The construct validity of the *h*-index

Bibliometrics aims to apply objective, scientific methods to the analysis of citation data. However, the discipline depends on measuring a range of theoretical constructs — such as research impact — which are not directly observable. In this context, it is surprising that researchers in the field pay little, if any, attention to construct validation. This is in contrast to their colleagues in social and behavioural sciences, who regard construct validity as a necessary condition for the development and application of new theories. The purpose of this article is to show the benefits of a stronger program of construct validity to bibliometric enquiry. As evidence, it looks at the construct validity of the *h*-index, a metric which has emerged as a major focus of research in the last decade.

Keywords: bibliometrics construct validity, *h*-index, Hirsch index, measurement, research impact, validity

What is the *h*-index?

In 2005, the physicist Jorge Hirsch proposed a new index of a scientist's research impact (Hirsch 2005). He defined his index in the following manner:

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 simple terms, a researcher with 10 published articles, each of which has received at least 10 citations, has a *h*-index of 10. Hirsch expressed his hope that his new index would have practical applications. In particular, he proposed that it

may provide a useful yardstick to compare different individuals competing for the same resource when an important evaluation criterion is scientific achievement. (Hirsch 2005, p. 16572)

The influence of the *h*-index

Many bibliometricians have been drawn to study the *h*-index by its inherent simplicity. The result has been a flood of research. The effect of the metric is hard to overestimate. Many 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 is used in countless studies to measure the research output of, not only of individual scientists, but also research groups, universities, and even whole nations (e.g., Jacsó 2009; Prathap and Gupta 2009; Lazaridis 2010). The extension of the metric to the measurement of journal impact (eg., Schubert and Glänzel 2007) was perhaps only a matter of time. More striking has been the trend to apply the *h*-index in new and unexpected areas, from the level of research interest in pathogens (McIntyre, Hawkes et al. 2011; Sanni, Safahieh et al. 2013; Cox, McIntyre et al. 2014) to the popularity of YouTube channels (Hovden 2013).

There has been scattered criticism, but "little serious analysis" (Adler, Ewing et al. 2008). Most researchers in the have been content to take the *h*-index at face value. With very few exceptions (Adler, Ewing et al. 2008; Gaster and Gaster 2012; Gingrais 2014), there has been no sustained effort to look more deeply. Efforts at validation have largely been restricted to studies of correlations between the *h*-index and other measures of research impact.

Tweaking the h-index

This general disinterest in methodological issues is illustrated by the immense research effort directed to the design of *h*-index variants. Most researchers agree that the *h*-index has a number of obvious shortcomings. These include:

- a bias against younger researchers (Kelly and Jennions 2006);
- susceptibility to manipulation through self-citation (Burrell 2007; Schreiber 2007; Zhivotovsky and Krutovsky 2008; Bartneck and Kokkelmans 2011);
- the lack of any adjustment to allow for multiple authorship (Burrell 2007; Schreiber 2008)
- the difficulty of making comparisons between researchers in different disciplines (Batista, Campiteli et al. 2006; Alonso, Cabrerizo et al. 2009); and
- the inconsistent *h*-indices for the same individual reported by different citation databases (Jacsó 2008).

The existence of these flaws is not in serious doubt. Hirsch himself readily acknowledges each point (Hirsch 2005; Hirsch 2007; Hirsch 2010; Hirsch and Buela-Casal 2014). Similar criticisms can be made of a wide range of citation-based measures of research impact.

The main effect of these shortcomings has been to encourage"an almost bewildering explosion of publications proposing *h*-index variants" (Harzing, Alakangas et al. 2014). By 2011, there were at least 37 main published variants, most of which are so highly correlated with Hirsch's original index as to be largely redundant (Bornmann, Mutz et al. 2011; Bornmann 2012). New specimens continue to be added to the "h-index zoo" at a regular basis (e.g. Zhang 2013; Wan 2014). Much of this effort would be pointless, however, if the *h*-index itself suffered from fundamental problems. This is a question, however, which is rarely, if ever, raised in this context.

The effect of the *h*-index outside bibliometrics

The effect of the *h*-index has been felt far beyond bibliometrics. The metric is now commonly used for a range of purposes. It has become the "most popular quantitative measure of a researcher's productivity and impact" (Penner, Petersen et al. 2013, p. 8). Exactly as Hirsch hoped, this metric is used in many countries as a guide to resource allocation in higher education. It is employed to determine the success or failure of grant proposals, the outcome of applications for promotion, fellowship or tenure, and even the level of government funding for institutions (Barnes 2014). Although this trend has generated a good deal of unease (Burrows 2012; Pillay 2013; Nunn and Pillay 2014), 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).

What is construct validity?

Construct validity is the central idea behind current approaches to measurement in the social and behavioural sciences (Kane 2001; Colliver, Conlee et al. 2012). In simple terms, construct validity is "the degree to which a measure assesses the construct it is purported to assess" (Peter 1981, p. 134). Although this notion is strongest in fields such as sociology, education and psychology, it is also useful outside these disciplines. Construct validity is relevant whenever attempts are made to measure theoretical concepts otherwise not directly observable or measurable (Cronbach and Meehl 1956; Sechrest 2005; Cook 2006). The concept, for example, is frequently employed in the information sciences (Boudreau, Gefen et al. 2001; Kim 2009; Agarwal 2011).

