Journal of Documentation

The construct validity of the h-index

Cameron Stewart Barnes

The construct validity of the *h*-index

Cameron Stewart Barnes
University of New England, Armidale, Australia

Received 10 October 2015
Revised 2 April 2016
Accepted 3 April 2016

Abstract

Purpose – The purpose of this paper is to show how bibliometrics would benefit from a stronger programme of construct validity.

Design/methodology/approach – The value of the construct validity concept is demonstrated by applying this approach to the evaluation of the *h*-index, a widely used metric.

Findings – The paper demonstrates that the *h*-index comprehensively fails any test of construct validity. In simple terms, the metric does not measure what it purports to measure. This conclusion suggests that the current popularity of the *h*-index as a topic for bibliometric research represents wasted effort, which might have been avoided if researchers had adopted the approach suggested in this paper.

Research limitations/implications – This study is based on the analysis of a single bibliometric concept.

Practical implications – The conclusion that the *h*-index fails any test in terms of construct validity implies that the widespread use of this metric within the higher education sector as a management tool represents poor practice, and almost certainly results in the misallocation of resources.

Social implications – This paper suggests that the current enthusiasm for the *h*-index within the higher education sector is misplaced. The implication is that universities, grant funding bodies and faculty administrators should abandon the use of the *h*-index as a management tool. Such a change would have a significant effect on current hiring, promotion and tenure practices within the sector, as well as current attitudes towards the measurement of academic performance.

Originality/value – The originality of the paper lies in the systematic application of the concept of construct validity to bibliometric enquiry.

Keywords Measurement, Impact, Bibliometrics, *h*-index, Construct validity, Hirsch index

Paper type Conceptual paper

Introduction

Bibliometrics aims to apply objective, scientific methods to the analysis of citations. Its practitioners frequently express the hope that their discipline will eventually be recognized as a "hard" social science. In this context, it is surprising that most bibliometricians pay so little attention to issues of construct validity. This attitude is in contrast with the practice elsewhere in the social sciences, where researchers place great weight on construct validation when designing new measures. This paper examines how bibliometrics might benefit from a stronger programme of construct validity. As evidence, it illustrates the value of the construct validity approach in the evaluation of a well-known and widely used metric: the *h*-index.

What is the *h*-index?

The *h*-index is a comparative measure of an individual's research impact proposed by physicist Jorge Hirsch. According to Hirsch (2005):

> A scientist has index *h* if *h* of his or her *Np* papers have at least *h* citations each and the other (*Np*−*h*) papers have ⩽*h* citations each (p. 16569).

In layperson's terms, a researcher with ten published articles, each of which has received at least ten citations, has a *h*-index of 10. From the beginning, Hirsch (2005)

Journal of Documentation
Vol. 72 No. 5, 2016
pp. 878-895

0022-0418
DOI 10.1108/JD-10-2015-0127

intended his metric to be used as decision-making tool, expressing the hope that it "may provide a useful yardstick to compare different individuals competing for the same resource when an important evaluation criterion is scientific achievement" (p. 16572).

The influence of the *h*-index on bibliometrics

The *h*-index has been an extremely popular topic of bibliometric research over the last decade. Enthusiasm for the *h*-index among many bibliometricians is so great that some observers have even gone so far as to "divide the research field into a pre and post Hirsch period" (Bartneck and Kokkelmans, 2011, p. 86). The *h*-index has been used in hundreds of studies to measure the research output of individual scientists, research groups, universities and even whole nations (e.g. Jacsó, 2009; Lazaridis, 2010; Prathap and Gupta, 2009). The extension of the metric to topics such as the measurement of journal impact (Schubert and Glänzel, 2007) was perhaps only a matter of time. More striking has been the recent trend to apply the *h*-index to unexpected areas: from the level of research interest in particular diseases and pathogens (McIntyre *et al.*, 2011; Sanni *et al.*, 2013) to the popularity of YouTube channels (Hovden, 2013).

The *h*-index outside bibliometrics

The *h*-index is commonly used as decision-making tool in universities across the globe. There are claims that it has become the "most popular quantitative measure of a researcher's productivity and impact" (Penner *et al.*, 2013, p. 8). The metric is frequently employed to determine the success or failure of grant proposals, the outcome of applications for promotion, fellowship or tenure, and the even level of government funding for institutions (Barnes, 2014). Although this trend has generated unease among some observers (Burrows, 2012), the consensus regarding the *h*-index seems to be "Whether you or I like it or not, it is here to stay" (Schreiber, 2014, p. 9).

Despite its increasing influence, there have been long concerns that the *h*-index has received "little serious analysis" (Adler *et al.*, 2008). With few exceptions, there has been little effort to look more deeply at the metric and its construction. Efforts at validation have been sparse, and have been largely restricted to attempts to show convergent validity, the degree to which the *h*-index appears correlated to other measures (Bornmann and Daniel, 2009; Bornmann *et al.*, 2008).

The *h*-index zoo

The relative lack of interest in fundamental issues is most evident in the proliferation of *h*-index variants. Like other citation-based measures of research impact, the *h*-index suffers from a number of inherent limitations. Hirsch himself readily acknowledges this point (Hirsch, 2005, 2007, 2010; Hirsch and Buela-Casal, 2014). These shortcomings include:

- a built-in bias against early career researchers (Kelly and Jennions, 2006);
- susceptibility to inflation through self-citation (Bartneck and Kokkelmans, 2011; Burrell, 2007; Schreiber, 2007; Zhivotovsky and Krutovsky, 2008);
- the absence of adjustments for multiple authorship (Burrell, 2007; Hirsch, 2010; Schreiber, 2008);
- the lack of any means of field-normalization (Alonso *et al.*, 2009; Batista *et al.*, 2006); and
- the dissimilar *h*-index scores for individuals generated by different citation databases (Jacsó, 2008).

The main effect of these shortcomings has been to encourage an explosion of *h*-index variants. By 2011, there were at least 37 of these (Bornmann *et al.*, 2011). New specimens continue to be added to the "*h*-index zoo" on a regular basis (e.g. Wan, 2014; Zhang, 2013). However, few of these *h*-index variants have attracted much research interest. Even Hirsch's own suggested modification of the *h*-index in order to account for co-authorship (Hirsch, 2010) has failed to excite any serious enthusiasm. Part of the reason is that most of these *h*-index variants are so highly correlated with Hirsch's original index as to be largely redundant (Bornmann, 2012; Bornmann *et al.*, 2011).

