performance. In each of these topics we discuss articles of all publication types (review articles, empirical studies, other studies).

Validity issues concerning SET. The validity of SET scores is, as we expected, the main research theme in the field of SET. 18 out of 25 review studies on our list focused on the validity of SET as a means to evaluate teaching effectiveness. Whereas many of these studies provide a narrative review on several validity issues, seven meta-analyses address the discriminant validity and convergent validity of SET as they focus on its relationships with one particular variable at the student level (e.g. student achievement, student learning), the teacher level (teacher's gender) and the class-level (e.g. class size). Also, 33 highly-cited empirical studies include the relationships between SET and one or more variables such as student achievement, student's expected grades, student's gender, student's prior interest, student's experience, teacher's gender, teacher's behavior, teacher's personality traits, teacher's physical attractiveness, teacher's research productivity, course workload, and course type. Many of these studies fit in the 'bias literature', i.e. the search for factors that are

unrelated to good teaching yet affect SET scores (Centra & Gaubatz, 2000). Whereas correlations between SET and certain characteristics such as teacher's gender or teacher's physical attractiveness should be considered as bias as these characteristics have nothing to do with good teaching, for other factors it is not clear whether they can be defined in the same way. The (in most studies positive) relationships between student achievement or student's expected grade and SET scores, for instance, can be seen as bias (as a result from grade inflation or grading leniency, see a.o. Greenwald & Gillmore, 1997 in our sample) but could also support the validity of SET, suggesting that good teaching leads to good learning (i.e., higher grades) and, as a consequence, higher SET scores (see a.o. Marsh & Roche, 2000; Centra, 2003 in our sample). Due to both the mixed results from such studies and the great variety in explanations for the observed relationships, bias studies receive much attention and certainly will continue to be in the center of the SET debate. Still, several authors found that the effect of such possible biasing factors appears to be rather small (Beran & Violato, 2005; Marsh & Roche, 2000; Spooren, 2010; Smith et al., 2007), which strengthens the argument to pay much more attention to other issues that might challenge the validity of SET-scores (i.e., poorly constructed questionnaires, mis-collection of data, inappropriate or ill-considered use of SET-scores) in future research and practice.

SET instruments. Five review studies and fifteen empirical studies deal with the construction and validation of SET-instruments. Much attention goes to (a) the number and the nature of dimensions of effective teaching than can and should be captured by such instruments, and (b) the results of procedures aimed at validating both new and existing SET-instruments, with Marsh's SEEQ and Ramsden's CEQ as the best documented questionnaires. These procedures included advanced quantitative research techniques such as (exploratory) structural equation modeling, (exploratory and confirmatory) factor analyses, and mixed-

method analyses. Still, these studies show mixed results concerning both the dimensions in SET-forms and the outcomes of validation procedures. This is largely due to the absence of a common conceptual framework concerning effective teaching upon which these instruments can be grounded (Devlin & Samarawickrema, 2010; Ory & Ryan, 2001; Penny, 2003). This forces questionnaire designers to rely heavily on both their own (or their institution's) views with respect to good university teaching and relatively data-driven decisions on the number and the nature of dimensions that are captured in their SET-instruments. This calls into question the structural validity of these instruments as they do not guarantee that effective teaching is (fully) represented in the items (Onwuegbuzie et al., 2009). Another important discussion includes the need of institutional boards to get a quick view on a teacher's teaching practice by means of, for instance, a single, global score (Apodaca & Grad, 2005). Although it is widely accepted that SET should be considered multidimensional (given that teaching consists of many aspects) and that SET instruments should capture this multidimensionality, some authors successfully explored the reliability of single item-type SET-instruments (see a.o. Wanous & Hudy, 2001 in our sample). Marsh (1987) however argued that single-item ratings are, in many cases, less accurate and less valid than multi-item scales. The challenge, then, is to develop short SET-instruments that consist of questions on several important dimensions of instruction and are tested extensively on both validity and reliability.