Before the 1950s, the validation of statistical tests focused on establishing criterion and content validity. This approach no longer represents current thinking (Messick 1987; John and Benet-

Martínez 2000; Kane 2001; Colliver, Conlee et al. 2012). Construct validity is now seen as "the whole of validity from a scientific view" (Loevinger 1957, p. 363). It has gradually subsumed all previous categories of validity within a single overarching framework (Clark and Watson 1995; Smith 2005; Cook 2006; Strauss and Smith 2009). This is the approach adopted here.

Why is construct validity important?

Construct validity is the primary method used to establish the validity of any new statistical instrument in the social and behavioural sciences (Kane 2001; Smith 2005). In Cronbach's words: "Questions of construct validity become pertinent the moment a finding is put into words" (1988, p. 13). Construct validity is seen as an indispensable concept which underpins recent progress in psychology and allied disciplines (Smith 2005). It is regarded as the basis of the systematic theory testing which constitutes the scientific method in these fields (Benson 1998; Kane 2001).

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

Construct validity must be established whenever the concept under investigation is an abstraction some steps removed from the raw test scores (Kane 2001). Any index in the social and behavioural sciences requires construct validation for this reason. By definition, 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).

The *h*-index fits this description. The construct in question is research impact, which is self-evidently an abstraction. The *h*-index provides a single summary score, which is a composite measure of output (number of published articles) and impact (citations). In these terms, evaluation of the *h*-index in terms of construct validity seems highly appropriate.

Strong and weak validation programs

Cronbach distinguished to different programs for construct validation (Cronbach 1988; 1989). These he termed the strong program and the weak program. Cronbach wrote 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 (Cronbach 1988, pp. 12-13)

The power of the strong program of construct validity comes from objectivity and scepticism. It involves testing "your interpretation of test scores against other plausible hypotheses" (Murphy and Davidshofer 1991, p. 123). More recently, this distinction has been elaborated in the following 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)

Another aspect of distinction between strong and weak validation programs is worth bearing in mind. Murphy and Davidshofer observe that:

A second difference between a strong and weak program of construct validation may be temporal. Weak construct validation research is done after the fact (i.e., after the test has been developed) to determine just what an existing test or assessment procedure measured. Strong programs of construct validation often begin at the test development stage (Murphy and Davidshofer 1991, p. 123).

As will be shown, the *h*-index typifies the prevalence of weak programs of construct validation in bibliometrics. First, the trend has been for researchers to seek confirmatory evidence for the metric's validity: there has been little interest in falsification or alternative explanations. Second, the *h*-index shows no sign of having emerged from a systematic process of hypothesis testing. There is every indication that Hirsch happened upon the idea of his index first, and then sought evidence after the fact.

The dominance of weak programs in *h*-index research

Researchers have not entirely neglected efforts to analyse the *h*-index in terms of its construct validity. However, the few published attempts may be characterised as examples of Cronbach's weak program. A good case is a recent series of studies on the *h*-index scores of medical researchers (Bould, Boet et al. 2011; Patel, Ashrafian et al. 2013; Sharma, Boet et al. 2013). These papers make explicit reference to construct validity. In practice, however, they are better described as attempts to measure convergent validity. The studies do no more than show a correlation between the *h*-index and other citation-based metrics.

Building a case for construct validity

It is important to understand the relationship between construct and convergent validity. Any convincing case for construct validity must rest on an accumulation of evidence. No single approach is enough on its own, and the demonstration of construct validity is an on-going process (Loevinger 1957; Campbell and Fisk 1959; Messick 1987; Benson 1998). Two forms of validity are often regarded as particularly important: convergent and discriminant validity. Many theorists argue that:

Convergent and discriminant relationships between operational measures of constructs (e.g. psychological tests) and other operational measures are necessary to establish the network that form the basis of construct validation (Derogatis and Melisaratos 1983, p. 600)

These two categories of validity are relatively easy to grasp:

Convergent validity is the degree to which multiple attempts to measure the same concept are in agreement. The idea is that two or more measures of the same thing should covary highly if they are valid measures of the concept. Discriminant validity is the degree to which measures of different concepts are distinct. The notion is that if two or more concepts are unique, then valid measures of each should not correlate too highly (Bagozzi, Yi et al. 1991, p. 425)

Convergent and discriminant validity are subordinate categories within the broader construct validity rubric. Evidence of both can strengthen the case for construct validity. However, they prove little if the underlying theory is flawed. To examine this point in relation to the *h*-index, it is essential to be clear about the nature of the construct which this metric is intended to measure.

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

In his original paper, Hirsch states that his index offers a measure of the "overall scientific impact" of an individual researcher (Hirsch 2005, p. 16569.). At the end of his paper he wrote that it is an "easily computable index" which

gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions. (Hirsch 2005, p. 16572)

Hirsch choose his words carefully. He is clearly writing about impact in a bibliometrics sense, rather than, say, the wider impact of research on society. This much is evident from the fact that his index is purely citation-based.

In 2007, Hirsch went so far as to state that the *h*-index did indeed serve as " a useful indicator of scientific quality" (2007, p. 19198). However, it would be unfair to judge his index on this basis. This would be to make his metric a relatively easy target in construct validity terms (as will be discussed below). Measurement of scientific quality was not the aim of his instrument as it was originally described. In addition, Hirsch has retreated somewhat from his claim regarding the *h*-index as a measure of quality. In 2014, Hirsch wrote that:

The h-index is an indicator of the impact of a researcher on the development of his or her scientific field …. It is logical to expect the quality of research to go hand in hand with its impact. Although this is often the case, there are also exceptions (Hirsch and Buela-Casal 2014, p. 163)

What is impact?