Many aspects of this research activity are troubling. The design of new *h*-index variants would be pointless if the original metric itself suffered from fundamental problems. This is a question, however, which is rarely, if ever, raised in this context. Waltman and van Eck (2012) have pointed out that the creation of new *h*-index variants usually proceeds in a relatively ad hoc manner. The inventor of a new variant typically singles out one of the *h*-index's shortcomings, and then proposes a new modification which adjusts for this factor. The new indicator is then justified on the grounds that its application produces results which appear to "be intuitively reasonable" (Waltman and van Eck, 2012, p. 406). Wider issues relating to the metric's validity are almost never addressed.

The same ad hoc approach is evident when bibliometricians apply the *h*-index to new uses. Researchers almost never ask whether the metric is fit for purpose. Its relevance to other areas of study is simply assumed. This approach is inconsistent with attitudes in the other social sciences, where investigators are far more cautious in extending the use of even well-established metrics to new purposes.

What is validity?

In the social scientific context, the term validity is often used as a synonym for test validity. This is the extent to which both evidence and theory support the interpretation placed on the test scores. In simple terms, validity is the extent to which an observer can be confident that a proposed measure, when applied to its intended purpose, produces reliable results (Cook and Beckman, 2006). This way of looking at validity has its origins in research into psychological measurement, primarily in the USA during the first part of the twentieth century (Thompson and Daniel, 1996).

Before the 1950s, psychologists and other social scientists usually understood test validity in a relatively narrow sense. The main measure of validity was often criterion validity: how well a test estimated or predicted other outcomes, such as task performance (Cronbach, 1989; Kane, 2001). Researchers also placed considerable weight on content validity: the match between the proposed test and the concept it was designed to assess (Kane, 2001). These different forms of validity are now generally subsumed within a single overarching concept: construct validity (Cronbach and Meehl, 1955; John and Benet-Martínez, 2000; Messick, 1995).

What is construct validity?

The modern idea of construct validity has its origins in the work of Cronbach and Meehl in the mid-1950s (Cronbach and Meehl, 1955). Working in the context of psychological testing, Cronbach and Meehl refocused attention on the question of how investigators could be sure that they were reliably scoring theoretical concepts ("constructs") not directly observable or measurable. Their answer to the question was to emphasize the importance of an underlying nomological network. In simple terms, a

nominological network arises from the results of previous observations. It describes the theory behind the construct under observation, explains how the focal construct is related to allied constructs, and defines exactly how changes in the construct will be reflected in the test scores (Cronbach, 1988, 1989; Cronbach and Meehl, 1955).

Construct validity was initially presented as an alternative to traditional criterion and content validity approaches (Kane, 2001). However, within a few years, it had begun to emerge as "the whole of validity from a scientific view" (Loevinger, 1957, p. 363). During the next decades, previous categories of validity were gradually absorbed under this single framework (Clark and Watson, 1995; Cook and Beckman, 2006; Kane, 2001). The construct validity approach favoured by Cronbach and Meehl is now widely seen as forming the basis of the scientific method within social enquiry (Benson, 1998; Smith, 2005), although other approaches to construct validation are popular, such as the multitrait-multimethod matrix (Campbell and Fiske, 1959).

The strength of the construct validity approach lies in its theoretical rigour. A defining feature of the construct validity approach is what has been described as "appropriate scepticism" (Smith, 2005, p. 397). The case for construct validity depends on an accumulation of evidence. Construct validation:

> […] is not only continuous (a matter of degree, not a categorical distinction between valid and invalid) but continual (a perpetual, self-refining process) (Drew and Rosenthal, 2003, p. 609).

Strong and weak validation programs

Before proceeding, it is useful to consider the distinction between weak and strong programs in construct validation (Benson, 1998; Cronbach, 1988, 1989). The former is generally regarded as unsatisfactory at best. Cronbach (1988) considers that the:

> […] weak program is sheer exploratory empiricism […] The strong program […] calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges (pp. 12-13).

The distinction between strong and weak programs has been described in these terms:

> Strong programs depend on precise theory that leads to specific predictions […] Weak programs, on the other hand, stem from less fully articulated theories and construct definitions. With weak validation programs there is less guidance as to what counts as validity evidence […] One result can be approaches in which almost any statistically significant correlation between a target measure another measure, of any magnitude, can be described as validation evidence (Smith and Zapolski, 2009, p. 85).

The paper will show that an extremely weak programme of construct validation has prevailed in relation to *h*-index research, a set of circumstances which may point towards a more general problem within the bibliometrics.

Is construct validity relevant to the *h*-index?

From a social sciences perspective, it is hard to see why the *h*-index has not been subject to rigorous construct validation. In social scientific terms, an index is "an abstract theoretical construct in which two or more indicators of the construct are combined to form a single summary score" (Carmines and Woods, 2004, p. 485). The *h*-index clearly fits this definition. The construct being measured is research impact, which is self-evidently an abstraction. Moreover, the *h*-index provides a summary score derived from two separate measures: output (number of published articles) and impact (citations).

What is the *h*-index intended to measure?

The first step in assessing the *h*-index in construct validity terms is to define the construct the metric is designed to measure. Unfortunately, Hirsch has not always been clear in regard to this point. The title of his original paper indicated that the goal of his metric was to "quantify an individual's scientific output" (Hirsch, 2005). However, in the paper itself, he focused attention the measurement of a researcher's impact. In his conclusion, Hirsch (2005) wrote that his index provided an "estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions" (p. 16572). In a latter paper, Hirsch states simply that the "*h*-index is an indicator of the impact of a researcher on the development of his or her scientific field" (Hirsch and Buela-Casal, 2014, p. 163). These statements would seem to indicate that the *h*-index is intended to measure research impact as understood by bibliometricians.

It is worth noting that many commentators have suggested a contrary interpretation. The view is often expressed that the *h*-index provides a measure of the quantity and quality of a researcher's publications (e.g. Bornmann and Marx, 2014; Egghe, 2006; Kulasegarah and Fenton 2010; Tol, 2009). This is reasonable assumption in terms of the *h*-index's construction. However, such an interpretation is contrary to Hirsch's intentions. He has explained that his metric:

> […] originates from the assumption that the number of citations received by a scientist is a better indicator of the relevance of his or her work than the number of papers he or she publishes or the journals where they are published. The fact that every paper published has its own number of citations implies having many numbers for each scientist. I had the idea of developing the h-index as a way of condensing all that information into one single number to facilitate comparisons between scientists (Hirsch and Buela-Casal, 2014, p. 161).

This statement leaves little doubt that the purpose of the *h*-index is indeed to measure research impact as bibliometricians understand the concept. The decision to count only a sub-set of citations based on the intersection point between outputs and citations is simply Hirsch's mechanism to facilitate comparisons between different researchers in terms of their citation histories.

What is impact?