Use of SET. Although almost all highly-cited papers contain conclusions or guidelines for the implementation of SET in higher education practice, twelve studies explicitly focus on the use of SET. McKeachie (1979, 1997) insists on a proper collection and use of SET data when evaluating teaching performance. After all, even if all of the above mentioned potentially biasing variables are under control and even if SET-instruments provide valid scores

concerning the quality of teaching, it is still possible for SET-scores to be administered and used in inappropriate ways. This undermines the potential value of SET to improve teaching quality (Penny, 2003). Although misuse and mis-collection of data might have serious consequences for both formative and summative evaluation of the teachers involved, little research is available concerning this topic.

Other authors focused on the effect of SET, combined with other activities (such as consultancy or self-assessment) or not, on future teaching practice. The results of these studies indicate that providing feedback by SET reports alone (without consultation) is far less effective. Kember, Leung & Kwan (2002) found no evidence that SET procedures improved the overall quality of teaching as perceived by the students. Marsh & Roche (1993) showed by means of an experimental design that, although SET scores improved over time in all groups, teachers who received consultation with respect to their SET scores at the end of the semester improved more than teachers who received no consultation at all. Cohen (1980) concurred by finding that feedback had a modest but significant effect on improving instruction (i.e., instructors who received mid-semester feedback had higher overall ratings than did instructors receiving no feedback).

Other highly-cited studies with respect to the use of SET include the traditionally low response rates in online SET (compared to the more classic paper-and-pencil methods). Dommeyer et al. (2004) compared SET that were completed in-class with those collected online. They found that response rates to the online survey were indeed lower than those in-class, and that a grade incentive used to encourage response to the online surveys led to response rates that were comparable with those to the in-class surveys. Still, online evaluations did not produce significantly different mean evaluation scores than the in-class evaluations (even when different incentives were offered to the students to achieve higher response rates). Nulty (2008) reviewed studies related to survey response rates and then

suggested a number of guidelines to help achieve acceptable response rates. These guidelines include (1) the use of multiple methods to get survey response rates as high as possible, (2) considering the effects survey design and method might have on the make-up of the respondents, and (3) use multiple methods of evaluation rather than relying on one method (for instance, an online survey) alone.

Discussion and Conclusion

The aim of the present study was to deepen our insights into the history of SET research by selecting and exploring a number of studies that have (or had) a large impact on the work of SET researchers. By means of a citation analysis, we identified the most cited articles in the SET research field. This procedure was followed in three databases (Google Scholar, Scopus, and Web of Science) and resulted in a list of 75 articles. The use of three major databases allowed us to map some basic structures of SET research. In line with previous research, our findings suggest that it is indeed advisable to use more than one database to assess the impact of research output. Specifically, one third of the selected articles are included as a top-ranked article in only one database, while two thirds appeared in at least two of the three databases under study. Also, when relying on Web of Science or Scopus alone, a reader would miss 23% (17 papers) or 13% (10 papers) respectively from our sample as these studies (although highly cited in Google Scholar) are not included in these databases. Moreover, in the field of SET research, Google Scholar appeared to be a valuable repository of knowledge, since Pearson's correlations of citations between Google Scholar and Scopus ($r=.84$) and Web of Science ($r=.78$) in our sample are robust. Future research exploring the contrast of Google Scholar, Web of Science and Scopus with other widely used databases such as ERIC or PSYCINFO might be useful. For instance, it has been shown previously that Google Scholar may yield more scholarly content than library databases (Howland et al., 2009). Extending

this research to other important research themes in the field of education would be of great help when scholars need to prove the value of their research.

A thematic analysis revealed three major topics in the field, including (a) validity issues concerning SET, (b) the construction and validation of SET instruments (including the dimensionality debate), and (c) the use of SET when evaluating teaching performance. There is a strong interest in the improvement of education in this research literature. It is fueled by the belief that progress can be obtained by remaking education on the basis of scientific or 'objective' knowledge. The findings of this research are used to organize the implementation of SET in educational practice. They create possibilities for administering higher education in particular ways; they define the ways in which students can have an influence on the educational practices in which they have to take part.