If the *h*-index measures impact, then what is impact? Bibliometricians rarely define the term. Most researchers in the field are quite happy to use the word impact dozens of times in the same paper without ever venturing a definition (e.g., Leydesdorff and Bornmann 2011). The working definition of the term is often circular: impact is usually whatever citations measure. This practice does not mean that the concept of impact is empty of meaning. A perfectly good definition of impact exists. Decades ago, Eugene Garfield 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)

Garfield's definition of impact has clear support from empirical studies of citation context (Moravcsik and Murugesan 1975; Moravcsik 1988). It is equally consistent with research into the citing behaviour of individual researchers (Brooks 1985; Brooks 1986; Shadish, Tolliver et al. 1995; Case and Higgins 2000) and the day-to-day experience of researchers (Adler, Ewing et al. 2008).

In addition, Garfield's definition has another great virtue. It is acceptable to almost all bibliometricians, whatever their respective allegiance regarding the two main rival theories of citation: the normative and social constructivist (Feist 1997; Cronin 1998; Nicolaisen 2007; Bornmann and Daniel 2008; Riviera 2014). According to normative theory, citations represent the payment of intellectual debts (Kaplan 1965; Merton 1973; Merton 1988). The social constructivist theory holds researchers cite the works of others for a range of pragmatic and rhetorical reasons (Cozzens 1989; Erikson and Erlandson 2014). Whatever their theoretical orientation, however, 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).

Is impact the same as quality?

The notion that impact is a direct measure of research quality is somewhat controversial. For social constructivists, the idea is simply contrary to the evidence. It is simply ruled out of court. However, the gap between the social constructivist and normative theories in recent years is much less than often recognised. Most supporters of the normative theory of citations now concede that any direct association between impact and research quality applies only at "a high aggregation level" (Bornmann and Daniel 2008, p. 70). Even at this level, impact can be no more than a proxy for quality (Van Raan 2005). Even supporters of the normative theory agree that there are simply too many confounding factors at lower levels of aggregation (Cronin 2005; Bornmann and Daniel 2008; Bornmann and Marx 2014; Riviera 2014). In relation to a single paper or an individual researcher:

one cannot assume that a high number of citations correlate with a high level of usefulness (Bornmann and Marx 2014)

However, this is the level of analysis to which the *h*-index was originally designed to apply. For this reason, we should be extremely cautious about claims that the *h*-index measures research quality, as Hirsch himself now recognises.

Does the *h*-index measure impact?

If impact is essentially a question of utility, then does the *h*-index measure the usefulness of a piece of research to other researchers? The answer is that it does not appear to do so. In terms of traditional bibliometrics, the notion that the *h*-index measures impact involves a conundrum. 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, Hankin 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 have long been regarded as the clearest evidence of a researcher's impact.

Surprisingly, many champions of the *h*-index defend its insensitivity to highly cited articles as one of the metric's great strengths. 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).

Multiple authorship, however, is something of a red herring. In the jargon of bibliometricians, the citation record of even the most successful researchers is "right-skewed". The typical scientist, even the researcher who writes without co-authors, has a few highly cited papers and many more rarely cited ones (Seglen 1992; Bornmann and Daniel 2009; Cerchiello and Giudici 2014). Hirsch himself is a case in point. He is the sole author of his own "big hit", the original 2005 article on the *h*-index. If multiple authorship was such a serious problem, it would have made more sense to design a metric that corrected for this factor.

The insensitivity of the *h*-index to highly cited articles poses insoluble problems at every level of analysis. A mere handful of papers receive most citations in any field, while the bulk of articles

receive very few citations or none (Seglen 1992; Adler, Ewing et al. 2008; Radicchi, Fortunato et al. 2008). For this reason, the *h*-index fails to capture the great mass of citation data relevant to any discipline, as it excludes high cited paper. In these terms, it is hard to see how the *h*-index can reasonably be used to study impact at higher levels of aggregation, such as at the research group or institutional level, or across a whole field of study.

The idea that citation-based measures of impact should ideally exclude highly cited articles appears to have been Hirsch's own. This was a new departure from traditional bibliometrics. The current enthusiasm for this idea seems a rationalisation. Those researchers who have praised this approach do so exclusively the context of their use of the *h*-index, and cite Hirsch's problematic argument as justification.

Arbitrariness of the h-index

The construction of the *h*-index is essentially capricious. Hirsch gives his readers no reason for his decision to determine that

individual's h-index score 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)

In particular, he 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 the first place. Those researchers who have considered this point agree that the construction of the *h*-index is inherently arbitrary (Glänzel 2006; Lehmann, Jackson et al. 2008; Franceschini and Maisano 2010; Ravallion and Wagstaff 2011; Waltman and van Eck 2012; Ellison 2013; Schreiber 2013; Gingrais 2014). In simple terms, the issue is that Hirsch

assumes an 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, Jackson et al. 2008, p. 377).

The highly arbitrary nature of the *h*-index now widely recognised: surprisingly this has even been seen as a decided advantage (Ellison 2013). In terms of construct validity, however, this fact means that the *h*-index falls at the first hurdle. The construction of the *h*-index gives us no reason to believe that it measures something essential about a researcher's impact. An individual's *h*-index score is nothing more two lines on a graph, a geometric shape.