If the *h*-index measures research input, then what does this term mean? The problem here is that bibliometricians rarely define the term. Many are quite happy to use the word impact dozens of times in the same paper without further elaboration (e.g. Leydesdorff and Bornmann, 2011). In practice, many bibliometricians write in apparently circular terms: impact is whatever citations measure.

Since the work of Martin and Irvine (1983), many bibliometricians have made a deliberate distinction between the impact (as measured by citations) of a paper and its importance or quality. The latter are characteristics which cannot easily be measured through citations alone. For non-bibliometricians this distinction is disconcerting: if impact is not importance or quality, then what is it? In fact, an excellent definition of impact exists. Decades ago, Eugene Garfield (1979) observed that:

> People talk about citation counts as being a measure of the "importance", or "impact" of scientific work, but those who are knowledgeable about the subject use the words in a very pragmatic sense: what they really are talking about is utility (1979, p. 363).

The utility to which Garfield refers is primarily the usefulness of a work in the context of the production of future scientific outputs, which typically take the form of papers.

Garfield's definition has many advantages. Garfield wrote in the context of the older, normative theory of citations. This is the idea that citations represent the payment of intellectual debts (Kaplan, 1965; Merton, 1973). However, his definition is also compatible with the ideas of those researchers who favour the social constructivist approach: the notion that authors cite the works of others for a range of pragmatic and rhetorical reasons (Cozzens, 1989; Erikson and Erlandson, 2014). Followers of both theories can at least agree that "articles are highly cited if they are useful to a large number of scientists" (Shadish, 1989, p. 415).

Does the *h*-index have face validity?

An initial step in determining construct validity is to look at a proposed metric in terms of its face validity. Face validity is the extent to which a measure seems "on its face" to be a good translation of a particular construct (Eysenck, 2004). If impact is essentially a question of utility, then how well does the *h*-index seem to measure the usefulness of an individual's publications to other researchers? Over the last decade, different commentators have pointed out three main challenges to the *h*-index in terms of face validity:

(1) the theory underlying the *h*-index rests on a simplistic model, one which is clearly inconsistent with the real world (Glänzel, 2006);

(2) the construction of the *h*-index makes no sense in terms of traditional bibliometrics, as almost all the evidence for a researcher's impact is discarded in the process of calculating an individual's *h*-index (Anderson *et al.*, 2008); and

(3) the construction of the *h*-index appears to be inherently arbitrary (Gingras, 2014; Lehmann *et al.*, 2008; Petersen and Succi, 2013; Schreiber, 2013a; Waltman and van Eck, 2012).

The next part of this paper will examine each of these points in turn.

Is the *h*-index based on flawed theory?

Soon after the publication of Hirsch's original paper, it was claimed that there existed a fundamental flaw in Hirsch's argument (Glänzel, 2006; Jensen *et al.*, 2009). Hirsch's (2005) model of the manner in which a researcher accumulates papers and citations rests on the assumption that an individual "produces papers of similar quality at a steady rate over the course of their careers" (p. 16570). On this basis, Hirsch proposes a "simple deterministic model" of a researcher's publication history in which there is a "constant annual growth of both publications and citations" (Glänzel, 2006, p. 316). According to this model, a researcher's total publications and citations increase in a linear fashion, in lock-step with the rise in his or her *h*-index. However, this model is inherently unrealistic. Bibliometricians have known for decades that the publication histories of most researchers do not fall into this pattern.

Hirsch's model departs from reality in a number of respects. First, it ignores the fact that most researchers' publication histories are heavily right-skewed. The typical scientist has a few highly cited papers and many more rarely cited ones (Anderson *et al.*, 2008; Bornmann and Daniel, 2009; Cerchiello and Giudici, 2014; Seglen, 1992). Hirsch's model is also contrary to the accumulated evidence for age-related effects on both citations and publication frequency (e.g. Cole, 1979; Hall *et al.*, 2007; Levin and Stephan, 1991). Finally, Hirsch assumes that, once published, papers continue to

accumulate citation at a steady rate. This assumption runs counter to the overwhelming evidence that citation rates in many disciplines show a rapid decrease over time (Jensen *et al.*, 2009; Kelly and Jennions, 2006; Redner, 2006).

Does the construction of the *h*-index make sense?

The second problem is that the construction of the *h*-index makes no sense in terms of traditional bibliometrics. The sticking point is that:

> An author's *h*-index cannot exceed his/her number of publications and will usually be considerably less. Thus, the vast majority of the hundreds or even thousands of citations that accompany the most highly cited papers effectively contribute zero […] Moreover, articles that have received many citations, but which fall just short of the number required to score for *h* […] also count for nothing in the sense that *h* is not affected by them (Anderson *et al.*, 2008, p. 578).

By design, the *h*-index throws away almost all of the evidence for a researcher's usefulness to his or her colleagues. What is worse, the metric is insensitive to highly cited articles, which for decades have been regarded as the clearest evidence of a researcher's impact (Aksnes, 2006; Anderson *et al.*, 2008; Egghe, 2006).

Hirsch himself defends this insensitivity on the grounds that citations to such articles:

> […] may be inflated by as small number of "big hits", which may not be representative of an individual if he or she is a co-author with many other on these papers (Hirsch, 2005, p. 16569).

However, multiple authorship is probably something of a red herring. As pointed out above, the citation record of the typical researchers is heavily skewed. Hirsch (2005) himself is a case in point. He is the sole author of his own "big hit", his original article on the *h*-index.

Is the construction of the *h*-index arbitrary?

There is a growing consensus that the construction of the *h*-index is essentially arbitrary in logical terms (Gingras, 2014; Lehmann *et al.*, 2008; Petersen and Succi, 2013; Schreiber, 2013a; Waltman and van Eck, 2012). The reason is that Hirsch does not explain why an individual's *h*-index is determined by:

> […] the "number of citations" (y) versus "paper number" (x) curve and the $y = x$ line, which leads to an x-shaped graph (Hirsch and Buela-Casal, 2014, p. 162).

Hirsch fails to explain why this particular intersection point is better than any other, or why it is necessary to graph citations and numbers of papers in this manner. Moreover, in statistical terms, Hirsch:

> […] assumes equality between incommensurable quantities. An author's papers are listed in an order of decreasing citations with paper i having C(i) citations. Hirsch's index is determined by the equality, $h = C(h)$, which posits an equality between two quantities with no evident logical connection (Lehmann *et al.*, 2008, p. 377).

Some critics have gone so far as to claim that confidence in the validity of the *h*-index involves a willing suspension of disbelief. In their report on citation statistics, Adler and his co-authors suggest the following simple thought experiment:

> Think of two scientists, each with 10 papers with 10 citations, but one with an additional 90 papers with 9 citations each; or suppose one has exactly 10 papers of 10 citations and the other exactly 10 papers of 100 each. Would anyone think them equivalent? (Adler *et al.*, 2008, p. 13).

