UNIVERSIDAD
COMPLUTENSE
MADRID

Advanced International Seminar

Assessment of Research and Teaching Outcomes at Higher Education Institutions

Colegio Oficial
de Psicólogos
de Madrid

Facultad de Psicología
Universidad Complutense de Madrid

16-17th November 2015

Monday, November 16th

9:00 Presentation

9:30 John Mingers, Kent Business School, University of Kent, UK
Approaches to evaluating research performance: Peer review and scientometrics – The case of the UK Research Evaluation Framework (REF)

10:30 Coffee break

11:00 Wolfgang Glänzel, Faculty of Economics and Business, KU Leuven, Belgium
Thoughts and facts on bibliometric indicators in the light of new challenges in their applications

12:00 Anne-Wil Harzing, Business School, Middlesex University London, UK
Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison

Tuesday, November 17th

9:30 Philip C. Abrami, Centre for the Study of Learning and Performance, Concordia University, Canada
The assessment of teaching in higher education

10:30 Coffee break

11:00 Pieter Spooren, Department of Social Sciences, Antwerp University, Belgium
On the (in)validity of student evaluations of teaching (SET): A state of the art of the research and suggestions for future practice

12:00 Anthony J. Onwuegbuzie, Department of Educational Leadership, Sam Houston State University, USA
A cross-cultural mixed research meta-framework for assessing teaching effectiveness

(Abstracts appended)

The Spanish Journal of Psychology Ψ

Approaches to evaluating research performance: Peer review and scientometrics – The case of the UK Research Evaluation Framework (REF)

John Mingers
Kent Business School, University of Kent, UK

For many years it was sufficient for academics to communicate their research by publishing in journals and books, and little attention was paid to the quality of the work beyond noting if it was published in a particularly prestigious journal. However, in recent decades this has changed and there is now a huge amount of effort devoted to attempting to evaluate the quality of researchers' work in order to monitor the quality of the researcher and their department or institution. This has resulted in some very elaborate and hugely costly national assessments programmes such as the UK's Research Assessment Exercise (RAE), now called the Research Excellence Framework (REF). This approach has also been adopted by other countries such as Australia and New Zealand. It has huge effects on individuals, departments and universities as a whole, partly in terms of research money but even more so in terms of prestige and league tables. It is only likely to grow in the future. However, there are many problems apart from the cost in these peer review exercises and alternative approaches have been proposed based on quantitative bibliometric techniques although so far these have not replaced peer review. In this talk I will combine together these two strands by firstly reflecting on what I see as the major shortcomings and dysfunctions of peer review, typified by the REF. Then, in the second part, I will try to illustrate how scientometrics has the potential to improve the situation, not by itself but in conjunction with peer review.

Thoughts and facts on bibliometric indicators in the light of new challenges in their applications

Wolfgang Glänzel
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven, Belgium

Research evaluation has become one important but not the only field of application of bibliometric methods. In particular, bibliometrics depicts essential aspects of scientific activities by quantitative and, more specifically, statistical methods, and its output proved to be a valuable supplement to qualitative methods such as peer reviews. Bibliometricians have developed or adopted methods and indicators, i.e., measures of various aspects of research output at different levels of aggregation. In this context bibliometrics is faced with a number of challenges. In the course of this presentation I would like to highlight and discuss four of those.

Bibliometrics/scientometrics has gradually evolved from a sub-discipline of library and information science to an instrument for evaluation and benchmarking. This implies that several scientometric tools became used in a context for which they were originally not designed. The journal impact factor might stand pars pro toto for such tools. The second issue refers to the focus shift away from macro studies down to meso and micro studies. This proved in many regards – not only in terms of statistical reliability of indicators – a true challenge. The third problem emerges from the extension of bibliometric studies beyond their original domain, among others to the social sciences, humanities and the web including social networks. All these developments result in a fourth, more general issue: Tools and indicators need to meet the basic requirements that characterise all scientific methods, namely, meaningfulness, validity, replicability and robustness. In particular, indicators should be insensitive to marginal changes in the aspects they aim to measure, should be meaningful measures of what they are applied to and, of course, under the same conditions and using the same data and methods, the same values and results should be obtained.