Critics have pointed out 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, Ewing et al. 2008, p. 13.)

There is broad agreement that identical *h*-indexes can have widely differing publication histories (Vinkler 2007; Bornmann and Daniel 2009; Cacioppo and Cacioppo 2012). In this context, it not surprising that the *h*-index has been scathingly dismissed as "a single metric with low information content" (Evidence Ltd 2007, p. 14).

Internal consistency is an essential element in construct validation. If an index or other statistical test produces inconsistent results, then it fails a basic test of operational utility. At this point, it should come as no surprise that the h-index is fails in this area also. The *h*-index behaves in a thoroughly inconsistent manner under a range of scenarios (Waltman and Van Eck 2009; Waltman and van Eck 2012). The inconsistency of the *h*-index is exactly what would expect from an index that lacks any logical basis for its construction.

The convergent validity of the *h*-index

The arbitrariness of the *h*-index has not deterred researchers from seeking to demonstrate the metric's convergent validity. One favourite method has been to show researchers' *h*-index scores correspond to the judgement of their peers. This first such attempt was made by Hirsch in his original article. He argued that an individual's *h*-index score has high predictive value whether a scientist has won honours, such as the Nobel Prize or Membership of the National Academy of Science (NAS). To demonstrate this point, he calculated the *h*-index for a group of Nobel Laureates in Physics over the previous twenty years. He found that 84 per cent of this group had an *h*-index at least 30. Hirsch then looked at the *h*-index scores for the newly elected members of the NAS in Physics and Astronomy. He found that that the median *h*-index of the members of this sample was 46. On this basis, Hirsch claims that:

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

Many commentators accept Hirsch's conclusion without comment (Panaretos and Malesios 2009) . However, his arguments do not sustain closer examination. As Adler and his co-authors point out:

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, Ewing et al. 2008, p. 13)

In fact, Hirsch's argument at this point provides good reason to question the convergent validity of his metric. The *h*-indices of the Nobel Prize winners in Hirsch's sample ranged from 22–79. Not only is this an extremely wide range, there is the additional problem there are many physicists with no chance of winning the Noble Prize who have *h*-indexes far higher than 22. The overlap between Hirsch's two sample populations in terms of their *h*-indexes was very high. In terms of the range, means and standard deviations the Nobel Prize winners and the members of the NAS were almost identical. This is not what we would expect if Hirsch's metric was closely related to the opinion of physicists. Nobel medallists are far more of scientific elite than the members of the NAS. In terms of peer judgement, the two groups are worlds apart.

An interesting aspect is Hirsch cannot have derived the construction of the *h*-index from his evidence. The *h*-index is no sense a best-fit solution to the data-points available to him. Instead, it is hard to resist the temptation that Hirsch developed the idea of the *h*-index first and then looked around for evidence in support of it. This idea may do Hirsch an injustice, but it is difficult to escape.

Can the *h*-index distinguish between researchers?

The arguments in Hirsch's 2005 paper strongly suggest that the *h*-index metric is insensitive to the real differences between researchers in terms of their impact. 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, research over the last ten years has 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 power at any point in a researcher's career. It is equally insensitive when applied to groups of highly cited researchers and when used to examine the publication careers of the least productive (Egghe 2006; Costas and Bordons 2007; Waltman and van Eck 2012). It cannot even "discriminate among average scientists" (Jin, Liang et al. 2007, p. 856). If so, then what logical claim for validity can be made on its behalf?

Further studies of the convergent validity of *h*-index

In the last decade there have been a number of studies which attempt to demonstrate the convergent validity of the *h*-index. The authors of these papers have typically sought to demonstrate a high correlation between the *h*-index and other impact indicators such as total number of citations (Cronin and Meho 2006; Van Raan 2006; Costas and Bordons 2007). However, it would be remarkable if no correlation existed between citation counts and *h*-index scores "since all these variables are functions of the same basic phenomenon— publications" (Adler, Ewing et al. 2008, p. 14). As Bornmann and his co-authors point out:

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, Wallon et al. 2008, p. 155)

Equally questionable are attempts to demonstrate the convergent validity between the *h*-index and the results of peer review (Van Raan 2006; Bornmann, Wallon et al. 2008; Lovegrove and Johnson 2008; Norris and Oppenheim 2010). Such studies often show statistically significant correlations, but the strength of such correlations is typically weak. However, any case for convergent validity depends critically on the strength of the correlation: mere statistical significance is not enough. None of these studies indicate any convincing evidence that *h*-index scores and the standing of individuals according to their peers are related in any meaningful way. At best, the published correlations suggest a simple hypothesis. Researchers who do not publish, or who publish only a few, rarely cited papers, will tend to have low *h*-indices. They are also unlikely to receive fellowships, be promoted or be highly regarded as researchers by their peers (Barnes 2014).