Many observers have pointed out that the *h*-index behaves in a logically inconsistent manner under a range of scenarios (Gingras, 2014; Ravallion and Wagstaff, 2011; Vinkler, 2007; Waltman and Van Eck, 2012). It is asserted that this inconsistency is exactly what we would expect from an index lacking any logical basis for its construction (Gingras, 2014; Waltman and Van Eck, 2012).

Even Hirsch concedes that his metric has problems in identifying highly cited researchers. He admits that: "for an author with a relatively low h that has a few seminal papers with extraordinarily high citation counts, the *h*-index will not fully reflect that scientist's accomplishments" (Hirsch, 2005, p. 16571). However, subsequent studies have revealed that the problem goes far deeper. The issue is not just that the *h*-index cannot "distinguish ground-breaking scientific papers from more conventional scientific studies" (Gaster and Gaster, 2012, p. 630). This would be bad enough. The difficulty is that the *h*-index has extremely weak discriminative power at any point in a researcher's career. There are claims that the *h*-index cannot even "discriminate among average scientists" (Jin *et al.*, 2007, p. 856).

Hirsch's proposed test of the convergent validity of the *h*-index

Convergent validity is often regarded as one of the most important forms of evidence for construct validity. In simple terms, convergent validity is "the degree to which multiple attempts to measure the same concept are in agreement" (Bagozzi *et al.*, 1991, p. 425). There have been a number of studies which have purported to demonstrate the convergent validity of the *h*-index. The first of these is contained in Hirsch's original paper.

In his 2005 article, Hirsch proposed a simple test for convergent validity. He compared the *h*-indexes of two populations: physicists and mathematicians recently elected to the National Academy of Science (NAS) in the USA; and recent Nobel Laureates in Physics. Hirsch found that the median *h*-index for the newly elected members of the NAS was 46 and that 84 per cent of his sample of Nobel Prize winners in Physics had a *h*-index at least 30. On this basis, he claimed that:

> These examples further indicate that the *h*-index is a stable and consistent estimator of scientific achievement (Hirsch, 2005, p. 1657).

A number of commentators have taken Hirsch's conclusions at face value (e.g. Aragón, 2013; Panaretos and Malesios, 2009). However, his argument does not sustain closer examination. It has been observed:

> One can conclude that it is likely a scientist has a high *h*-index given the scientist is a Nobel Laureate. But without further information, we know very little about the likelihood someone will become a Nobel Laureate or a Member of the National Academy, given that they have a high h-index. That is the kind of information one wants in order to establish the validity of the *h*-index (Adler *et al.*, 2008, p. 13).

Hirsch's proposed test of the convergent validity of his metric is particularly unfortunate in another respect. The figures he quotes indicate a surprising degree of overlap in the *h*-indexes of NAS members and Nobel Prize winners. In terms of their ranges, means and standard deviations, the *h*-indexes for the two groups are almost identical (Hirsch, 2005). However, Nobel medallists are worlds apart from NAS members in terms of their scientific reputation. The lack of any noticeable difference between the two groups in relation to their *h*-index scores is exactly what we would not expect if the metric actually possessed convergent validity (Barnes, 2014).

More recent studies of *h*-index's convergent validity

In the years immediately after the publication of Hirsch's (2005) paper, a number of studies were published which argued for a convergent validity on the basis that statistically significant associations existed between the *h*-index and other bibliometric measures, such as total number of citations (Bornmann *et al.*, 2008; Costas and Bordons, 2007; Cronin and Meho, 2006; Saad 2006; van Raan, 2006). This approach soon fell out of favour. It was not long before researchers realised that such study designs were unlikely to provide convincing evidence for the convergent validity of the *h*-index. The reason was that it would be remarkable if no correlation existed:

> Since the h index combines number of publications and citations counts in one single index, very large correlation coefficients between the measures are not surprising (Bornmann *et al.*, 2008, p. 155).

Research interest has therefore shifted to studies of convergent validity based on comparison between *h*-index data and peer judgements. On the whole the results have been relatively unimpressive. Findings include:

(1) mean *h*-indexes for successful applicants for fellowships through the German Boehringer Ingelheim Fonds were higher than the mean *h*-indexes for unsuccessful applicants (Bornmann and Daniel, 2005);

(2) there was a statistically significant correlation between *h*-index scores and selection committee ratings for applicants for the European Molecular Biology Organization (EMBO) long-term fellowships and young investigator programmes: at $r = 0.21$ and $r = 0.28$, respectively (Bornmann *et al.*, 2008);

(3) differences in *h*-indexes explained 37 per cent ($r^2 = 0.376$) of the variance in ratings assigned to 163 South African botanists and zoologists by the National Research Foundation (Lovegrove and Johnson, 2008);

(4) variations in *h*-index totals explained 21 per cent ($r^2 = 0.2161$) of the variance in the composite ratings assigned to 147 chemistry research groups in Europe (van Raan, 2006);

(5) French researchers within the Centre national de la recherche scientifique system with higher *h*-index scores were more likely to be promoted than their peers (Jensen *et al.*, 2009);

(6) success in obtaining at least one Institutes of Health (NIH) grant among academic radiologists in the USA was associated with a higher *h*-index, but that overall "funding status and related metrics were not highly correlated with *h*-index" (Rezek *et al.*, 2011, p. 1339);

(7) mean *h*-indexes for academic ophthalmologists in the USA who were successful in applications for NIH grants were higher than the means for unsuccessful applicants (Svider *et al.*, 2014);

(8) mean *h*-index scores for clinical anesthesiologists (Bould *et al.*, 2011), urologists (Benway *et al.*, 2009) and otolaryngologists (Svider *et al.*, 2013) rose with academic rank; and

(9) academic rank accounted for 33.3 per cent of the variance in general surgeons' *h*-index scores (Sharma *et al.*, 2013).

In every case but one (Rezek *et al.*, 2011), the authors of these studies have reported their results as evidence of convergent validity. However, such interpretations are problematic on a number of grounds.

The willingness of many researchers to see such low or moderate correlations as evidence of convergent validity indicates a misunderstanding of the concept. If observed correlations between two tests which are presumed to measure an identical construct are moderate or low, there are two likely explanations. The first is that the results indicate that the metric under observation measures something other than the desired construct. The second is that the study design may be flawed: either the test under observation is being compared to one that measures a different construct, or there are other experimental factors affecting the outcome (Podsakoff *et al.*, 2009). In either case, such results should be regarded as doubtful evidence of construct validity. Both explanations may be relevant to the papers in question.

