Research Evaluation, 2026, **35(1)**, rvag023
https://doi.org/10.1093/reseval/rvag023
Article

OXFORD

Towards a more informed and balanced use of scientific performance metrics

Jaap J. A. Denissen[1,*], Klaas Sijtsma[2] and Wil M. P. van der Aalst[3]

[1]Department of Psychology, Utrecht University, Utrecht 3584 CS, the Netherlands
[2]Department of Methodology and Statistics, Tilburg University, Tilburg 5037 AB, the Netherlands
[3]Process and Data Science, RWTH Aachen University, Aachen 52056, Germany
*Corresponding author: Jaap J. A. Denissen, Department of Developmental Psychology, Utrecht University, Utrecht 3584 CS, the Netherlands.
E-mail: jjadenissen@gmail.com.

Abstract

The goal of scientific assessment is to predict which individuals can make optimal use of limited resources within a specific context to make optimal allocation decisions. In academic contexts that pertain to individual-level allocations, this is most relevant for decisions on whom to hire for academic positions, nominate for awards, or whose research projects to fund. The current perspective paper draws upon insights from decades of psychometric research and more recent research on scientific performance to derive a set of five psychometric criteria that should be met for optimal assessment procedures in academia. Although data-driven decision making has gained popularity in most domains, there is increasing resistance against using quantitative measurements in scientific assessment. Recently, several stakeholders have proposed to jettison such measurements and focus instead on qualitative indicators or narratives. We argue that both quantitative and qualitative assessment do not always meet our five criteria, but solely relying on qualitative indicators appears to be a suboptimal strategy. We argue instead that there are smarter ways to use quantitative indicators so that they become more reliable, predictive, and ultimately also more efficient and equitable. We conclude with a set of recommendations for scientific quality assessment that is based on the most recent psychometric and scientific insights. In an appendix, we apply these recommendations to a Dutch case study of how researcher information is considered in the application procedure for a prestigious individual grant.

Keywords assessment; bibliometrics; performance indices, scientometrics; work performance

1. Introduction

Confronted with limited resources, in academia, like many other professional settings, difficult decisions about individual allocations need to be made. Who gets the coveted professorship or the competitive grant? Such decisions are often based on a variety of criteria (e.g. excellence, diversity, feasibility), but an assessment of research potential is often chiefly among them. In scientometrics, research potential is often based on citations as the unit of analysis (Garfield 1955). During the previous decades, a growing emphasis on quantitative indices based on citations, such as the *h*-index (Hirsch 2005), has developed, and many different indices have been proposed. Such indices have sometimes been used as a measure of scientific quality (i.e. a quality intrinsic to the researcher), academic reputation (i.e. how positively the researcher is regarded by her peers), or research impact (i.e. how often other researchers use works of the researcher in their own work). Early studies (Bornmann and Daniel 2007; Hirsch 2007; Ruscio et al. 2012) have suggested that the *h*-index aligns with indicators of peer reputation, and therefore has at least some validity. For example, researchers with high *h*-indices are more likely to be selected as members of prestigious scientific academies or as winners of scholarly awards.

Still, many authors (e.g. Barnes 2016) have questioned aspects of the validity of the *h*-index and other bibliometric indices. Limitations mentioned include incomplete coverage of databases and misspellings of journal and author names, causing incorrect citation counts. In addition, Aksnes et al. (2019) have argued that the *h*-index might not be related to scientific impact and relevance. The authors found little, if any, relationship between citation metrics and generally accepted quality indicators of scientific work, which are plausibility, originality, scientific value (Polanyi 1962), and societal value. Finally, the *h*-index is not field-normalized (although corrections are readily available; Ioannidis et al. 2016), and alternative indices (e.g. *P*(top 10%) or beamplots) have been developed to address this, as outlined below.

Partly because of these criticisms, recently, the conviction has grown that targeting the optimization of quantitative indices

goes at the expense of research quality (Barnes 2016). In response to this conviction, a movement has developed that dismisses quantitative indices and promotes the use of narratives produced by researchers explaining their intentions and visions on research to replace the indices (e.g. Torres-Salinas et al. 2023; Waks and Kramer 2023). In 2012, the influential San Francisco Declaration on Research Assessment (DORA) stated: 'Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions' (DORA 2012). Although the DORA declaration primarily focused on abuse and misuse of the Journal Impact Factor (JIF), it was generally interpreted more broadly and, in some cases, led to a ban on reporting published papers and the number of citations. In Torres-Salinas et al. (2023), this phenomenon is described as 'bibliometric denialism'.

The current article is partly inspired by our experiences in the Dutch context of research evaluation, because this is the context that we know best (for a similarly informed focus on the Italian context, see Abramo 2024). In the Netherlands, the Dutch Research Council (acronym NWO) invests almost one billion euros each year in curiosity-driven research, research related to societal challenges, and research infrastructure. According to their Strategy 2023–26 (NWO 2022), NWO endorses the DORA declaration. While the NWO strategy technically supports the responsible (i.e. balanced) use of metrics, in practice, this has resulted in a more radical shift towards qualitative evaluation, at least when evaluating individual researchers. Although citations are allowed for individual publications, they are no longer allowed at aggregate levels, such as the individual researcher.[1] While the UK REF has adopted a more balanced approach for decades, combining both qualitative and quantitative indicators, this approach is under threat in other countries. For example, New Zealand has cancelled its performance-based research fund (PBRF) quality evaluation, partly as a result of criticisms of quantitative assessment (Collins and Simmonds 2024).

In this article, we argue that scientometrics done carefully and effectively provides a reasonable starting point for the interpretation of the contribution of an individual researcher's work to her research area. Even though we thus argue that it can be reasonable to first check quantitative information and then follow up with qualitative analysis, our main point is that quantitative indices and interpretations of contribution and value are complementary. In isolation, each provides an incomplete and often biased picture of someone's contribution to the scientific enterprise (for a similar argument, see Hicks et al. 2015; Ruscio et al. 2012). Our article offers a novel perspective by contextualizing the recent backlash against research metrics as an overcorrection, driven in part by misunderstandings about statistical properties of research metrics (see also Abramo 2024) and exaggerations of the possible contribution of metrics, inviting negative reactions to abandon them completely, thereby ignoring whatever value they do have. Our main focus is on psychometrics because we primarily focus on the evaluation of individual researchers' performance (rather than the performance of a journal or an academic institution).

Throughout our paper, we developed a set of five criteria that assessment of scientific performance should meet. These criteria were based on the literature on psychometrics, which offers a well-developed toolkit that can be used to evaluate the robustness of local practices (Drenth and Sijtsma, 2006). First, assessment procedures should be bias-free and fair, meaning that the criteria and the language used to describe them should not confer unfair advantage to certain demographics (*Fairness Criterion*). Second, assessment should be standardized so that all candidates provide relevant information in approximately the same way (*Standardization Criterion*). Third, procedures need to be fact-based, grounded in verifiable facts or examples that should be easily understandable to judges and difficult to manipulate (*Objectivity Criterion*). Fourth, procedures should be comparable across fields to enable comparisons across scientific (sub-)disciplines (*Field Comparability Criterion*). Fifth and finally, procedures must be scientifically sound, meaning that criteria need to be limited to scientifically proven predictors of the outcome (*Predictive Validity Criterion*). By introducing these criteria and applying them to a case study (see Appendix A), we move beyond generic though useful calls for 'balanced metrics' by proposing concrete principles for combining qualitative and quantitative indicators, including guidance on the order in which quantitative and qualitative information should be taken into account, and how they should be integrated.