The divergent validity of the h-index

The *h*-index also fails the test of discriminant validity. If the metric is intended as a measure of impact, as Hirsch has stated this repeatedly, this is not how it behaves. Researchers have pointed out that "it combines a measure of quantity and impact in a single indicator" (Costas and Bordons 2007). This has been seen as one of its main advantages. In practice, the metric cannot do both. The *h*-index is constructed in such a way as to be

closely correlated with total publication output; thus, it will generally result in the same assessment as one based on counting publications (Kelly and Jennions 2006, p. 169)

The level of correlation can be alarmingly high if one is inclined to see the *h*-index as a measure of impact. For a population of 248 Danish professors in the health-sciences, the correlation coefficient between number of their papers and their *h*-index scores was r=0.93 ($p<0.001$) (Gaster and Gaster

2012, p. 830). This outcome is consistent with the results of other studies. The correlation between number of publications and *h*-index scores for groups of individual researchers in different fields is typically high: examples include 0.89 (Spearman) and 0.77–0.85 (Pearson) (Bornmann, Wallon et al. 2008). This state of affairs arises because a researcher's *h*-index can never be higher than the total number of his or her publications. Inevitably, a researcher's *h*-index is more strongly determined by the number of published papers than the number of citations these papers receive. The metric overwhelmingly measures output, not impact. Once again, the evidence in terms of discriminant validity against the construct validity of the *h*-index is damning.

The predictive power of the *h*-index

In 2007, Hirsch sought to demonstrate the validity of the *h*-index from another perspective. He argued that his metric was superior in terms of predictive power. Hirsch looked at the publication history of two convenience samples of physicists, analysing them through a number of statistical tests of correlation. His study indicated that:

- researchers with a high *h*-index 12 years after first publication were likely to have a high *h*-index after 24 years; and
- the *h*-index was better than total number of articles and total number of citations to articles in predicting its future cumulative value.

He concluded therefore that his index was the best method of predicting a researcher's future achievements. (Hirsch 2007).

This claim been accepted by many without comment. However, there are a number of problems with this interpretation. Hirsch believes that the high stability of the *h*-index over time is good evidence for its validity. However, if we look closer at the theory of citations, it is likely that the stability of the *h*-index is stronger proof of the metric's inherently low information content. The problem is we have no reason to believe that the typical academic's impact is stable over time. In this context is worth considering Cronbach and Meehl comments in their early paper on construct validity:

High correlations and high stability may constitute either favourable or unfavourable evidence for the proposed interpretation, depending on the theory surrounding the construct (1956, p. 200.)

This point is of particular relevance. Hirsch's assumption that the typical physicist's publication career is characterised by stability it not simply an over-simplification. It runs counter to the accumulated evidence of age-related effects (Cole 1979; Levin and Stephan 1991; Hall, Mairesse et al. 2007; Anderson, Hankin et al. 2008). It is exactly what we would *not* expect after decades of bibliometrics research.

When looked at more closely, Hirsch's argument suffer from even more glaring weaknesses in statistical terms. Shreiber points out Hirsch ignore 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. Specifically, I show examples where 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, 2013, p. 1)

Schreiber demonstrated that this conclusion applied even in relation to Hirsch himself. He found that:

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 2013)

In short, the apparent predictive power of the *h*-index according to Hirsch 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, 2013, p. 3)

García-Pérez and Núñez-Antón observe that Hirsch fails to take into account the fact that an individual's *h*-index cannot decline. For that reason the elaborate significance tests in his 2007 article are simply not meaningful (García-Pérez 2013; García-Pérez and Núñez-Antón 2013). More recent attempts to the show that the *h*-index can be used to predict future scientific careers (Acuna, Allesina et al. 2012) have been criticised on identical grounds (García-Pérez 2013; Penner, Pan et al. 2013; Penner, Petersen et al. 2013). In simple terms, a

cumulative non-decreasing measures like the h-index contain intrinsic autocorrelation, resulting in significant overestimation of their 'predictive power' (Penner, Pan et al. 2013, p. 1)

The flaws in Hirsch's case for the predictive validity of his metric are interesting for a number of reasons. The worrying fact is that Hirsch's arguments for the predictive power of his index are so obviously weak in both empirical and methodological terms. It is hard to believe that practising researchers in the field could accept an argument so clearly based on faulty arguments. The continued attractiveness of the *h*-index in absence of convincing evidence points towards the dominance of weak program of construct validation in bibliometrics. Many bibliometricians have simply looked the other way when offered alternative hypotheses.

Conclusion

The goal of this paper is not to show that the *h*-index lacks construct validity on any level. This point should have been apparent a decade ago. The purpose of this article is to demonstrate how the idea of construct validity can be usefully applied to bibliometric enquiry. The enormous popularity of the *h*-index as a research topic is good evidence, if any is required, that many researchers in the field do not give adequate attention to the validation of indices and other statistical instruments. Tests for convergent validity seem the main tool in their arsenal, yet these are applied with little regard to underlying theory or the constructs under examination. The result is a great deal of wasted effort. If bibliometrics is to advance as a discipline, and to justify repeated claims to scientific objectivity, its practitioners need to become more thorough in their approach. In truth, bibliometrics may never be one of the hard sciences, as it deals with so many abstractions, but at least it could live up to the standards of the best research in the social and behavioural disciplines.

References

Acuna, D. E., S. Allesina, et al. (2012). "Future impact: Predicting scientific success." Nature **489**(7415): 201-202.