Taken at face value, the results of these studies should probably be seen as evidence against convergent validity. The mere existence of a statistically significant correlation between *h*-index scores and expressions of peer esteem tell us little on its own. Convincing evidence for convergent validity depends on showing that two separate measures are highly correlated. For example, where Pearson's correlation coefficient is employed, observed values of at least $r = 0.70$ are desirable (Carlson and Herdman, 2012). Some researchers insist on even higher correlation coefficients in order to allow for experimental error (Podsakoff *et al.*, 2009).

In contrast, the authors of many of these studies are prepared to accept extremely weak associations as evidence of convergent validity. Vinkler (2007) has pointed out that although Bornmann and his team (Bornmann and Daniel, 2005) show a difference in means between successful and unsuccessful applicants for fellowships, the difference is extremely small. He observes that:

> The difference in means given is, however, not significant in each year at $p < 0.05$ level. For example in 1992: the mean *h*-index for researchers awarded was 2.92 ($n = 13$; SD 2.29) and for researchers rejected: 2.70 ($n = 57$; SD 2.17), significance between the two means was: $p = 0.80$. In 1994: 2.83 ($n = 12$; SD = 1.27) and 2.46 ($n = 52$; SD = 2.11), respectively; $p = 0.56$ (Vinkler 2007, p. 490).

Similar criticisms, for example, can be made of the conclusion drawn by Bormann and his team in relation to the association between *h*-index scores and EMBO selection committee ratings. Their study argues that values as low as $r = 0.21$ and $r = 0.28$ show convergent validity (Bornmann *et al.*, 2008). However, these figures indicate that *h*-index scores are extremely poor predictors of EMBO selection committee ratings in practice. The reason is that:

> […] a correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is a 100% perfectly predictable relationship between X and Y, correlation of 0.5 does not mean that you can make predictions with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus a correlation of $r = 0.5$ means that one variable partially predicts the other, but the predictable portion is only $r^2 = 0.5^2 = 0.25$ (or 25%) of the total variability (Gravetter and Wallnau, 2013, p. 521).

Moreover, it is arguable that the 11 studies listed above do not directly address the convergent validity of the *h*-index. As the *h*-index is intended to measure research impact, any test of convergent validity must be based on comparison with another measure of the same construct. Peer evaluations, career outcomes or ratings from grant

bodies do not self-evidently constitute such measures. They are presumably based on perceptions relating to the importance and quality of an individual's contribution to his or her field, characteristics which, as discussed above, may be imperfectly correlated with impact (Martin and Irvine, 1983).

Finally, there is another, common-sense objection to the assumption that these studies demonstrate the convergent validity of the *h*-index. Such loose associations as they reveal are explicable by a simple counter-hypothesis. Researchers who do not publish (or write only a few, rarely cited papers) will tend to have low *h*-indexes. They are also unlikely to receive fellowships, to be promoted or be highly regarded as researchers by their peers (Barnes, 2014).

The discriminant validity of the *h*-index

Another widely accepted indication of construct validity is discriminant validity. In a test of discriminant validity, one calculates the extent to which a proposed measure correlates negatively with another test that measures an opposite or totally different construct (Vogt, 2007). The *h*-index comprehensively fails such a test. The reason is that the metric, although intended to measure impact, in effect functions as a measure of output (i.e. number of papers). This is due to the fact that a researcher's *h*-index can never be higher than the total number of his or her publications. As a result, an individual's score is more strongly limited by the total number of published papers than the number of citations these papers receive.

This point has been noted by a number of commentators. A decade ago, Kelly and Jennions (2006) pointed out the *h*-index is "closely correlated with total publication output; thus, it will generally result in the same assessment as one based on counting publications" (p. 169). This observation has been confirmed by subsequent studies (e.g. Kulasegarah and Fenton, 2010). In one recent paper, the observed correlation coefficient between the published papers and *h*-index scores for 248 Danish professors in the health-sciences was strikingly high: $r = 0.93$ (Gaster and Gaster, 2012, p. 830). The contrast here with the low correlation coefficients reported in studies arguing for the convergent validity of the *h*-index is worthy of remark.

The predictive validity of the *h*-index

Predictive validity is usually demonstrated by showing that a test is a good predictor of a future outcome (Vogt, 2007). In 2007, Hirsch argued for the predictive validity of his metric on the basis of the publication history of two convenience samples of physicists. He demonstrated that:

(1) physicists with a high *h*-index 12 years after their first publication were still likely to have a high *h*-index 12 years later; and

(2) the *h*-index was better than either a physicist's total number of publications or total number of an author's citations in predicting its future cumulative value (Hirsch, 2007).

Hirsch (2007) therefore concluded that his index was the best method of predicting a researcher's future achievements. This conclusion has been accepted by a number of commentators (Acuna *et al.*, 2012; Benway *et al.*, 2009; Bornmann *et al.*, 2008; Egghe, 2010). However, there are serious problems with Hirsch's argument. The first objection is a technical one. Predictive validity is usually demonstrated by showing a high coefficient of correlation between the test under observation and another indicator.

In contrast, Hirsch is comparing the *h*-index against itself. Most importantly, he ignores the fact that:

> […] the increase of the h-index with time after a given point of time (e.g. the time of appointment or the time of allocating resources) is not necessarily related to the scientific achievements after this date […] the growth of the h-index is the same, irrespective of whether the investigated researcher had performed as he or she did or whether he (she) had not published any further work (Schreiber, 2013b, p. 1).

This observation applies even in relation to Hirsch himself:

> If Hirsch had stopped working in 2001, his index would have been unaffected in 2010 and even in 2012 deviate only by one index point. From 2005 onwards no change would have resulted except a deviation of one index point in the year 2009 (Schreiber, 2013b, p. 3).

In short, the apparent predictive power of the *h*-index is an illusion: "the increase of the *h*-index does not necessarily depend on the factual performance for several years in the future, but is more likely to result from previous, often rather old publications" (Schreiber, 2013b, p. 3). Rather than a reliable predictor of the future, the *h*-index is "clearly a measure of a researcher's past accomplishments" (Penner *et al.*, 2013, p. 1).

In fairness, it should be noted this last characteristic is not unique to the *h*-index. It is inherent in all citation-based metrics. Due to age and career-related effects, past citations are inherently poor predictors of future performance. However, this does not alter the invalidity of Hirsch's proposed test of his metric in this instance.

Conclusion

The *h*-index comprehensively fails any test of construct validity. The metric is arbitrary in its construction, and lacks any convincing theoretical justification. It behaves in a logically inconsistent manner when applied in the real world, and has limited discriminatory power. Claims that the *h*-index has convergent or predictive validity do not sustain close examination, whereas the metric clearly fails the simplest test of discriminant validity.