Herbert Marsh has been found to be the most impactful author in the field with 11 (co-) authored papers. His papers that were included in the list also received the highest number of citations. Apart from the highly cited special issue on the validity of SET in *American Psychologist* (1997), the journals *Journal of Educational Psychology* (15 papers), *Research in Higher Education* (11 papers), and *Assessment and Evaluation in Higher Education* (7 papers) published several impactful studies and can be considered important sources for information on the research traditions and results in the field of SET.

Our analysis further reveals that some of the most impactful studies on SET research date back to the 1960s and that its coming of age situates in the (golden) seventies. Since then the evaluation of teachers by students - as an institutionalized practice in higher education and as a research domain - became increasingly visible upon the front stage. The stable and relatively high proportion of impactful articles since 2000 indeed suggests a trend of continuous growth in SET research. At the same time, the historical knowledge in the form of classic studies on SET lives on through the many citations in recent studies.

With the present paper we hope to have contributed to this already extensive body of knowledge on SET by selecting and characterizing some of its most impactful research articles in the last half a century. Future research can elaborate on our work by looking at other than journal publications such as books. Future reflexive studies on SET could also identify patterns of consensus and conflict within SET, as well as benefit from a longitudinal thematic analysis. This might contribute to a better, reflective understanding of the main dynamics and limitations of the research field. Finally, in order to determine the impact of SET beyond the inner circle, we believe that it is worthwhile to further examine the diffusion of SET research to (sub)disciplines other than its close neighbors.

References

- *Abrami, P.C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, 82, 219-231. doi:10.1037/0022-0663.82.2.219
- * Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research*, 52, 446-464. doi: 10.2307/1170425
- *Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13, 153-166. doi:10.1023/A:1008168421283
- 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. doi:10.1080/03075070500340101
- *Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, 87, 656-665. doi: 10.1037/0022-0663.87.4.656
- *Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, 79, 308-314. doi: 10.1037/0022-0663.79.3.308
- * Beckman, T. J., Ghosh, A. K., Cook, D. A., Erwin, P. J., & Mandrekar, J. N. (2004). How reliable are assessments of clinical teaching? *Journal of General Internal Medicine*, 19, 971-977. doi: 10.1111/j.1525-1497.2004.40066.x
- *Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74, 170-179. doi: 10.1037/0022-0663.74.2.170

- 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. doi:10.1080/02602930500260688.
- 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. doi: 10.1007/s11092-011-9126-2
- *Bryant, J., Comisky, P. W., Crane, J. S., & Zillmann, D. (1980). Relationship between college teachers' use of humor in the classroom and students' evaluations of their teachers. *Journal of Educational Psychology*, 72, 511-519. doi: 10.1037/0022-0663.72.4.511
- *Bolliger, D. U. (2004). Key factors for determining student satisfaction in online courses. *International Journal on E-learning*, 3, 61-67.
- *Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, 14, 17-24.
doi:10.3102/00028312014001017
- Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco, CA: Jossey-Bass.
- *Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44, 495-518.
doi:10.1023/A:1025492407752.
- *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. doi: 10.1080/02602930301683

Clarivate Analytics (2017). Web of Science fact book. Retrieved from www.clarivate.com

*Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321-341. doi: 10.1007/BF00976252

*Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309. doi:10.3102/0034654305100328

*Copeland, H. L., & Hewson, M. G. (2000). Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. *Academic Medicine, 75*, 161-166. doi:10.1097/00001888-200002000-0001

Corby, K. (2001). Method or madness? Educational research and citation prestige. *Portal: Libraries and the Academy, 1*, 279-288. doi:10.1353/pla.2001.0040

*Costin, F., Greenough, W. T., & Menges, R. T. (1971). Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *Review of Educational Research, 41*, 511-535. doi:10.2307/1169890

*d'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198-1208. doi: 10.1037/0003-066X.52.11.1198

* Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*, 611-623. doi:10.1080/02602930410001689171

*Entwistle, N., & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education, 19*, 169-194. doi: 10.1007/BF00137106

- * Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, 4, 69-111. doi: 10.1007/BF00991462
- * Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, 5, 243-288. doi:10.1007/BF00991967
- * Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education*, 6, 223-274. doi: 10.1007/BF00991288
- * Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9, 199-242. doi: 10.1007/BF00976997
- *Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, 21, 45-116. doi: 10.1007/BF00975035
- * Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education*, 28, 291-329. doi: 10.1007/BF01006402
- *Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, 30, 583-645. doi: 10.1007/BF00992392
- *Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34, 151-211. doi:10.1007/BF00992161
- * Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external