2. Quantitative indices

In the following, we will first discuss some basic properties of quantitative indices. We have opted to use the *h*-index as an example, but not because this is the most advanced or suitable index, but because it is still very popular (in spite of criticism) and relatively easy to understand. Further below, we do in fact discuss alternative indices, so this section is not meant as an endorsement of the *h*-index, but rather as an illustration that it simply is a statistic comparable to other well-known statistics that researchers in many areas use almost daily. They use the statistics for what they can accomplish and accept their limitations, compensating for these limitations using other information. This is what we will also recommend.

2.1. Statistics provide limited information

Quantitative indices like *h* provide only limited information about what one eventually desires to know. This is no different in statistics. No researcher well-versed in data analysis will claim that the mean of a variable informs her of everything there is to know about the variable's distribution, or that a correlation reveals all about the exact form of a relationship between two or more variables. With performance metrics, it is no different: They provide incomplete information about what one wants to know, in this case, the quality of a researcher's work. Just as no one expects statistics to work miracles, one must not expect performance metrics to reveal all there is to know about research quality, and it stands to reason that a full performance analysis requires several sources complementing one another. We make a point of this because, in the world of scientific performance metrics, there is a strong undercurrent focusing on what impact indicators *cannot* accomplish. Although this is no different from indicators in other domains, it is used as an argument to resort to more subjective impressions: Metrics should

be dropped in favor of ad-hoc evaluation schemes using informal criteria (Woolston 2021). We disagree with such an approach and rather prefer two alternatives: (a) optimize the robustness and accuracy of indices, and (b) complement incomplete indices with additional information, either quantitative or qualitative.

Without knowing anything else about a researcher—a starting point that would be highly undesirable when hiring someone for a job or comparing nominations for an award—a performance indicator like the *h*-index (or an alternative index) might represent our best (though imperfect) estimate of a researcher's future productivity and impact. In the case of the *h*-index, the assumption would be that if the researcher has shown a steady rate of productive and impactful work in the past, and to the extent that it has been observed and recorded by others, then this past performance is a reasonable predictor for future scientific performance (Hönekopp and Khan 2012). In this regard, quantitative indicators would seem to score high on the Predictive Validity Criterion.

However, additional information might cast doubt on this predictive validity. For example, one could imagine a very ambitious early career researcher ratcheting up high-profile publications that are widely cited. That person, however, might have the ambition to use a full professorship, once it is attained, as a springboard for political ambitions. Alternatively, he or she might become exhausted by a focus on extrinsic performance and stop publishing after some time. Even more, it might be that past performance was facilitated by the availability of resources (e.g. a helpful supervisor; ample research time) that are no longer available in the new position. In all instances, the researcher's past performance would no longer be predictive of future performance. However, these and other possibilities can only be identified by other sources of information, such as probing the motivation of the researcher, and exploring (and weighting) the circumstances in which academic achievements were realized. While this again illustrates the need to avoid blindly relying on simple indices, it is also a reality that, in most cases, past performance records, especially if they were obtained in situations that are representative of the future setting, are predictive of future performance (Meehl 1989; but see Schmidt et al. 1992).

2.2. Deconstructing quantitative indices

We introduce the popular distinction in the social and behavioral sciences between formative and reflective measures that we believe is also relevant in the present context (other research areas may use different distinctions; the present one is generally accepted among psychometricians, but see Edwards and Bagozzi 2000, for more discussion). An index is a formative measure because it is defined by one or more observable indicators (Edwards and Bagozzi 2000). The *h*-index depends solely on an individual's number of publications and their citation frequencies–as such, the *h*-index scores highly on the Objectivity Criterion. Other examples of formative indices are the Amsterdam Exchange Index (AEX) defined by the exchange rate movement of the 25 greatest stock funds and socioeconomic status (SES) defined by income, education, and occupation. Whereas formative measures or simple indices do not have surplus meaning beyond the indicators that define them, reflective measures are manifestations of underlying attributes defined by theory (Edwards and Bagozzi 2000). An example of a reflective measure is the intelligence quotient resulting from a psychological test consisting of a set of problems derived from the theory of intelligence, but that do not define the attribute. Essential is that the theory would allow different or partly different sets of problems for measuring the attribute. Thus, one might claim that the intelligence test refers to an attribute, general intelligence, that has surplus meaning beyond the operationalization the test represents. In contrast, an index coincides with the elements that define it. For example, deleting income from the SES would lead to a different definition.

Now we move to the *h*-index and other quantitative indices, and notice that, exactly like the mean, the correlation, and other statistics, they inform us only about a feature of a distribution of data without telling the whole story. We will argue that quantitative indices can be useful for what they have to say but that additional information is needed for a complete picture of the data. Later, we will focus on scientific accomplishment beyond the data on which the *h*-index and other quantities are based. To deconstruct such indices, we must first identify the data on which they are based. For bibliometric analysis, the raw data comprise a complete list of a researcher's scientific publications and, for each publication, a count of the number of times it is cited in the academic literature. In a later section, we will comment on the data source used and the major role it plays in realizing an index' value (e.g. Van der Aalst 2022, Van der Aalst et al. 2023). Here, we simply assume a reasonable choice was made as far as these data sources permit (for a comparison across widely used sources, see Adriaanse & Rensleigh, 2013; Martín-Martín et al. 2018; Singh et al. 2021), and that the procedure is equal for all individual researchers included in the system. In the case of the *h*-index, if we sort this list by number of citations for each publication, we can compute a Rank variable. Then, the *h*-index equals the number or publications for which the number of Citations is at least as large as their Rank. Table 1 shows a simple example for someone with 6 publications and $h = 4$. Because this logic is the same for each scientist, the index scores highly on our Standardization Criterion.

Hirsch (2005) discussed the *h*-index meticulously, studying features of functions representing the relationship between publications ranked by decreasing numbers of citations (in a graph, on the abscissa) and citation frequency (ordinate) and pointing out interesting additional data features that help to

Table 1 Number of citations in decreasing magnitude for six publications.

Publication	Cites	Rank	Cites ≥ rank
A	327	1	true
b	53	2	true
C	21	3	true
D	8	4	true
E	4	5	false
F	0	6	false

Note. *h* is the number of publications of a total of *N* that are cited at least *h* times in other publications, while the other $N - h$ publications are each cited no more than *h* times. In the table, $h = 4$.

understand researchers' performance records. An example is $m = h/n$, where n is the number of years someone has been active, and so m is the slope parameter for a linear model that connects publications and citation counts. Furthermore, h-values may depend on whether an author works alone or in a large group producing many multi-authored papers (he has since proposed alternative indices to account for the latter; Hirsch 2010; 2019). Finally, citation rates can vary greatly across different scientific fields, which is a reason why field-normalized indices are often recommended. Examples include the P(top 10%), which ignores the volume of publication output, and percentile beamplots, which show the entire distribution of data points and thereby provide a nuanced and rich source of information.

As stated, many authors have criticized the h-index (e.g. see Waltman and van Eck 2012). The recent literature on the h-index and similar indices has extensively focused on its limited scope, but this is more likely due to its misuse in practice than its theoretical foundation. Hirsch (2005: 16571) himself noticed that 'a single number can never give more than a rough approximation to an individual's multifaceted profile', necessitating consulting other sources of information in addition to what the h-index and similar statistics can provide. This criticism might be seen as a variation of the Predictive Validity Criterion, as it highlights the existence of other predictors that might meaningfully predict scientific performance, in addition to the h-index. In fact, Hirsch (2019) himself recently introduced a novel index that can be used alongside the h-index and that indicates academic leadership. Still, it could be argued that the index can be valuable as a summary indicator of the extent to which an author regularly publishes papers that are cited in the literature. This can be achieved by adopting either a high-volume strategy with the occasional (and perhaps lucky) paper that obtains a large number of citations or a strategy that prioritizes a smaller number of (potential) high-impact papers.