The aim of this paper is not simply to show that the *h*-index lacks construct validity. Its purpose has been to demonstrate the power of construct validity as a heuristic tool when applied to bibliometric enquiry. When looked at in terms of construct validity, the weaknesses of the *h*-index become glaringly obvious. Tests for convergent validity have been the main tool used to evaluate this metric, yet these tests have been applied with little or no regard to underlying theory of the constructs under examination.

If bibliometrics is to advance as a discipline, its practitioners will benefit from greater emphasis on construct validation. There are many areas where such an approach might be usefully applied. Despite the enormous impact that the *h*-index has made on the field, it is far from the standard indicator in bibliometrics. For most bibliometricians, field-normalized indicators remain the method of choice when used to evaluate researcher impact (Bornmann and Marx, 2014). Despite decades of effort, however, there remain a host of unanswered statistical questions regarding normalization methods (Waltman and van Eck, 2013). Doubt has even been expressed whether normalization is an achievable goal in the real world (Seglen, 1998). If so, the use of field-normalized indicators may be one of many bibliometric practices that might be usefully reviewed from a construct validity perspective.

References

Acuna, D.E., Allesina, S. and Kording, K.P. (2012), "Future impact: predicting scientific success", *Nature*, Vol. 489 No. 7415, pp. 201-202.

Adler, R., Ewing, J. and Taylor, P. (2008), "Citation statistics: a report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)", International Mathematical Union, Berlin.

Aksnes, D.W. (2006), "Citation rates and perceptions of scientific contribution", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 2, pp. 169-185.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E. and Herrera, F. (2009), "h-Index: a review focused in its variants, computation and standardization for different scientific fields", *Journal of Informetrics*, Vol. 3 No. 4, pp. 273-289.

Anderson, T.R., Hankin, R.K.S. and Killworth, P.D. (2008), "Beyond the Durfee square: enhancing the h-index to score total publication output", *Scientometrics*, Vol. 76 No. 3, pp. 577-588.

Aragón, A.M. (2013), "A measure for the impact of research", *Scientific Reports*, Vol. 3, Article No. 1649, pp. 1-5.

Bagozzi, R.P., Yi, Y. and Phillips, L.W. (1991), "Assessing construct validity in organizational research", *Administrative Science Quarterly*, Vol. 36 No. 3, pp. 421-458.

Barnes, C. (2014), "The emperor's new clothes: the *h*-index as a guide to resource allocation in higher education", *Journal of Higher Education Policy and Management*, Vol. 36 No. 5, pp. 456-470.

Bartneck, C. and Kokkelmans, S. (2011), "Detecting *h*-index manipulation through self-citation analysis", *Scientometrics*, Vol. 87 No. 1, pp. 85-98.

Batista, P.D., Campiteli, M.G. and Kinouchi, O. (2006), "Is it possible to compare researchers with different scientific interests?", *Scientometrics*, Vol. 68 No. 1, pp. 179-189.

Benson, J. (1998), "Developing a strong program of construct validation: a test anxiety example", *Educational Measurement: Issues and Practice*, Vol. 17 No. 1, pp. 10-17.

Benway, B.M., Kalidas, P., Cabello, J.M. and Bhayani, S.B. (2009), "Does citation analysis reveal association between h-index and academic rank in urology?", *Urology*, Vol. 74 No. 1, pp. 30-33.

Bornmann, L. (2012), "Redundancies in *h* index variants and the proposal of the number of top-cited papers as an attractive indicator", *Measurement: Interdisciplinary Research and Perspectives*, Vol. 10 No. 3, pp. 149-153.

Bornmann, L. and Daniel, H.-D. (2005), "Does the *h*-index for ranking of scientists really work?", *Scientometrics*, Vol. 65 No. 3, pp. 391-392.

Bornmann, L. and Daniel, H.-D. (2009), "The state of *h* index research", *EMBO Reports*, Vol. 10 No. 1, pp. 2-6.

Bornmann, L. and Marx, W. (2014), "How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations", *Scientometrics*, Vol. 98 No. 1, pp. 487-509.

Bornmann, L., Wallon, G. and Ledin, A. (2008), "Is the *h* index related to (standard) bibliometric measures and to the assessments by peers? An investigation of the *h* index by using molecular life sciences data", *Research Evaluation*, Vol. 17 No. 2, pp. 149-156.

Bornmann, L., Mutz, R., Hug, S.E. and Daniel, H.-D. (2011), "A multilevel meta-analysis of studies reporting correlations between the *h* index and 37 different *h* index variants", *Journal of Informetrics*, Vol. 5 No. 3, pp. 346-359.

Bould, M.D., Boet, S., Sharma, B., Shin, E., Barrowman, N.J. and Grantcharov, T. (2011), "*h*-indices in a university department of anaesthesia: an evaluation of their feasibility, reliability, and validity as an assessment of academic performance", *British Journal of Anaesthesia*, Vol. 106 No. 3, pp. 325-330.

Burrell, Q. (2007), "Should the h-index be discounted?", in Glänzel, W., Schubert, A. and Schlemmer, B. (Eds), *The Multidimensional World of Tibor Braun: A Multidisciplinary Encomium for his 75th Birthday*, ISSI, Leuven, pp. 65-68.

Burrows, R. (2012), "Living with the *h*-index? Metric assemblages in the contemporary academy", *The Sociological Review*, Vol. 60 No. 2, pp. 355-372.

Campbell, D.T. and Fiske, D.W. (1959), "Convergent and discriminant validation by the multitrait multmethod matrix", *Psychological Bulletin*, Vol. 56 No. 2, pp. 81-105.

Carlson, K.D. and Herdman, A.O. (2012), "Understanding the impact of convergent validity on research results", *Organizational Research Methods*, Vol. 15 No. 1, pp. 17-32.

Carmines, E.G. and Woods, J. (2004), "Index", in Lewis-Beck, M.S., Carmines, E.G. and Woods, J. (Eds), *The SAGE Encyclopedia of Social Science Research Methods*, Sage, Thousand Oaks, CA, pp. 486-487.

Cerchiello, P. and Giudici, P. (2014), "On a statistical *h* index", *Scientometrics*, Vol. 99 No. 2, pp. 299-312.

Clark, L.A. and Watson, D. (1995), "Constructing validity: basic issues in objective scale development", *Psychological Assessment*, Vol. 7 No. 3, pp. 309-319.

Cole, S. (1979), "Age and scientific performance", *American Journal of Sociology*, Vol. 84 No. 4, pp. 958-977.

Cook, D.A. and Beckman, T.J. (2006), "Current concepts in validity and reliability for psychometric instruments: theory and application", *The American Journal of Medicine*, Vol. 119 No. 2, pp. 7-16.

Costas, R. and Bordons, M. (2007), "The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level", *Journal of Informetrics*, Vol. 1 No. 3, pp. 193-203.