- (neutral) observers. *Research in Higher Education*, 30, 137-194. doi:
 10.1007/BF00992716
- *Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science*, 182(4107), 83-85. doi:10.1126/science.182.4107.83
- Galbraith, C., Merrill, G., & Kline, D. (2012). 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. doi:10.1007/s11162-011-9229-0
- *Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52, 1182-1186. doi: 10.1037/0003-066X.52.11.1182
- *Greenwald, A. G. & Gillmore, G. M. (1997) Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1209-1217. doi:10.1037/0003-066X.52.11.1209
- *Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, 89, 743-751. doi:10.1037/0022-0663.89.4.743
- *Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24, 369-376. doi:10.1016/j.econedurev.2004.07.013
- Harzing, A. W. (2010). *The Publish or Perish Book*. Melbourne: Tarma software research.
- Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2): 787-804. doi:10.1007/s11192-015-1798-9

- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89. doi: 10.1080/19312450709336664
- Howland, J., L., Wright, T., C., Boughan, R., A., & Roberts, B., C. (2009). How scholarly is Google Scholar? A comparison to library databases. *College & Research Libraries, 70*, 227-234. doi: 10.5860/0700227
- *Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., Baerwaldt, J. W., & Zinn, K. L. (1964). Dimensions of student evaluations of teaching. *Journal of Educational Psychology, 55*, 344 -351. doi:10.1037/h0042551.
- * Irby, D. M. (1978). Clinical teacher effectiveness in medicine. *Academic Medicine, 53*, 808-815.
- * Irby, D., & Rakestraw, P. (1981). Evaluating clinical teaching in medicine. *Academic Medicine, 56*, 181-186.
- *Irby, D. M., Ramsey, P. G., Gillmore, G. M., & Schaad, D. (1991). Characteristics of effective clinical teachers of ambulatory care medicine. *Academic Medicine, 66*, 54-55. doi:10.1097/00001888-199101000-00017
- Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*, 419-434. doi:10.1080/713699176
- *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. doi:10.1080/0260293022000009294
- *Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology, 80*, 342-344. doi:10.1037/0022-0663.80.3.342

- Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, 30, 411-433. doi: 10.1111/j.1468-2958.2004.tb00738.x
- *Kulik, J. A. (2001). Student ratings: validity, utility and controversy. *New Directions for Institutional Research*, 27, 9-25. doi:10.1002/ir.1
- *Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, 3, 210-240.
- *Litzelman, D. K., Stratos, G. A., Marriott, D. J., & Skeff, K. M. (1998). Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. *Academic Medicine*, 73, 688-695. doi: 10.1097/00001888-199806000-00016
- *Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal*, 17, 219-237. doi:10.3102/00028312017002219
- *Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52, 77-95. doi:10.1111/j.2044-8279.1982.tb02505.x
- *Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, 75, 150-166. doi:/10.1037/0022-0663.75.1.150
- *Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, 76, 707-754. doi:10.1037/0022-0663.76.5.707
- *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. doi:10.1016/0883-0355(87)90001-2

- *Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83, 285-296. doi:10.1037/0022-0663.83.2.285
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York, NY: Springer.
- * Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness: Complementary, antagonistic, or independent constructs? *The Journal of Higher Education*, 73, 603-641. doi: 10.1353/jhe.2002.0047
- *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, 439-176. doi:10.1080/10705510903008220
- *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. doi:10.3102/00028312030001217
- *Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias and utility. *American Psychologist*, 52, 1187-1197. doi:10.1037/0003-066X.52.11.1187
- *Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology*, 92, 202-228. doi:10.1037/0022-0663.92.1.202
- *McKeachie, W.J. (1979). Student ratings of faculty – Reprise. *Academe*, 65, 384-397.