The construct validity (Cronbach and Meehl 1955; Lissitz 2009) of the h-index and other indices has been the topic of much debate (e.g. Barnes 2016), but it is imperative to notice that, being indices, such quantities are defined by their constituent indicators. The question then becomes whether productivity, number of citations, authorship positions and distributions, and any other measurable bibliometric aspect are valid indicators of scientific performance (Predictive Validity Criterion). This is precisely the issue on which the debate must focus: How is scientific performance defined and, in addition, is there enough agreement among the scientific community about a definition? To us, it is clear that without quantifying productivity in some way it is impossible to assess scientific performance. Also without citations, it is clear that a researcher's work has little to no impact on the research discipline. Our view is that productivity and impact through citations are necessary conditions for scientific performance, but we will also argue that they are not sufficient, because scientific performance has surplus meaning beyond productivity and impact, and that additional information is needed.

2.3. Which data source to use?

The h-index and other quantitative measures discussed later use information about published papers, the authors, and the number of times the paper was cited. There are several data sources that can be used to get these numbers. Most widely used for this purpose are Web of Science (WoS, www.webofscience.com), Google Scholar (GS, scholar.google.com), and Scopus (www.scopus.com). WoS is owned by Clarivate Analytics and originates from an information retrieval tool in 1964 by Eugene Garfield from the Institute of Scientific Information (ISI). Google Scholar was launched in 2004 and uses a web crawler instead of depending on lists of selected sources (mostly journals). This inclusive approach gives GS potentially more comprehensive coverage of scientific and scholarly literature. Scopus was also launched in 2004. Scopus is owned by Elsevier and, compared to WoS, provides better coverage in humanities, social sciences, business, economics, and computer science. There are also focused databases such as PubMed (limited to biomedical and life sciences literature, pubmed.ncbi.nlm.nih.gov) and new initiatives such as Dimension (www.dimensions.ai) launched in 2018. The latter provides an API for non-commercial scientometric research projects.

The different databases have been compared in various articles. Martín-Martín et al. (2018) compared WoS, GS, and Scopus using 2,299 highly-cited papers (covered by all three databases) from 252 subject categories published in 2006. Only 46.9% of all citations were found by all three databases. GS found the most citations, including most of the citations found by WoS and Scopus. In contrast, only 6% of all citations were found by WoS and/or Scopus, and not by GS. An additional 10.2% of all citations were found by both GS and Scopus (7.7%), or GS and WoS (2.5%). Over a third (36.9%) of all citations were only found by GS (Martín-Martín et al. 2018). Singh et al. (2021) investigated the journal coverage of WoS, Scopus, and Dimensions. They found that almost all journals indexed in WoS are also covered by Scopus and Dimensions. Scopus indexed 66.07% more unique journals as compared to WoS and Dimensions covered 82.22% and 48.17% more unique journals as compared to WoS and Scopus, respectively.

The coverage of the different databases greatly varies per discipline. For example, for researchers in physics or chemistry, the coverage of WoS is generally quite good. However, this is not the case for, for example, computer scientists. It is not uncommon that GS reports almost *four times* as many citations as WoS and a much higher h-index. Note that WoS does not include books and provides limited coverage for conferences and workshops, thus explaining the differences. Although GS provides the best coverage, it has obvious drawbacks. For example, different publications may be merged into one when titles are similar, or the same citation is counted twice because different versions of the same paper are not merged. Hence, the higher citation counts of GS are not necessarily superior to those of WoS and Scopus.

Next to differences in coverage per discipline and data source, there are also remarkable differences in publication traditions (e.g. the number of authors and the ordering of authors). In five years (2012–2016), the number of papers in the Nature Index with more than a thousand authors increased from zero to a hundred, mostly in physics (Sijp 2018). Also in other fields (astronomy, genetics, but increasingly also in the social sciences), it is common to have many authors. This has had a large impact on the h-index, and it makes for a poorer performance on Field

Comparability Criterion. However, this does not imply that one should not use quantitative indicators anymore. These problems can be addressed by selecting a representative data source and correcting for the number of authors, which is easy to accomplish as we will see shortly.

2.4. Alternative indices

Many bibliometric alternatives to the highly popular and much criticized *h*-index have been proposed, and some already existed previously to *h* (Alonso et al. 2009; Ruscio et al. 2012). Since it is not our purpose to provide yet another review of such indices, we focus on three alternative approaches we consider promising. Moreover, we show that it is relatively easy to include desirable features in an index and exclude other features considered undesirable. We do not claim that these three approaches represent the only possibilities, but we think together they offer a convincing argument of how indices can be modified to counter alleged weak spots and boost favorable features that improve quality assessment.

2.4.1. Correcting the *h*-index

We first consider the standardization of the raw indices, and focus on *h*. For researcher *i* with h_i and the maximum $h_{\max}$ available in the database we use for comparison, a possible standardization (Ioannidis et al. 2019) is

$$y_h = \frac{\log(h+1)}{\log(h_{\max}+1)},$$

with values $y_h = 0$ if $h = 0$ (meaning the individual has no publications that are cited at least once, a truly trivial case) and $y_h = 1$ if $h = h_{\max}$ (meaning the individual is at the top of her field). As *h* increases, the standardized *h*-index denoted y_h levels off until its maximum is reached. For example, if $h_{\text{John}} = 12$, $h_{\text{Mary}} = 37$, and $h_{\text{Heather}} = 62$ (increments of 25), while the field maximum is $h_{\max} = 112$, John, Mary and Heather have y_h-values 0.54, 0.77, and 0.88 (increments of 0.23 and 0.11), respectively. Thus, the greatest gain is in increments in the smaller counts, meaning that having modest impact is clearly valued over having little impact whereas having larger impact produces a diminishing return. The adapted index could serve to standardize the assessment for all candidates and keep subjectivity from inviting unequal treatment at arm's length. Other manipulations are feasible. For example, the concave logarithmic transformation function can be replaced with another function to vary the weighing of different performance features. For example, the transformation could be convex but bounded, emphasizing that later increases in *h* are harder to attain.

As another example of a modified index, Koltun and Hafner (2021) discussed fractional allocation, correcting *h* and other indices for hyperauthorship involving hundreds of co-authors. One could also include a penalty for self-citations that starts to become active as the frequency exceeds a preset maximum, thereby acknowledging that some self-citations are unavoidable when a researcher's work accumulates across time. Such measures would improve the performance of quantitative indices on various psychometric criteria (e.g. it would contribute positively to the Field Comparability Criterion because hyperauthorship differs between fields). We do not intend to develop such indices here, only to suggest that different data features can be quantified in numerous indices to be used for different assessment goals.

2.4.2. Multi-component indices

The second approach we discuss was put forward by Ioannidis et al. (2016; also, Ioannidis et al. 2019; Ioannidis et al. 2020), who proposed combining six components into their *c*-index that appreciates personal accomplishment (desirable) and corrects for piggybacking due to multi-co-authorship. The six components are: (1) the total number of citations (N_c), (2) the *h*-index (Hirsch 2005), (3) the Schreiber h_m-index (*h* adjusted for co-authorship; Schreiber 2008; see Hirsch 2010; 2019, for alternative adjustments), (4) the number of citations to papers as a single author (N_{cs}), (5) the number of citations to papers as single or first author (N_{csf}), and (6) the number of citations to papers as first, single, or last author (N_{csfl}). Each of the six components is standardized logarithmically, as discussed in the standardization of the *h*-index to obtain quantities ranging from 0 and 1, and denoted y_c, y_h, y_{h_m}, y_{cs}, y_{csf}, and y_{csfl}. Index *c* is the sum of the six standardized components, $c = y_c + y_h + y_{h_m} + y_{cs} + y_{csf} + y_{csfl}$, and thus, ranges from 0 to 6. The higher each of the six quantities, *ceteris paribus* the higher *c*. The *c*-index combines bulk impact (N_c, *h*) and author order and co-author adjusted impact (h_m, N_{cs}, N_{csf}, N_{csfl}). Critical readers may notice that alternatives for *c* consisting of different components and thus emphasizing different features of authorship are readily defined. We focus on *c* because it shows convincingly how one can (de)emphasize authorship features and adapt a performance index to the needs of a research area and the appreciation of performance. Next, we will have closer look at *c*.