Adler, R., J. Ewing, et al. (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). Berlin, International Mathematical Union.
Agarwal, N. K. (2011). "Verifying survey items for construct validity: A two-stage sorting procedure for questionnaire design in information behavior research." Proceedings of the American Society for Information Science and Technology **48**(1): 1-8.
Alonso, S., F. J. Cabrerizo, et al. (2009). "h-Index: A review focused in its variants, computation and standardization for different scientific fields." Journal of Informetrics **3**(4): 273-289.
Anderson, T. R., R. K. S. Hankin, et al. (2008). "Beyond the Durfee square: Enhancing the h-index to score total publication output." Scientometrics **76**(3): 577-588.
Bagozzi, R. P., Y. Yi, et al. (1991). "Assessing Construct Validity in Organizational Research." Administrative Science Quarterly **36**(3): 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 **36**(5): 456-470.
Bartneck, C. and S. Kokkelmans (2011). "Detecting h-index manipulation through self-citation analysis." Scientometrics **87**(1): 85-98.
Batista, P. D., M. G. Campiteli, et al. (2006). "Is it possible to compare researchers with different scientific interests?" Scientometrics **68**(1): 179-189.
Benson, J. (1998). "Developing a strong program of construct validation: A test anxiety example." Educational Measurement: Issues and Practice **17**(1): 10-17.
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 **10**(3): 149-153.
Bornmann, L. and H.-D. Daniel (2008). "What do citation counts measure? A review of studies on citing behavior." Journal of Documentation **64**(1): 45-80.
Bornmann, L. and H.-D. Daniel (2009). "The state of h index research." EMBO reports **10**(1): 2-6.
Bornmann, L. and W. Marx (2014). "The wisdom of citing scientists." Journal of the Association for Information Science and Technology **65**(6): 1288-1292.
Bornmann, L., R. Mutz, et al. (2011). "A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants." Journal of Informetrics **5**(3): 346-359.
Bornmann, L., G. Wallon, et al. (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 **17**(2): 149-156.
Boudreau, M.-C., D. Gefen, et al. (2001). "Validation in information systems research: a state-of-the-art assessment." MIS Quarterly **25**(1): 1-16.
Bould, M., S. Boet, et al. (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 **106**(3): 325-330.
Brooks, T. A. (1985). "Private acts and public objects: An investigation of citer motivations." Journal of the American Society for Information Science **36**(4): 223-229.
Brooks, T. A. (1986). "Evidence of complex citer motivations." Journal of the American Society for Information Science **37**(1): 34-36.
Burrell, Q. (2007). Should the h-index be discounted? . The multidimensional world of Tibor Braun: A multidisciplinary encomium

for his 75th birthday. I. W. Glanzel, A. Schubert and B. Schlemmer. Leuven, ISSI**:** 65-68.
Burrows, R. (2012). "Living with the h-index? Metric assemblages in the contemporary academy." The Sociological Review **60**(2): 355-372.
Cacioppo, J. T. and S. Cacioppo (2012). "Metrics of scholarly impact." Measurement: Interdisciplinary Research and Perspectives **10**(3): 154-156.
Campbell, D. T. and D. W. Fisk (1959). "Convergent and discriminant validation by the multitrait multmethod matrix." Psychological Bulletin **56**: 81-105.

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

Case, D. O. and G. M. Higgins (2000). "How can we investigate citation behavior? A study of reasons for citing literature in communication." Journal of the American Society for Information Science **51**(7): 635-645.

Cerchiello, P. and P. Giudici (2014). "On a statistical *h* index." Scientometrics **99**(2): 299-312.

Clark, L. A. and D. Watson (1995). "Constructing validity: Basic issues in objective scale development." Psychological assessment **7**(3): 309.

Cole, S. (1979). "Age and scientific performance." American journal of sociology: 958-977.

Colliver, J. A., M. J. Conlee, et al. (2012). "From test validity to construct validity … and back?" Medical Education **46**(4): 366-371.

Cook, D. A. (2006). "Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application." The American Journal of Medicine **119**(2): 7-16.

Costas, R. and M. Bordons (2007). "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level." Journal of Informetrics **1**(3): 193-203.

Cox, R., K. M. McIntyre, et al. (2014). "Comparison of the h-Index Scores Among Pathogens Identified as Emerging Hazards in North America." Transboundary and Emerging Diseases(4).

Cozzens, S. E. (1989). "What do citations count? The rhetoric-first model." Scientometrics **15**(5): 437-447.

Cronbach, L. J. (1988). Five Perspectives on Validity Argument. Test Validity. H. Wainer and H. I. Braun. Hillsdale, New Jersey, Lawrence Erlbaum**:** 3-18.

Cronbach, L. J. (1989). Construct Validation After Thirty Years. Intelligence: Measurement, Theory and Public Policy. R. L. Linn. Urbana-Campaign, Illinois, University of Illinois Press.

Cronbach, L. J. and P. E. Meehl (1956). Construct validity in Psychological Tests. The Foundations of Science and the Concepts of Psychology and Psychoanalysis. H. Feigl and M. Scriven, University of Minnesota Press. **1:** 174-204.

Cronin, B. (1998). "Metatheorizing citation." Scientometrics **43**(1): 45-55.

Cronin, B. (2005). "A hundred million acts of whimsy?" Current Science **89**(9): 1505-1509.

Cronin, B. and L. Meho (2006). "Using the h-index to rank influential information scientists." Journal of the American Society for Information Science and Technology **57**(9): 1275-1278.

Derogatis, L. R. and N. Melisaratos (1983). "The Brief Symptom Inventory: An introductory report." Psychological Medicine **1983**(13): 595-605.

Egghe, L. (2006). "Theory and practise of the g-index." Scientometrics **69**(1): 131-152.