- *McKeachie, W.J. (1997) Student ratings: the validity of use. *American Psychologist*, 52, 1218-1225. doi:10.1037/0003-066X.52.11.1218
- *McKeachie, W. J., Lin, Y., & Mann, M. (1971). Student Ratings of Teacher Effectiveness: Validity Studies. *American Educational Research Journal*, 8, 435-445.
doi:10.3102/00028312008003435
- *Min, Q., & Baozhi, S. (1998). Factors affecting students' classroom teaching evaluations. *Teaching and Learning in Medicine*, 10, 12-15. doi: 10.1207/S15328015TLM1001_3
- Meho, L.I., & Yang, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58, 2105-2125.
doi:10.1002/asi.20677.
- Moed, H. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer.
- *Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75, 138- 149.
doi:10.1037/0022-0663.75.1.138.
- *Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, 82, 250- 261. doi:10.1037/0022-0663.82.2.250
- *Naftulin, D. H., Ware Jr, J. E., & Donnelly, F. A. (1973). The Doctor Fox Lecture: a paradigm of educational seduction. *Academic Medicine*, 48, 630-635.
- *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. doi: 10.1080/02602930701293231

- 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. doi:10.1007/s11135-007-9112-4
- Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, 109, 27-44. doi:10.1002/ir.2
- *Ozkan, S., & Koseler, R. (2009). Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers & Education*, 53, 1285-1296. doi:10.1016/j.compedu.2009.06.011
- Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, 8, 399-411. doi:10.1080/13562510309396
- Ramsden, P. (1979). Student learning and perceptions of the academic environment. *Higher Education*, 8, 411-427. doi:10.1007/BF01680529
- * Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, 16, 129-150. doi: 10.1080/03075079112331382944
- * Ramsden, P., & Moses, I. (1992). Associations between research and teaching in Australian higher education. *Higher Education*, 23, 273-295. doi: 10.1007/BF00145017
- * Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, 30, 387-415. doi:10.1080/02602930500099193.
- * Rodin, M. & Rodin, B. (1972). Student evaluations of teachers. *Science*, 177, 1164-1166. doi: 10.1126/science.177.4055.1164

- * Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25, 397-405. doi:10.1080/713611436
- Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of Communication. *Women's Studies in Communication*, 30, 64-77. doi:10.1080/07491409.2007.10162505
- 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. doi:10.1016/j.stueduc.2011.02.001
- 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, 598-642. doi: 10.3102/0034654313496870
- *Sullivan, A. M., & Skanes, G. R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology*, 66, 584-590. doi:10.1037/h0036929.
- * Teven, J. J., & McCroskey, J. C. (1997). The relationship of perceived teacher caring with student learning and teacher evaluation. *Communication Education*, 46, 1-9. doi: 10.1080/03634529709379069
- * Thurmond, V. A., Wambach, K., Connors, H. R., & Frey, B. B. (2002). Evaluation of student satisfaction: Determining the impact of a web-based environment by controlling for student characteristics. *The American Journal of Distance Education*, 16(3), 169-190. doi: 10.1207/S15389286AJDE1603_4

*Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review.

Assessment and Evaluation in Higher Education, 23, 191-210.

doi:10.1080/0260293980230207

*Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension.

Organizational Research Methods, 4(4), 361-375. doi: 10.1177/109442810144003

*Ware Jr, J. E., & Williams, R. G. (1975). The Dr. Fox effect: a study of lecturer

effectiveness and ratings of instruction. *Academic Medicine*, 50, 149-156.

Wilensky, H. L. (1964). The professionalization of everyone?. *American Journal of*

Sociology, 70(2), 137-158. doi: 10.1086/223790

*Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and

application of the Course Experience Questionnaire. *Studies in Higher Education*, 22,

33-53. doi: 10.1080/03075079712331381121

Highlights

- SET (student evaluation of teaching) is used as a measure of teaching performance in almost every institution for higher education
- Since the introduction of SET, thousands of research studies on the validity, reliability and utility of SET were written
- A citation analysis (Google Scholar, Scopus, Web of Science) was used to map the high impact studies, the leading researchers, and the key journals in this research field
- The analysis reveals three main topics in the SET literature: the use of SET, validity issues concerning SET, and the construction and validation of SET instruments
- The high proportion of impactful articles since 2000 suggests a trend of continuous growth in SET research