Ioannides and colleagues present the *c*-index almost in passing but two issues deserve attention. The first issue is that the inclusion of the components, y_{h_m}, y_{cs}, y_{csf}, and y_{csfl}, although formally unequal, all correct for authorship inflation and essentially reward (different types of) scientific leadership/seniority. It might be necessary to add different components in future revisions (e.g. giving credit to authors who employ open science practices and focus on consensually important research problems; Schönbrodt et al. 2025; Leising et al. 2022). Thus, it should become clear that one can choose components that suit one's purpose best. This could mean replacing, deleting, or adding components, and thus changing *c*'s interpretation. A flexible attitude could help enormously in accepting quantitative approaches to assessment while not ignoring the value of interpretation.

One could also use a profile of different indices that each provide different information on research performance, instead of one summary index, such as *c*. Based on the six components and the open data provided by Ioannidis (2022), one can easily select or adapt the weights. For example, in some disciplines, it is customary to order authors alphabetically, and in others, the sorting is based on contribution. It is important to apply local knowledge about such disciplinary differences to obtain indicators with maximal validity.

Finally, Ioannidis (2022) assigns authors to the most common scientific field and the two most common scientific subfields based on the author's publications, which improves the

performance on our Field Comparability Criterion. This way, authors are assigned to 22 scientific fields (e.g. Chemistry, Biology, and Information & Communication Technologies), and these are further divided into 176 subfields according to the Science-Metrix journal classification system (e.g. Organic Chemistry, Plant Biology & Botany, and Artificial Intelligence & Image Processing). This enables comparisons between different fields and authors within a field, thus addressing the criticism that bibliometric indices are field-dependent.

2.4.3. More complex modeling strategies

The third approach is exemplified by Sinatra et al. (2016). They asked how citation impact changes over a scientific career and whether results can be used to facilitate the prediction of the timing of a scientist's outstanding accomplishment. They found that an author's highest-impact paper appears randomly in the sequence of her publications over time; it might be the first or the last publication or any paper in-between with almost equal probability. Sinatra et al. (2016) call this the random-impact rule; it is valid for any discipline, career length, decades of activity, and number of authors, alone or in teams. This result and other results, not discussed here for reasons of brevity, trigger the question of whether a scientist's individual ability affects the quality of her performance.

To study this issue, Sinatra et al. (2016) present a statistical model that includes a parameter for the ability of a researcher to transform the potential of a randomly chosen research topic into a highly cited research paper. The key finding was that citation impact is the result of a random choice of a research topic/paper with impact potential P from a distribution that is the same for each scientist (meaning that scientists, in general, do not have a hunch for impact potential) and a person-dependent parameter Q_i that quantifies scientist i's ability to enhance ($Q_i > 1$) or diminish ($Q_i < 1$) a paper's impact potential. Because different scientists can identify the same topics and may select a particular topic and write a corresponding paper, the difference between them resides in their ability to take advantage of the available information. Intuitively, this is the same in music, painting, and other creative activities, where artists with varying abilities will drive the same starting point to results highly differing in colleagues' appreciation. It makes a huge difference whether Lennon and McCartney worked on an initial simple idea of a folk song compared to most other musicians working on that same idea as a point of departure.

Analysis showed that the Q value is constant for most scientists, and Q-based rankings of scientists predict Nobel-winning careers better than the h-index and other indices. The consequences of the results Sinatra and colleagues presented are highly relevant to the assessment of research performance, especially for our Predictive Validity Criterion. Of course, without any productivity, a high (latent) Q-value will not produce a high impact. The dependence of the results on the quality of education, the size of the research field, gender, and publication habits, to name a few variables, is unknown. Much additional research is needed to study the complexities of what makes researchers successful in the sense that they contribute to the solution in ways that are picked up by their colleagues and have an impact on the solution of societal problems. Although still in its infancy, further research on the determinants of Q may provide interesting insights useful in training young scientists.

2.5. Sociological problems

In this section, we review sociological problems of quantitative indices, by which we mean that their use can have undesirable social consequences. Some of these problems directly relate to our Fairness Criterion, because the use of quantitative indices can benefit some demographic groups unfairly. This is a paradoxically a result of the high score on the Objectivity Criterion: The h-index makes abundantly clear what causes a high value; that is, many publications that are often cited frequently. This means that for individuals, it is obvious what they must accomplish to realize a high h-index. This can introduce perverse incentives: People might become inclined to increase the indicators themselves to increase chances of success, such as tenure, promotion, and increased access to research money. This can be a problem if researchers deploy publication strategies that only aim at increasing chances of success at the expense of research and publication quality. For example, a researcher might divide a comprehensive study into several smaller studies and publish them separately as short articles for no other reason than that she expects this to increase her number of publications and their total impact, even if this obscures meaningful associations between the parts. Whether such a strategy really increases her h-index is uncertain, but what counts is that for colleagues it would be more convenient to have all the results in one comprehensive paper and not to be bothered with an author's publication strategy.

On a collective level, such perverse incentives call to mind the so-called Goodhart's law (Goodhart 1975): 'Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes'. In economics, where the law was first formulated, this means that when a certain quantitative index is chosen to track performance, it is quickly distorted. A historical example comes from the British colonial administration in India, which wanted to reduce the size of the cobra population and offered money for each cobra skin, assuming people would start to hunt cobras. Eventually, however, people started to breed cobras for their skin, thus ironically increasing the cobra population. A short while later, social psychologist Donald Campbell (1979) formulated a similar principle: 'The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor'. Basically, people will adapt their behavior to what the measure or the index expects of them, such that they increase the perceived profit.

There is evidence that such effects have taken place. For example, Moed (2008) found that British researchers adopted their publication practices because of indicators used in the UK Research Assessment Exercises (RAE), and Michels and Schmoch (2014) demonstrated the same for German researchers. For the h-index, Koltun and Hafner (2021) even demonstrated that it ceased to correlate with indicators of scientific reputation a short while after its introduction in 2005. Next to phenomena such as unduly dividing a comprehensive paper into several shorter papers, the primary mechanism of h's declining validity

is authorship inflation: the increasing number of co-authors per academic paper. Although new indices, such as the *c*-index, can be developed that correct for authorship inflation (Ioannides et al., 2016, 2019), the sociological criticism is broader and would assume that these new indices, over time, will also become corrupted by humans' tendency to act strategically.

Another way in which quantitative indicators can influence human performance is by undermining intrinsic motivation. In developmental psychology, Ryan and Deci (2000) showed that children lose interest in solving a puzzle when they are paid to do so. The explanation for this effect is that the extrinsic reward crowds out the experience of pleasure over solving the puzzles. For example, children would start to infer that they were solving the puzzles to earn money instead of experiencing pleasure, and when the money stopped, they would cease playing. While this example has parallels to the scientific enterprise, there are also obvious differences. For example, nobody suggests stopping paying scientists a salary to increase their intrinsic motivation. However, the example suggests the possibility that intrinsic factors, such as curiosity, could be undermined by more extrinsic features that could be exacerbated by quantitative tracking, especially when this the tracking is comparative (for an analysis of sociological effects of quantifying performance, see Espeland and Sauder 2007; Espeland and Stevens 1998; Lamont 2012).