Ellison, G. (2013). "How Does the Market Use Citation Data? The Hirsch Index in Economics." American Economic Journal: Applied Economics **5**(3): 63-90.

Erikson, M. G. and P. Erlandson (2014). "A taxonomy of motives to cite." Social studies of science **44**(4): 625-637.

Feist, G. J. (1997). "Quantity, quality, and depth of research as influences on scientific eminence: Is quantity most important?" Creativity Research Journal **10**(4): 325-335.

Franceschini, F. and D. A. Maisano (2010). "Analysis of the Hirsch index's operational properties." European Journal of Operational Research **203**(2): 494-504.

García-Pérez, M. A. (2013). "Limited validity of equations to predict the future h index." Scientometrics **96**(3): 901-909.

García-Pérez, M. A. and V. Núñez-Antón (2013). "Correlation between variables subject to an order restriction, with application to scientometric indices." Journal of Informetrics **7**(2): 542-554.

Garfield, E. (1979). "Is citation analysis a legitimate evaluation tool?" Scientometrics **1**(4): 359-375.

Gaster, N. and M. Gaster (2012). "A critical assessment of the h-index." BioEssays **34**(10): 830-832.

Gingrais, Y. (2014). Criteria for evaluating indicators. Beyond bibliometrics: harnessing multidimensional indicators at scholarly impact. B. Cronin and C. R. Sugimoto. Cambridge, Mass, MIT Press.

Glänzel, W. (2006). "On the h-index-A mathematical approach to a new measure of publication activity and citation impact." Scientometrics **67**(2): 315-321.
Hall, B. H., J. Mairesse, et al. (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 **16**(2): 159-177.
Harzing, A.-W., S. Alakangas, et al. (2014). "hIa: an individual annual h-index to accommodate disciplinary and career length differences." Scientometrics **99**(3): 811-821.
Hirsch, J. (2005). "An index to quantify an individual's scientific research output." Proceedings of the National Academy of Sciences **102**: 16569-16572.
Hirsch, J. E. (2007). "Does the h index have predictive power?" Proceedings of the National Academy of Sciences **104**(49): 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 **85**(3): 741-754.
Hirsch, J. E. and G. Buela-Casal (2014). "The meaning of the h-index*." International Journal of Clinical and Health Psychology **14**(2): 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 **64**(11): 2326-2331.
Jacsó, P. (2008). "The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases." Online Information Review **32**(2): 266-283.
Jacsó, P. (2009). "The h-index for countries in Web of Science and Scopus." Online Information Review **33**(4): 831-837.
Jin, B., L. Liang, et al. (2007). "The *R*-and *AR*-indices: Complementing the *h*-index." Chinese science bulletin **52**(6): 855-863.
John, O. P. and V. Benet-Martínez (2000). Measurement: Reliability, Construct Validation, and Scale Construction. Handbook of Research Methods in Social and Personality Psychology. H. T. Reis and C. M. Judd. New York, Cambridge University Press**:** 339-369.
Kane, M. T. (2001). "Current Concerns in Validity Theory." Journal of Educational Measurement **38**(4): 319-342.
Kaplan, N. (1965). "The norms of citation behavior: Prolegomena to the footnote." American documentation **16**(3): 179-184.
Kelly, C. D. and M. D. Jennions (2006). "The *h index* and career assessment by numbers." Trends in Ecology & Evolution **21**(4): 167-170.
Kim, Y.-M. (2009). "Validation of psychometric research instruments: The case of information science." Journal of the American Society for Information Science and Technology **60**(6): 1178-1191.
Lazaridis, T. (2010). "Ranking university departments using the mean h-index." Scientometrics **82**(2): 211-216.
Lehmann, S., A. D. Jackson, et al. (2008). "A quantitative analysis of indicators of scientific performance." Scientometrics **76**(2): 369-390.
Levin, S. G. and P. E. Stephan (1991). "Research productivity over the life cycle: Evidence for academic scientists." The American Economic Review: 114-132.
Leydesdorff, L. and L. Bornmann (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 **62**(11): 2133-2146.
Loevinger, J. (1957). "Objective tests as instruments of psychological theory." Psychological reports **3**(3): 635-694.
Lovegrove, B. G. and S. D. Johnson (2008). "Assessment of research performance in biology: how well do peer review and bibliometry correlate?" Bioscience **58**(2): 160-164.
Ltd., E. (2007). The Use of Bibliometrics to Measure Research Quality in UK Higher Education Institutions. London, Universities UK.