The lecture will discuss both the theoretical background and typical applications of these topics. In this context, also a few methodological caveats and pitfalls will be discussed.

Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison

ANNE-WIL HARZING
Middlesex University
The Burroughs, Hendon, London NW4 4BT
Email: anne@harzing.com
Web: www.harzing.com

SATU ALAKANGAS
University of Melbourne
Parkville Campus, Parkville VIC 3010, Australia

Based on a sample of 146 senior academics, we provide a longitudinal and cross-disciplinary comparison of three major bibliometric databases: Google Scholar, Scopus and the Web of Science. Our longitudinal comparison of the rate of growth of publications and citations shows a consistent quarterly growth across all three databases. Our cross-disciplinary comparison of four key research metrics (publications, citations, h-index, and hI,annual, an annualised individual h-index) across five major disciplines (Humanities, Social Sciences, Engineering, Sciences and Life Sciences) shows that the data source and the specific metrics used have a significant influence on cross-disciplinary comparisons.

More specifically, we find that when using the h-index as a metric and the Web of Science as a data source, Life Science and Science academics dramatically outperform their counterparts in Engineering, the Social Sciences and Humanities. However, when using the hI,annual (see Harzing, Alakangas & Adams, 2014) and Google Scholar or Scopus as a data source, Life Science, Science, Engineering and Social Science academics show a very similar research performance; even the average Humanities academic has a hI,annual that is half to two thirds as high as the other disciplines. We thus argue that a fair and inclusive cross-disciplinary comparison of research performance is possible, provided we use Google Scholar or Scopus as a data source, and the recently introduced hI, annual - a h-index corrected for career length and co-authorship patterns - as the metric of choice.

Harzing, A.W., Alakangas, S., & Adams, D. (2014). hIa: An individual annual h-index to accommodate disciplinary and career length differences, Scientometrics, 99(3), 811–821.

The assessment of teaching in higher education

Philip C. Abrami
Centre for the Study of Learning and Performance,
Concordia University, Montreal, Quebec

At many colleges and universities worldwide, an instructor who applies for contract renewal, tenure, or merit needs to provide evidence about his or her teaching. Teaching dossiers developed for this purpose should include a multitude of sources and types of information from a number of courses and over a number of years. These sources may include course outlines, teaching notes, websites and blogs, student examination results and other evidence of student learning, and student ratings of teaching. Evaluation committees are especially impressed when there is consistent and uniform evidence from multiple sources of successful efforts to be an effective instructor.

Often the single most important source of evidence is from student ratings. Because student ratings are gathered after students have had substantial experience with an instructor, because ratings represent the views of a large number of students, and especially because ratings are the single most reliable and valid source of evidence on teaching effectiveness, student rating results are weighted heavily by evaluation committees.

But student ratings need to be *interpreted properly* if they are to be well used. On the one hand, there are critics of ratings who believe they should have no weight in promotion and tenure decisions. And at the other extreme, there are instances where ratings are used as if they were capable of making very fine distinctions among teachers. Neither of these extreme views is correct.

To ignore ratings entirely is to ignore the decades of research and thousands of studies conducted to date on the evaluation of teaching **and the majority view which supports the validity of student ratings**. But to go overboard the other way is also to do injury to the wise use of ratings. In short, *ratings should be used as general guides to teaching effectiveness—great, good, poor—and not more.*

Why? Well, the accumulated evidence tells us that ratings predict teacher-produced student learning imperfectly. The relationship between ratings and learning is moderately positive but that's all. Since ratings are imperfect predictors of a teacher's quality and impact, experts recommend that *we use student ratings to make general judgments and not judgments in tenths or hundredths of decimal points*. This is why multiple sources of evidence accumulated over years is so important.