Such comparative tendencies could be made especially salient when the *h*-index is transformed into a population-normative statistic. For example, the *m*-index (Hirsch 2005) can be interpreted normatively, such as $m = 2$ corresponding to 'outstanding scientists, likely to be found only at the top universities or major research laboratories' (ibid., p. 16571). A normative approach represents a zero-sum partitioning of researcher populations, which creates winners and losers and therefore might erode some individual researchers' self-esteem as well as systemic trust among scholars. That said, this is likely not true for all individuals, and we believe that healthy competition can have positive motivational effects for some people (Denissen and Sijtsma 2022).

3. Qualitative indicators

As stated earlier, some people have used the problems inherent in quantitative indicators to argue for a switch to qualitative indicators. Qualitative indicators have a long history in scientific assessment. For example, consider awards. These are essentially qualitative labels, such as being a Nobel laureate, that demonstrate that individuals are deemed representative of an idealized profile of various dimensions, including creativity and rigor. In the laudation, an assessment committee would formulate the rationale for the distinction, often highlighting landmark contributions to the field in question. Such narrative evaluations have several important advantages, which makes it unsurprising that they are increasingly considered as sources of performance assessment. For example, the Dutch NWO has shifted to narrative CV's when evaluating the potential of grant applicants to successfully carry out the proposed research (Gossink-Melenhorst 2019). In Appendix A, we outline this shift in more detail, and also apply our five criteria to evaluate the evolving evaluation practices at NWO.

We acknowledge four advantages of narrative approaches compared to quantitative indicators. First, narratives are better able to describe a *Gestalt*, which is the totality of an individual's profile. For example, more than a quantitative index, a narrative description might better convey that a researcher is not only a very original thinker but also a gifted mentor and a rigorous empirical scholar. While this could, in theory, also be achieved by a dashboard of quantitative indicators, as we propose further below, such a nuanced picture cannot be easily provided by a single quantitative index. In theory, this could improve qualitative indices' performance on our Predictive Validity Criterion.

Second, a narrative approach is potentially better at sketching the *contextual* factors that have given rise to a person's achievements or lack thereof. For example, a narrative approach can outline personal or institutional factors (e.g. pregnancy or sick leave, working in a country with structural economic disadvantages) that have influenced a person's scientific work. This can boost qualitative indices' performance on our Fairness Criterion.

Third, a narrative approach allows for an *interpretation* of a person's scientific achievement. For example, it might describe the type of substantive contribution that a particular scholar has made to the field. Consider a scholar who has long swum against the tide, working on a novel but highly risky methodological paradigm that eventually revolutionized her field of research (e.g. an often-cited example would be Katalin Karikó, who received the Nobel prize for her contribution to mRNA technology). Such a contribution could never be highlighted by a quantitative indicator. For example, Einstein currently has $h = 128$ (https://scholar.google.com/citations?user=qc6CJjYAAAAJ&hl=en), which is outstanding but not unique, but it would be hard for anyone with a similar or even higher value to top Einstein's contribution to science. It is difficult to map this advantage on our psychometric criteria, though perhaps it has some potential to improve our Field Comparability Criterion.

The fourth advantage of a narrative approach is that it is potentially more compatible with *human motivation*. This advantage is not covered by our five psychometric criteria, though it does align with motivational insights from developmental psychology and sociology, among others. As stated above, just as in other areas of professional life, also in science, individuals might be motivated by competition, and this should not be discouraged in any wholesale fashion. For some, research is like competitive sports, where competition leads to extraordinary achievements. For many individuals, however, narratives can provide direction and meaning in shaping individual lives. This is an important feature that makes narrative discourse a powerful motivational force. For example, a researcher's narrative might highlight how her achievements serve a deeply-held mission to alleviate human suffering or to understand the mysteries of particle physics. By introducing narrative approaches, individual researchers might become more connected to deeply held personal inspirations in a way that quantitative indicators would never be able (McAdams 2008).

3.1. Problems of qualitative approaches

Just like quantitative approaches, however, narratives have several problems that limit their usefulness in scientific assessment

(for an overview of empirical research into the use of narrative CVs, we refer the reader to Bordignon et al. 2023; and Strinzel et al. 2021). First, narratives are influenced by *self-serving biases*: human tendencies to portray themselves in an overly positive light (Taylor and Brown 1988). This relates to a poorer performance on our Objectivity Criterion: Researchers might skimp on performance areas that represent relative weaknesses in their profiles or might cherry-pick the indicators that best serve their purpose. For example, a researcher with few, rather mediocre papers but one highly-cited paper might highlight the latter as demonstrating the enormous impact that her work has had on a certain field. Such bias is hard to evaluate in case an assessment committee is not knowledgeable of important developments in the field in question. This might happen if the committee mainly includes generalists, as is often the case.

Second, a problem in using narrative approaches for selection purposes related to the first problem is that a narrative-based assessment may favor persons who are prone to self-enhancement as a personality trait, but also persons who understand well what might optimize their chances, scientifically relevant or not, and opportunistically target their narrative at the assessors, mainly for effect rather than substance. This poses a problem for our Fairness Criterion. Universities and funding agencies obviously do not intend to stimulate persons with a particular personality profile (in the worst case, a narcissistic profile; Morf et al. 2011) and persons skilled at manipulating narratives to obtain a desired result no matter what to reach their goal, such as a job, a promotion, or a grant assignment. On the contrary, most universities and funders are actively trying to avoid selfish and manipulative behavior (O'Reilly and Chatman 2020) and stimulate academic achievement.

A third problem is that narratives concern a particular individual and therefore do not allow easy comparison between individuals. After all, each narrative is an individual story (although it can of course also be influenced by sociological factors, such as the habitus of a particular field; Hutaibat et al. 2021). If not properly structured, narratives are difficult to generalize, thus limiting the comparison with other persons. This problem makes that narratives score very poorly on our Standardization Criterion. Relatedly, it has been shown by van den Besselaar and Mom (2022) that the use of certainty words and a narrative (as compared to an analytic) writing style in ERC applications predicted the favorability of peer review evaluations. To the extent that these stylistic features are uncorrelated with quality, this poses a problem for our Predictive Validity Criterion. Especially the finding that 'certainty words' predicted peer review scores exacerbates our concern that certain qualitative formats provide biased opportunities for highly confident individuals, which might especially apply if unstructured formats are implemented in a (semi-)public setting, such as an interview stage.

A fourth problem is that unstructured narratives are difficult to refine using psychometric methods, which makes it more difficult to improve their relative standing on the Predictive Validity Criterion. This is decidedly different in the case of quantitative indices. For example, if an index is unreliable (e.g. if the number of citations per year fluctuates strongly), then measures can be undertaken to alleviate this problem (e.g. aggregate across multiple publication years). Furthermore, if an index is invalid, then efforts might be undertaken to boost validity. An example is the research that has suggested that the *h*-index is no longer valid if researchers engage in authorship inflation practices (Koltun and Hafner 2021). By dividing the number of citations by the number of authors, one fixes this problem in a straightforward and valid way. This can lead to a continuous improvement of measurement quality that is the hallmark of solid science. Until now, similar efforts have not been undertaken for qualitative indices, and their Gestalt character might make such efforts more difficult.

Thus far, we have mainly focused on narrative CV's or other kinds of self-assessments of one's scientific work and its impact. One might interject that this is not the only possibility: One might simply ask knowledgeable reviewers to actually read representative papers that are used to back up a candidate's quality claims. While this might appear reasonable at first sight, it suffers from momentous problems. To begin with, this takes a lot of time, and because reviewers are often extremely busy and (often) not or only symbolically compensated for their work, it would probably decrease the pool of willing reviewers if a requirement were to read and compare a substantial number of papers, which is a logistical (vs. a psychometric) problem. Second, there is the problem of representativity: When only a single paper is read per candidate, this introduces a lot of random bias. For example, what if the paper in question is highly atypical of the researcher's performance? This is especially true because the contribution of an individual paper is partly dependent on chance, as Sinatra and her colleagues (2016) have convincingly shown. By focusing on a single 'best' paper by a researcher, one maximizes this source of error thus faring more poorly on our Predictive Validity Criterion. Third, to truly compare individual contributions, reviewers of heterogeneous panels would have to be knowledgeable of many different fields and specializations within fields (for similar arguments and analysis, see Abramo 2024), which threatens our Field Comparability Criterion.