Cozzens, S.E. (1989), "What do citations count? The rhetoric-first model", *Scientometrics*, Vol. 15 No. 5, pp. 437-447.

Cronbach, L.J. (1988), "Five perspectives on validity argument", in Wainer, H. and Braun, H.I. (Eds), *Test Validity*, Lawrence Erlbaum, Hillsdale, NJ, pp. 3-18.

Cronbach, L.J. (1989), "Construct validation after thirty years", in Linn, R.L. (Ed.), *Intelligence: Measurement, Theory and Public Policy*, University of Illinois Press, Urbana-Champaign, IL, pp. 147-171.

Cronbach, L.J. and Meehl, P.E. (1955), "Construct validity in psychological tests", *Psychological Bulletin*, Vol. 52 No. 4, pp. 281-302.

Cronin, B. and Meho, L. (2006), "Using the *h*-index to rank influential information scientists", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp. 1275-1278.

Drew, W. and Rosenthal, R. (2003), "Quantifying construct validity: two simple measures", *Journal of Personality and Social Psychology*, Vol. 84 No. 3, pp. 608-618.

Egghe, L. (2006), "How to improve the h-index", *The Scientist*, Vol. 20 No. 3, p. 15.

Egghe, L. (2010), "The Hirsch index and related impact measures", *Annual Review of Information Science and Technology*, Vol. 44 No. 1, pp. 65-114.

Erikson, M.G. and Erlandson, P. (2014), "A taxonomy of motives to cite", *Social Studies of Science*, Vol. 44 No. 4, pp. 625-637.

Eysenck, M.W. (2004), *Psychology: An International Perspective*, Psychology Press, Hove.

Garfield, E. (1979), "Is citation analysis a legitimate evaluation tool?", *Scientometrics*, Vol. 1 No. 4, pp. 359-375.

Gaster, N. and Gaster, M. (2012), "A critical assessment of the h-index", *BioEssays*, Vol. 34 No. 10, pp. 830-832.

Gingras, Y. (2014), "Criteria for evaluating indicators", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, MIT Press, Cambridge, MA, pp. 109-125.

Glänzel, W. (2006), "On the h-index – a mathematical approach to a new measure of publication activity and citation impact", *Scientometrics*, Vol. 67 No. 2, pp. 315-321.

Gravetter, F.J. and Wallnau, L.B. (2013), *Statistics for the Behavioral Sciences*, Wadsworth, Belmont, CA.

Hall, B.H., Mairesse, J. and Turner, L. (2007), "Identifying age, cohort, and period effects in scientific research productivity: discussion and illustration using simulated and actual data on French physicists", *Economics of Innovation and New Technology*, Vol. 16 No. 2, pp. 159-177.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences*, Vol. 102 No. 46, pp. 16569-16572.

Hirsch, J.E. (2007), "Does the h index have predictive power?", *Proceedings of the National Academy of Sciences*, Vol. 104 No. 49, pp. 19193-19198.

Hirsch, J.E. (2010), "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship", *Scientometrics*, Vol. 85 No. 3, pp. 741-754.

Hirsch, J.E. and Buela-Casal, G. (2014), "The meaning of the h-index", *International Journal of Clinical and Health Psychology*, Vol. 14 No. 2, pp. 161-164.

Hovden, R. (2013), "Bibliometrics for internet media: applying the h-index to YouTube", *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 11, pp. 2326-2331.

Jacsó, P. (2008), "The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases", *Online Information Review*, Vol. 32 No. 2, pp. 266-283.

Jacsó, P. (2009), "The h-index for countries in Web of Science and Scopus", *Online Information Review*, Vol. 33 No. 4, pp. 831-837.

Jensen, P., Rouquier, J.-B. and Croissant, Y. (2009), "Testing bibliometric indicators by their prediction of scientists promotions", *Scientometrics*, Vol. 78 No. 3, pp. 467-479.

Jin, B., Liang, L., Rousseau, R. and Egghe, L. (2007), "The *R*-and *AR*-indices: complementing the *h*-index", *Chinese Science Bulletin*, Vol. 52 No. 6, pp. 855-863.

John, O.P. and Benet-Martínez, V. (2000), "Measurement: reliability, construct validation, and scale construction", in Reis, H.T. and Judd, C.M. (Eds), *Handbook of Research Methods in Social and Personality Psychology*, Cambridge University Press, New York, NY, pp. 339-369.

Kane, M.T. (2001), "Current concerns in validity theory", *Journal of Educational Measurement*, Vol. 38 No. 4, pp. 319-342.

Kaplan, N. (1965), "The norms of citation behavior: prolegomena to the footnote", *American Documentation*, Vol. 16 No. 3, pp. 179-184.

Kelly, C.D. and Jennions, M.D. (2006), "The *h index* and career assessment by numbers", *Trends in Ecology & Evolution*, Vol. 21 No. 4, pp. 167-170.

Kulasegarah, J. and Fenton, J.E. (2010), "Comparison of the *h* index with standard bibliometric indicators to rank influential otolaryngologists in Europe and North America", *European Archives of Oto-Rhino-Laryngology*, Vol. 267 No. 3, pp. 455-458.

Lazaridis, T. (2010), "Ranking university departments using the mean h-index", *Scientometrics*, Vol. 82 No. 2, pp. 211-216.

Lehmann, S., Jackson, A.D. and Lautrup, B.E. (2008), "A quantitative analysis of indicators of scientific performance", *Scientometrics*, Vol. 76 No. 2, pp. 369-390.

Levin, S.G. and Stephan, P.E. (1991), "Research productivity over the life cycle: evidence for academic scientists", *The American Economic Review*, Vol. 81 No. 1, pp. 114-132.

Leydesdorff, L. and Bornmann, L. (2011), "Integrated impact indicators compared with impact factors: an alternative research design with policy implications", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 11, pp. 2133-2146.

Loevinger, J. (1957), "Objective tests as instruments of psychological theory", *Psychological Reports*, Vol. 3 No. 3, pp. 635-694.

Lovegrove, B.G. and Johnson, S.D. (2008), "Assessment of research performance in biology: how well do peer review and bibliometry correlate?", *Bioscience*, Vol. 58 No. 2, pp. 160-164.

McIntyre, K.M., Hawkes, I., Waret-Szkuta, A., Morand, S. and Baylis, M. (2011), "The H-Index as a quantitative indicator of the relative impact of human diseases", *PloS ONE*, Vol. 6 No. 5, p. e19558.

Martin, B.R. and Irvine, J. (1983), "Assessing basic research – some partial indicators of scientific progress in radio astronomy", *Research Policy*, Vol. 12 No. 2, pp. 61-90.

Merton, R.K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press, Chicago, IL.