McIntyre, K. M., I. Hawkes, et al. (2011). "The H-Index as a quantitative indicator of the relative impact of human diseases." PloS one **6**(5): e19558.
Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. Chicago, University of Chicago Press.
Merton, R. K. (1988). "The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property." Isis **79**(4): 606-623.
Messick, S. (1987). Validity. Princeton, New Jersey, Educational Testing Service.
Moravcsik, M. J. (1988). "Citation context classification of a citation classic concerning citation context classification." Social studies of science **18**(3): 515-521.
Moravcsik, M. J. and P. Murugesan (1975). "Some results on the function and quality of citations." Social studies of science **5**(1): 86-92.
Murphy, K. R. and C. O. Davidshofer (1991). Psychological Testing: Principles and Applications. Eaglewood Cliffs, New Jersey, Prentice Hall.
Nicolaisen, J. (2007). "Citation analysis." Annual Review of Information Science and Technology **41**(1): 609-641.
Norris, M. and C. Oppenheim (2010). "Peer review and the h-index: Two studies." Journal of Informetrics **4**(3): 221-232.
Nunn, R. and A. Pillay (2014). "After invention of the h-index, is there a place for the teaching track in academic promotion?" Higher Education Research & Development **33**(4): 848-850.
Panaretos, J. and C. Malesios (2009). "Assessing scientific research performance and impact with single indices." Scientometrics **81**(3): 635-670.
Patel, V. M., H. Ashrafian, et al. (2013). "Enhancing the h index for the objective assessment of healthcare researcher performance and impact." Journal of the Royal Society of Medicine **106**(1): 19-29.
Penner, O., R. K. Pan, et al. (2013). "On the Predictability of Future Impact in Science." Scientific Reports **3**: 1-8.
Penner, O., A. M. Petersen, et al. (2013). "Commentary: The case for caution in predicting scientists' future impact." Physics Today **66**(4): 8-9.
Peter, J. P. (1981). "Construct Validity: A Review of Basic Issues and Marketing Practices." Journal of Marketing Research **18**(2): 133-145.
Pillay, A. (2013). "Academic promotion and the *h*-index." Journal of the American Society for Information Science and Technology **64**(12): 2598-2599.
Prathap, G. and B. Gupta (2009). "Ranking of Indian universities for their research output and quality using a new performance index." Current science **97**(6): 751-752.
Radicchi, F., S. Fortunato, et al. (2008). "Universality of citation distributions: Toward an objective measure of scientific impact." Proceedings of the National Academy of Sciences **105**(45): 17268-17272.
Ravallion, M. and A. Wagstaff (2011). "On measuring scholarly influence by citations." Scientometrics **88**(1): 321-337.
Riviera, E. (2014). "Testing the strength of the normative approach in citation theory through relational bibliometrics: The case of italian sociology." Journal of the Association for Information Science and Technology **64**(7): 1442-1453.
Sanni, S., H. Safahieh, et al. (2013). "Evaluating the growth pattern and relative performance in Nipah virus research from 1999 to 2010." Malaysian Journal of Library & Information Science **18**(2): 14-24.
Schreiber, M. (2007). "A case study of the Hirsch index for 26 non-prominent physicists." Annalen der Physik **16**(9): 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 **10**(4): 1-9.
Schreiber, M. (2013). "A case study of the arbitrariness of the h-index and the highly-cited-publications indicator." Journal of Informetrics **7**(2): 379-387.

Schreiber, M. (2013). "How relevant is the predictive power of the h-index? A case study of the time-dependent Hirsch index." Journal of Informetrics **7**(2): 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 **4**(1): 5-10.
Schubert, A. and W. Glänzel (2007). "A systematic analysis of Hirsch-type indices for journals." Journal of Informetrics **1**(3): 179-184.
Sechrest, L. (2005). "Validity of measures is no simple matter." Health services research **40**(5p2): 1584-1604.
Seglen, P. O. (1992). "The skewness of science." Journal of the American Society for Information Science **43**(9): 628-638.
Shadish, W. R. (1989). Perceptions and Evaluations of Quality in Science. Psychology of Science, Contributions to Metascience. B. S. Ghoulson, William R, R. A. Neimeyer and A. C. Houts. Cambridge, Cambridge University Press**:** 383-426.
Shadish, W. R., D. Tolliver, et al. (1995). "Author judgements about works they cite: three studies from psychology journals." Social studies of science **25**(3): 477-498.
Sharma, B., S. Boet, et al. (2013). "The h-index outperforms other bibliometrics in the assessment of research performance in general surgery: a province-wide study." Surgery **153**(4): 493-501.
Smith, G. T. (2005). "On Construct Validity: Issues of Method and Measurement." Psychological Assessment **17**(4): 396-408.
Smith, G. T. and T. C. B. Zapolski (2009). Construct Validation of Personality Measures. Oxford Handbook of Personality Assessment. J. N. Butcher. New York, Oxford University Press**:** 81-98.
Strauss, M. E. and G. T. Smith (2009). "Construct Validity: Advances in Theory and Methodology." Annual review of clinical psychology **5**: 1-25.
Van Raan, A. F. J. (2005). Measuring Science: Capita Selecta of Main Issues. Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems. H. F. Moed, W. Glänzel and U. Schmoch. Dordrecht, Kluwer**:** 19-50.
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 **67**(3): 491-502.
Vinkler, P. (2007). "Eminence of scientists in the light of the h-index and other scientometric indicators." Journal of Information Science **33**(4): 481-491.
Waltman, L. and N. J. Van Eck (2009). A taxonomy of bibliometric performance indicators based on the property of consistency, ERIM Report Series Research in Management.
Waltman, L. and N. J. van Eck (2012). "The inconsistency of the h-index." Journal of the American Society for Information Science and Technology **63**(2): 406-415.
Wan, X. (2014). "x-index: a fantastic new indicator for quantifying a scientist's scientific impact." arXiv preprint arXiv:1405.0641.
Zhang, C.-T. (2013). "The h'-index, effectively improving the h-index based on the citation distribution." PloS one **8**(4): e59912.
Zhivotovsky, L. A. and K. V. Krutovsky (2008). "Self-citation can inflate *h*-index." Scientometrics **77**(2): 373-375.