More fundamentally, while pondering the limitations of focused qualitative assessments of single papers, it seems odd in that case not to make use of a readily available aggregate assessment of different qualitative reviewers who, as a collective, have evaluated all individual papers of any candidate. Namely, by counting how many times scientific peers find a conclusion useful enough to cite in their own papers–provided that at least a substantial number of them have actually read the paper before they cite it, which seems a reasonable assumption. This would of course revive the same bibliometric indices that many critics would like to abolish!

The problems mentioned illustrate that abandoning quantitative research assessment ('bibliometric denialism') may lead to subjective and unverifiable decisions (Van der Aalst et al. 2023). By getting rid of quantitative indices, one discards advantages like objectivity and transparency in favor of qualitative indices that suffer from their own set of disadvantages.

4. What do We really want?

In this section, we put the value of quantitative information in perspective. First, instead of emphasizing all the weaknesses

and concluding that $h = 68$ (Mary) does not mean a lot when it comes to research quality, one could ask what it means when a senior scholar has $h = 12$ (John). As implied by the analysis of Sinatra et al. (2016), to contribute to the progress of a research area, researchers must come up with results and communicate these to their colleagues. When a senior researcher produces only a few publications that colleagues cite marginally, resulting in a small number of citations and a low *h*-index, this is difficult to interpret as a positive result. A plausible contextual explanation might help understand this result, but it can go either way, positive but also negative (e.g. John might work in a relatively obscure field, or Mary might be part of a citation network). Even though the number of citations and the *h*-index cannot tell the whole story, at least they show the degree to which the scientific community has picked up the work. Anyone trying to accomplish something new, whether she is a musician, a sculptor, a silversmith, or a scientist, must face the risk of failure but cannot hide from trying.

Second, to pursue the topic of quantitative measures, a review of their limitations must not downplay their value as summaries of an otherwise complex data set. When confronted with a table of raw data and asked to make sense of it, one of the first things people do is look for patterns and start counting. It helps tremendously to have statistics available as a first summary of interesting data features (just like one starts a complex statistical analysis by looking at some descriptive statistics) and replace the often-primitive attempts of a naïve observer trying to make sense of a batch of numbers. This is not to say that we should accept all limitations uncritically, but it does mean that one should be realistic about the availability of useful summaries, use them for what they can accomplish, and add information from different sources to compensate for their shortcomings, considering the goals one wishes to attain.

Finally, from personnel selection psychology, we know that rare birds are hard to catch (H. C. Taylor and Russell 1939; Wiggins 1973). This is the base-rate problem when brilliant applicants are rare in the population of potential applicants, which is common in science (Simonton 1988). This means that if very few individuals make groundbreaking discoveries and the vast majority must face the truth that they are laying the brickwork of a scientific area rather than providing the innovations. It stands to reason that at least modest *h*-indices are a necessary condition for filling scientific leadership positions. *Ceteris paribus*, a low *h* indicates little recognition of accomplishment, while a high value suggests at least a greater contribution. The rare bird who has made a groundbreaking contribution will be generally known and recognized for that feat, irrespective of their *h*-value. Our point is that quantitative indices are especially important for most scientists whose quality ranges from poor to excellent—that is, almost everybody except for the few geniuses that everybody knows.

To conclude, quantitative summary indices are useful in personnel assessment and hiring but they do not tell the whole story. Also, we have concluded that small publication records and few citations are reason for skepticism about scientific accomplishments especially in senior researchers, while very high production and citation records are rare and top researchers are well-known also without the bibliometric information. The question becomes how to value modest bibliometric indices among the vast majority in the middle of the distribution. What additional sources do we need? In the following, we aim to derive a few constructive recommendations (see also Appendix A for recommendation applied to a case study).

4.1. The future of quantitative indices

It seems ironic that prominent voices (CWTS, 2021; Woolston 2021) are calling for the abolishment of quantitative indices during a period in which computational power and scholarly insights make it both possible and easy to reap the benefits of increasingly sophisticated indices. If we do not want to throw out the baby with the bathwater, this means that proponents of quantitative indices need to be committed to upholding the strictest of standards. The most straightforward way to do this is for databases like Web of Science (WoS), Google Scholar (GS), and Scopus to invest in more sophisticated dashboards. For example, Harzing (2023a, 2023b) created Publish or Perish, which pulls alternative indices from Google Scholar profiles. It is surprising that such configuration options do not yet exist in WoS, GS, and Scopus. The *c*-index (Ioannidis et al. 2016) and the seminal work of Sinatra et al. (2016) provide helpful insights on how to configure the corresponding dashboards.

In our view, configuration options of bibliometric profiles should at least include the following. First, it would be beneficial to more easily compare quantitative indices based on different databases. For example, Google Scholar should ideally have a filter to only count citations and publications that are also included in Web of Science, and vice versa (in fact, both platforms have started to work together in 2022, making this easier to realize). Second, it should be possible to correct quantitative indices for the field(s) in which the researcher is primarily active, ideally using a hierarchical field classification (e.g. social sciences > psychology > social psychology > stereotype research). Third, a correction for academic age should be implemented, for example, by relating the *h*-index to a scholar's academic age, thus resulting in the *m*-index (Hirsch 2005). Fourth, an option should exist to remove most self-citations (leaving a limited number, acknowledging that an author's work often builds on her previous work) from a person's *h*-index. Finally, it is important to correct citation counts for the number of co-authors on a particular paper (Ioannides et al. 2016, 2019). Based on the requirements of the assessment procedure (or merely as a robustness check), assessors could toggle different options on or off and study the results. Obviously, these configuration options will result in more complexity, which likely requires more interpretative guidelines and the input of qualified bibliometricians.

We believe that, in the future, more sophisticated indices will become possible. Using propensity score matching or nomination procedures, it should be possible to identify 'benchmark individuals' with similar demographic, institutional, and career features. For example, one could identify researchers in a similar career phase, from the same World region, sharing the same demographic features, and who are active in the same field. The person to be evaluated could then be compared to these peers, and this would result in a relative comparison index, akin to a field-normalized citation score. This might offer an improvement compared to current practices, in which disadvantaged candidates have difficulty overcoming obstacles to produce scientific

articles that are widely cited at a steady rate. Furthermore, future work might develop useful bibliometric indicators of more specific dimensions of researcher quality, such as interdisciplinarity (Schölvinck et al. 2024), leadership (Hirsch 2019), independence (Van den Besselaar and Sandström 2019), and disruptive novelty (Bornmann et al. 2020).

We note that a more thorough psychometric analysis of quantitative indices is also helpful to point at the limits of such indices and suggest alternatives. One obvious example is the assessment of early-career researchers. It is well known from statistics that the mean of a distribution is more precise as the number of observations grows. The numbers of publications and citations are small at earlier career stages and do not mean a lot (but see Krauss et al. 2023). In such cases, traditional assessment procedures using standardized assessment techniques and/or work probes could be used, perhaps supplemented by assessment of open science practices (e.g. Schönbrodt et al. 2025) and one's contribution to solving scientific problems of consensual importance (Leising et al. 2022). Ability and personality assessment could be used to evaluate the likelihood of future success based on our knowledge of what predicts scientific excellence (Simonton 1988). An example of a work probe could be the production of a research proposal that is clearly written and contains original ideas.