Messick, S. (1995), "Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, Vol. 50 No. 9, pp. 741-749.

Panaretos, J. and Malesios, C. (2009), "Assessing scientific research performance and impact with single indices", *Scientometrics*, Vol. 81 No. 3, pp. 635-670.

Penner, O., Pan, R.K., Petersen, A.M., Kaski, K. and Fortunato, S. (2013), "On the predictability of future impact in science", *Scientific Reports*, Vol. 3, Article No. 3052, pp. 1-8.

Petersen, A.M. and Succi, S. (2013), "The Z-index: a geometric representation of productivity and impact which accounts for information in the entire rank-citation profile", *Journal of Informetrics*, Vol. 7 No. 4, pp. 823-832.

Podsakoff, N.P., Whiting, S.W., Podsakoff, P.M. and Blume, B.D. (2009), "Individual- and organizational-level consequences of organizational citizenship behaviors: a meta-analysis", *Journal of Applied Psychology*, Vol. 94 No. 1, pp. 122-141.

Prathap, G. and Gupta, B. (2009), "Ranking of Indian universities for their research output and quality using a new performance index", *Current Science*, Vol. 97 No. 6, pp. 751-752.

Ravallion, M. and Wagstaff, A. (2011), "On measuring scholarly influence by citations", *Scientometrics*, Vol. 88 No. 1, pp. 321-337.

Redner, S. (2006), "Citation statistics from 110 years of physical review", *Physics Today*, Vol. 58 No. 6, pp. 49-54.

Rezek, I., McDonald, R.J. and Kallmes, D.F. (2011), "Is the *h*-index predictive of greater NIH funding success among academic radiologists?", *Academic Radiology*, Vol. 18 No. 11, pp. 1337-1340.

Saad, G. (2006), "Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively", *Scientometrics*, Vol. 69 No. 1, pp. 117-120.

Sanni, S., Safahieh, H., Zainab, A., Abrizah, A. and Raj, R. (2013), "Evaluating the growth pattern and relative performance in Nipah virus research from 1999 to 2010", *Malaysian Journal of Library & Information Science*, Vol. 18 No. 2, pp. 14-24.

Schreiber, M. (2007), "A case study of the Hirsch index for 26 non-prominent physicists", *Annalen der Physik*, Vol. 16 No. 9, pp. 640-652.

Schreiber, M. (2008), "To share the fame in a fair way, h_m modifies h for multi-authored manuscripts", *New Journal of Physics*, Vol. 10 No. 4, pp. 1-9.

Schreiber, M. (2013a), "A case study of the arbitrariness of the h-index and the highly-cited-publications indicator", *Journal of Informetrics*, Vol. 7 No. 2, pp. 379-387.

Schreiber, M. (2013b), "How relevant is the predictive power of the *h*-index? A case study of the time-dependent Hirsch index", *Journal of Informetrics*, Vol. 7 No. 2, pp. 325-329.

Schreiber, M. (2014), "Is it possible to measure scientific performance with the h-Index or with another variant from the Hirsch index zoo?", *Journal of Unsolved Questions*, Vol. 4 No. 1, pp. 5-10.

Schubert, A. and Glänzel, W. (2007), "A systematic analysis of Hirsch-type indices for journals", *Journal of Informetrics*, Vol. 1 No. 3, pp. 179-184.

Seglen, P.O. (1992), "The skewness of science", *Journal of the American Society for Information Science*, Vol. 43 No. 9, pp. 628-638.

Seglen, P.O. (1998), "Citation rates and journal impact factors are not suitable for evaluation of research", *Acta Orthopaedica*, Vol. 69 No. 3, pp. 224-229.

Shadish, W.R. (1989), "Perceptions and evaluations of quality in science", in Ghoulson, B.S., William, R., Neimeyer, R.A. and Houts, A.C. (Eds), *Psychology of Science, Contributions to Metascience*, Cambridge University Press, Cambridge, pp. 383-426.

Sharma, B., Boet, S., Grantcharov, T., Shin, E., Barrowman, N.J. and Bould, M.D. (2013), "The h-index outperforms other bibliometrics in the assessment of research performance in general surgery: a province-wide study", *Surgery*, Vol. 153 No. 4, pp. 493-501.

Smith, G.T. (2005), "On construct validity: issues of method and measurement", *Psychological Assessment*, Vol. 17 No. 4, pp. 396-408.

Smith, G.T. and Zapolski, T.C.B. (2009), "Construct validation of personality measures", in Butcher, J.N. (Ed.), *Oxford Handbook of Personality Assessment*, Oxford University Press, New York, NY, pp. 81-98.

Svider, P.F., Choudhry, Z.A., Choudhry, O.J., Baredes, S., Liu, J.K. and Eloy, J.A. (2013), "The use of the *h-index* in academic otolaryngology", *Laryngoscope*, Vol. 123 No. 1, pp. 103-106.

Svider, P.F., Lopez, S.A., Husain, Q., Bhagat, N., Eloy, J.A. and Langer, P.D. (2014), "The association between scholarly impact and national institutes of health funding in ophthalmology", *Ophthalmology*, Vol. 121 No. 1, pp. 423-428.

Thompson, B. and Daniel, L.G. (1996), "Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines", *Educational and Psychological Measurement*, Vol. 56 No. 2, pp. 197-208.

Tol, R. (2009), "The h-index and its alternatives: an application to the 100 most prolific economists", *Scientometrics*, Vol. 80 No. 2, pp. 317-324.

van Raan, A.F.J. (2006), "Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups", *Scientometrics*, Vol. 67 No. 3, pp. 491-502.

Vinkler, P. (2007), "Eminence of scientists in the light of the *h*-index and other scientometric indicators", *Journal of Information Science*, Vol. 33 No. 4, pp. 481-491.

Vogt, W.P. (2007), *Quantitative Research Methods for Professionals*, Pearson Education, Boston, MA.

Waltman, L. and van Eck, N.J. (2012), "The inconsistency of the h-index", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 2, pp. 406-415.

Waltman, L. and van Eck, N.J. (2013), "Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison", *Scientometrics*, Vol. 96 No. 3, pp. 699-716.

Wan, X. (2014), "x-index: a fantastic new indicator for quantifying a scientist's scientific impact", available at: http://arxiv.org/abs/1405.0641 (accessed 31 March 2016).

Zhang, C.-T. (2013), "The h'-index, effectively improving the h-index based on the citation distribution", *PloS ONE*, Vol. 8 No. 4, p. e59912.

Zhivotovsky, L.A. and Krutovsky, K.V. (2008), "Self-citation can inflate *h*-index", *Scientometrics*, Vol. 77 No. 2, pp. 373-375.

Corresponding author
Cameron Stewart Barnes can be contacted at: cbarnes@une.edu.au