In this presentation, I will review numerous aspects of the research on student ratings of instruction and ways to use ratings including:

1. Multisection validity studies
2. The Dr. Fox effect or educational seduction
3. The grading leniency phenomenon
4. Other potential sources of bias
5. The dimensionality of student ratings
6. Global ratings for summative purposes
7. Cafeteria approaches for formative purposes
8. Statistical means for interpreting student ratings
9. Electronic teaching portfolios

On the (in)validity of student evaluations of teaching (SET)
A state of the art of the research and suggestions for future practice

Pieter Spooren
Department of Social Sciences, Antwerp University, Belgium

Nowadays, student evaluation of teaching (SET) is used as a measure of teaching performance in almost every institution for higher education throughout the world. Universities and university colleges have developed more or less complex procedures and instruments to collect, analyse and interpret these data as the dominant (and sometimes sole) indicator of teaching quality. This widespread use has much to do with their (apparent) ease of collecting the data and presenting and interpreting the results.

In most institutions, SET is obviously used for formative purposes (e.g., as feedback for the improvement of teaching) as well as for summative purposes (e.g., mapping teaching competence for administrative decision-making and institutional audits). These dual usages—and the unresolved tension between them—makes the use of SET fragile. On the one hand, many teachers are convinced of the usefulness of SET as an instrument for feedback on their teaching. SET results help them to improve the quality of their teaching as it provides them with useful insights in the strengths and weaknesses of their teaching practice, based on student opinions. On the other hand, it is argued that nowadays the principal purpose of SET lies in its use as a measure for quality monitoring, administrative policy-making and mapping whether or not teachers reach a certain required standard in their teaching practice. This justification for using SET in staff appraisals is related to an increasing focus on internal quality assurance and performance management in universities, which have become subject to the demands of consumer satisfaction. Teacher performance and the quality of teaching could be defined as the extent to which student expectations are met, thus equating student “opinions” with “teaching quality”.

For this reason, many faculty members have been questioning the validity and reliability of SET results for many years. 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”, (c) SET procedures and practices (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, et cetera), and (d) the psychometric value of the SET instruments.

This lecture provides a clear idea of the state of the art with regard to research on SET, thus allowing to formulate suggestions for both future research and SET-practice. The utility and validity ascribed to SET should continue to be called into question. Still, the baby should not be thrown out with the bathwater as SET remains a valuable and important source of data about (the quality) of teaching.

A cross-cultural mixed research meta-framework for assessing teaching effectiveness

Anthony J. Onwuegbuzie
Department of Educational Leadership, Sam Houston State University,
Huntsville, TX 77341, USA

Many institutions of higher education worldwide use some type of instrument to assess teaching effectiveness. These instruments are extremely important because administrators often use these instruments to make decisions about tenure, promotion, merit pay increases, and/or the like. Further, these instruments have the potential to provide faculty members with information that can help them optimize their instructional effectiveness in the future. Unfortunately, many of these instruments lack sufficient score validity—whether stemming from insufficient content-related validity, criterion-related validity, and/or construct-related validity. As a result, these instruments are subject to misuse and abuse by administrators. Thus, the purpose of this seminar is multifold. First, I will outline the strengths and limitations of teaching effectiveness instruments. As part of my discussion of measurement issues associated with teaching effectiveness instruments, I will provide evidence of the role that culture (e.g., gender, ethnicity, age) plays in the assessment process. Second, I will discuss how teaching evaluation instruments can be misused and abused. In so doing, I will use real data to illustrate several of the identified problems with these instruments. Third, I will provide guidelines for appropriate use and interpretation of scores stemming from teaching effectiveness instruments. In so doing, I contend that teaching effectiveness instruments should never be used in isolation to evaluate instructional effectiveness. Rather, they should be combined with other measures of teaching effectiveness. Finally, and central to my presentation, I will utilize an evaluation meta-framework that I very recently co-developed called a Mixed Methods Theory-Based Impact Evaluation—comprising 8 phases—wherein mixed methods research techniques are used at every phase, to illustrate how to develop an evidence-based teaching effectiveness instrument. I will discuss the implications that this meta-framework has for assessing teaching effectiveness in higher education.