4.2. Improving qualitative indices

We reiterate that such statistical truths do not imply that counts should be dismissed favoring narratives: The fact that some people are dissatisfied with one method—metrics—does not imply that another method—narratives—is superior without proof. Qualitative indicators have both benefits and pitfalls and are not the panacea that render quantitative indices superfluous, as some seem to suggest (e.g. https://www.nwo.nl/en/dora). In the assessment literature, it is well-known that unstructured job interviews are more biased than structured ones (Wiesner and Cronshaw 1988). The psychological literature on personnel selection and decision-making is replete with evidence-based warnings against unstructured interviews giving way to vast numbers of judgment biases (Kahneman 2011; Kahneman et al. 2021; Meehl 1954; Roe 1983; Schmidt and Hunter 1998), a problem encountered in psychological therapy as well (e.g. Dawes 1994). Given this knowledge, it is odd that narrative alternatives hardly limit the content of a narrative assessment. Committees should identify the key information that they are interested in and provide separate prompts to elicit input on this information. This recommendation calls for a formalization of qualitative data such as narratives into two categories, one consisting of information relevant for the prediction task at hand and another containing information that is irrelevant and must be ignored. The prompts needed to elicit relevant information may take the form of a structured questionnaire helping the researcher to provide precisely the information needed to make the best predictions. We propose two formalized approaches in the final section before the Discussion.

As suggested, the relevant information likely varies as a function of the prediction goal at hand, such as a job position. For example, if the focus is on hiring a researcher who can build bridges between disciplines, then a strong interdisciplinary narrative might be valued, but that does not alter the fact that narratives must contain standardized elements that can be compared across candidates. This requirement holds for any job selection, promotion assessment, and grant application, but the specific standardized elements are different for these three categories and also for different jobs, job levels, and types of grants. For example, the UK Résumé for Research and Innovation (R4RI) template offers four modules to describe relevant aspects of researcher quality, including innovative potential, establishing effective working relationships, contributions to the scholarly community, and societal impacts. As can be seen in Appendix A, the Dutch NWO forms have fluctuated a great deal in this regard, being originally quite well standardized in their 2015–19 forms, to a less well standardized form in 2020–22 and especially 2023, to a current form in which the categories 'general academic profile', and 'leadership and mentorship' are standardized to some extent.

In our view, the diverging nature of the input criteria (both between countries and across time) raises the question to what extent these different input categories are based on scientific research, thus questioning the Predictive Validity Criterion. Furthermore, however, instead of leaving researchers the choice of what to describe, a higher score on the Standardization Criterion would require a description of all these aspects, as well as a pre-determined weighting of the various dimensions. Finally, to increase performance on the Fairness Criterion, we think it is important that participants elaborate on the specific circumstances in which they have worked, indicating how much time they devoted to research and whether career breaks have occurred.

Standardizing these narrative elements or prompts, different candidates are more likely to provide similar information, thus increasing comparability and fairness. Researchers who specialize in qualitative research and structured assessment interviews can invest resources in fine-tuning the prompts, for example, by comparing different formulations of a question about research motivation and then considering which formulation produces the most valid outcomes (for an example, see Hartwell et al. 2019). In our view, it is surprising that psychometric research is so rarely used to improve qualitative decisions, for example, how to improve the reliability and validity of panel members' assessments of quality (Marsh et al. 2008; Van den Besselaar and Sandström 2015). In the next section, we identify two approaches to combining quantitative and qualitative data sources (for additional examples, see Hammarfelt et al. 2020; van den Besselaar and Sandström 2020).

4.3. Evaluating quality assessment procedures in practice: the case of the Dutch NWO Vici scheme

To check whether our psychometric criteria are suitable to evaluate specific assessment procedures, we tried to apply them to various iterations of the Dutch NWO's Vici scheme. This is a prestigious grant scheme that is aimed towards senior researchers. As we outline in Appendix A, the 2015–19 forms combined both qualitative and quantitative information, yet from 2020 onwards, narrative elements increasingly became important to the detriment of quantitative indicators.

As can be seen in Appendix A, our criteria are well able to evaluate the psychometric quality of the different form editions, and they provide some context to the various rounds of adjustments that NWO supposedly undertook to improve these forms' quality. In particular, the 2015–19 forms combined quantitative and qualitative information and scored relatively high on our psychometric criteria, but subsequent forms reflected a backlash against quantitative information and a promotion of a form of qualitative assessment that scored relatively poorly on our criteria.

The 2024 form apparently tried to rebalance qualitative and quantitative assessment input, coming closer the our proposal for a combined approach that is described in more detail below. The newest edition of the form also improved on various of our criteria, though our framework could be used to identify aspects where even this form still falls short on specific quality criteria. This hopefully demonstrates the practical usefulness of our approach to improve the quality of scholarly assessment procedures.

4.4. The royal road: combining different sources of information

As has become clear, our psychometric analysis reveals advantages and disadvantages of both qualitative and quantitative procedures to assess researcher quality, which are summarized in Table 2. Our recommendation is therefore to combine different sources of information for making high-stakes decisions (for a similar proposal, see Schönbrodt et al. 2025). We first discuss how to deal with quantitative indices and then we identify aspects of narratives that might be used in combination with the quantitative indices. Second, we discuss two approaches of combining all information and reach a well-founded conclusion.

For example, hiring a new full professor at age 40 means that this person will spend the next 20–30 years in a tenured position, consuming large quantities of resources, such as accumulated salary, as well as supporting personnel and research funds. In these and other situations, it would be wise not only to combine different sources of information but also to consult experts in performance assessment who are able to provide advice regarding the type of input that is available and that best matches the assessment purpose. For example, an expert in bibliometric assessment could help to provide necessary corrections to standard quantitative indices and suggest suitable narrative prompts to collect qualitative information from participants (for a similar argument, see Abramo 2024). It is surprising that institutions making many selection decisions, such as universities or grant agencies, do not routinely employ this information to improve committees' work, but instead assume that each committee reaches the optimal conclusions without such input.

Because of the pros and cons of both quantitative and qualitative approaches, the sound advice is to combine sophisticated input from both. Given the additional time that qualitative assessment requires, it seems prudent to first focus on robust quantitative indices to select candidates that meet a certain threshold of performance. As noted, it is important that such indices are corrected for biases, such as academic field or demographic factors. However, it is also important that different quantitative indices are compared by knowledgeable assessors. For example, what does it mean if someone has a high *h*-index when it implodes when correcting for the numbers of co-authors and self-citations? Van den Besselaar and Sandström (2020) discussed another approach to integrate qualitative (peer review) and bibliometric information, namely by requiring peer reviews to explicitly discuss and explain discrepancies between their evaluations and the quantitative indicators.

Recently, Schönbrodt et al. (2025) also proposed a two-step procedure, but with methodological rigor as an additional criterion in the first step, in addition to citations and publication volume. We agree with the addition of rigor as a primary criterion, though we think that this is better evaluated in more detail during the second phase of the procedure, as outlined below. In any case, once a thorough assessment has identified applicants that meet a certain cut-off, the assessment procedure could turn to more qualitative indicators that give more insight into the context in which scientific achievements have taken place, including the applicant's motivation. The assessment committee could be instructed on how to analyze this qualitative information, but also to address broader questions, such as concerning:

- The place a person's research occupies in the field in which she is active. Is it located at the heart of the general research effort addressing the topics generally considered urgent, or is it marginal or even esoteric?
- The contribution a researcher makes to the solution of a problem generally identified as relevant or the progress a field makes. Does she address the right questions, and is the research of sufficient quality to contribute?
- The methodological quality of the research design. Does the researcher choose the appropriate designs to study a

Table 2 Comparing quantitative (*h*-index), qualitative (e.g. narrative), and their combined approaches for scientific assessment.

	Quantitative	Narrative	Combination
Construct	Ability to frequently produce papers that are well-cited.	Ability to convince fellow researchers of the quality of one's work	Importance of both objective successes as well as their significance in broader research contexts
Strengths	Objective, difficult to fake	Meaningful, contextualized	Weaknesses of the other two balance each other out
Weaknesses	Strategic publication, erosion of intrinsic motivation, narrow interpretation without context	Easy to fake, self-enhancement, subjective, lack of verifiable assessment	More work

problem? Does she use representative samples based on well-defined populations?
- The quality of the statistical methods. For example, do the statistical methods match the data structure? Are formal proofs provided, if necessary? What is the quality of the reporting?
- The writing. Is it of sufficient quality in terms of flow and structure?
- The researcher's plans for the near future. Are these plans realistic? What are the resources required to execute the planned research? Are the resources available, or is there a realistic scenario available for acquiring grant money?

4.3.1. Approach 1: the statistical model

The statistical approach is the most rigorous of the two approaches we discuss. It consists of a detailed analysis of the requirement to make a job or job level fulfillment a success, as well as the execution of research financed with grant money. This analysis, in the psychological research literature known as criterion analysis (the criterion is what one intends to predict), should next identify the information a committee or an agency must collect to predict the criterion. This predictive information is formalized as a set of K predictor variables enumerated X_1, X_2 up to X_K, that are included in a statistical model to predict criterion Y. The predictors may, for example, be a performance index that corrects for certain features deemed undesirable in the prediction of research success (in grant application) denoted X_1, and quantifications of narrative aspects given in the list concluding the preceding section (quantification may simply be a recording of whether an aspect is present; e.g. using appropriate designs may be quantified as 1 = yes and 0 = no), denoted $X_2, \ldots, X_K$; K depends on the number of predictors used. Criterion variable Y may consist of a rating scale for expressing the degree to which the funded research proved successful, but may also be extended to a number of variables.

The K predictors and, say, the one criterion variable are included in a model, usually a linear regression model. Empirical research using a sample of grant applicants should provide evidence that the model indeed has enough predictive power. From the literature on job selection (e.g. Wiggins 1973), it is well known that this kind of research is feasible but also laborious and suffers from many methodological problems, the simplest being the collection of a sufficiently large sample. However, if an evidence-based model has been obtained and implemented, the scores of the grant applicant on the various predictors can be inserted in the model, which then produces a statistically optimal prediction of the criterion score. The Committee's task would then be limited to discussing predictions and making well-founded amendments the model is unable to consider. The final decision is made by the Committee.

4.3.2. Approach 2: the human model

In most cases, formal prediction models are not feasible, and one obvious reason is that the set of predictors may change in number and in content between goals (job selection, promotion decision, grants application) and for each goal between cases. In this case, the Committee takes over the role of the model. Criterion analysis and predictor selection remain necessary first steps. For the Committee work, the predictors need to be decided upon. Each Committee member receives a (paper or electronic) sheet with the predictors in the columns and the candidates to be assessed in the rows. Each member also has access to the data that are relevant for the task, which many contain performance indices, publications, details of research plans, and so on. An interview of the Committee with each candidate may provide the opportunity to settle for uncertainty or obtaining additional information. Each Committee member independently rates each candidate on each predictor using fixed rubrics (supplied to each member) that limit undesirable assessment flexibility introducing random error and judgment bias. The Committee chair assembles the sheets and computes the total number of credit points for each candidate. The candidate (s) with the highest scores win/wins, but this has to be evaluated and discussed first by the Committee.

Here, the Committee has a greater role than they had when the statistical model was in place, but the similarity resides in the methodology: In both cases, all the important decisions regarding criterion and predictors, how to reach a final score and how to make a decision have been taken before the selection, promotion, and grant assignment decisions are made. We expect a greater formalization of the procedure before the procedure takes place can keep most random error and disturbing biases at arms' length. An additional advantage is that the use of highly subjective narratives can be pushed back to the essentials for the task at hand.

Other formalized procedures may be possible, but the two discussed here can be considered ideal approaches to performance assessment. In both cases, the Committee has the final word, but to prevent the Committee relapsing to the judgment subjectivity we wish to reduce, we recommend controlling the possibilities for departing all too easily from the two procedures' objective outcomes. How to allow the Committee overruling the model while preventing the quality of decision-making to deteriorate calls for a separate study that is beyond the scope of the present study.

5. Conclusion

In this article, we compared different approaches to evaluate individual accomplishments and research quality. Although organizations increasingly use data-driven approaches to improve efficiency and effectiveness, there are forces trying to completely abandon the use of quantitative measures ('bibliometric denialism'). In some settings, it is even forbidden to mention numerical data (like the h-index and the number of citations) in applications or evaluations. However, application processes (for positions or funding) are competitive, and qualitative approaches suffer from the problems described in this article. We showed that there are possibilities to improve measures like the simple h-index, but these are rarely used. Alternatively, we propose to combine quantitative and qualitative approaches, and aim to improve both in terms of all of the five psychometric criteria discussed above.

We are aware that this is not an original position, as already Hirsch (2005) himself expressed this opinion, and a balanced approach is also endorsed and implemented by the UK REF

(Wilsdon 2015). However, more than ever, balance in the discussion seems necessary. Of note, we have based our call for a more balanced approach on theoretical grounds, drawing from relevant psychometric frameworks. While we claim that this approach will increase validity and improve performance, this claim can and should of course be tested by future empirical work (e.g. tracking the outcomes of implementing a more balanced policy).

We further note that the present developments toward team efforts do not imply that hiring an 'all-star team' guarantees collective performance at a high level reflecting on individuals as well. Well-functioning teams often differentiate roles, and in an academic context, this could mean that a strong writer and a creative individual will not reach their potential without individuals who can match creative ideas with suitable methods or who provide a constructive social atmosphere in which good science can thrive. Just like in musical ensembles who both compose and perform their own music, and The Beatles again make an excellent example, the whole can be much more than the sum of its parts (https://www.allmusic.com/artist/the-beatles-mn0000754032), especially when competition, commitment, and persistence are given space. The same goes for scientific excellence, where making goals unquantifiable and unspecific may lead to a poor usage of limited resources.

Taking the needs of the wider context into account is, of course, already possible (e.g. some academic job profiles call for a 'connector' who can bring together researchers from different disciplines), but this often occurs without a systematic analysis of the team's needs (e.g. is it really lacking a person who connects different research lines or are structural forces incentivizing researchers to go about their individualistic ways?). It then becomes difficult to assess the degree to which an applicant fulfills these needs (e.g. does the candidate merely claim that he or she is an open-minded and sociable person who can act as a bridge-builder?) Also in this domain, standardized assessment is fortunately available to select the best candidates that make optimal use of limited resources and optimally serve the organization's goals. It should be a priority to invest in the scientific refinement of such procedures because so much can depend on their optimal outcomes.

We express the hope that the use of performance metrics, as we discussed them here, is continued, improved, or restored, if only because quantification has done so much for the advancement of science through the ages. In the early sixteenth century, Da Vinci and Galilei used the rules of logic through quantification and formalization to change the face of physics forever, and this was pushed further by Newton and Huygens, to name just a few. Chemistry profited enormously in the nineteenth century from using insights from physics, and biology profited from categorization (Linnaeus) and employing biochemistry (DNA, Watson, Crick, Wilkins, and Franklin). What all these examples have in common is that quantification enabled researchers to create order of the apparent chaos of the real phenomena that surround us and are unintelligible without further assistance. These may look like big words in the context of scientific performance assessment but making sense—understanding the essentials—of someone's scientific work requires information that needs separation into signals and noise, even when the signal subsequently has to be interpreted and contextualized. This is where metrics, when used wisely, can assist tremendously—in combination with other sources of information.

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Note

1. For example, see the recent application form for the 2025 Replication Studies call "DORA aims to call a halt to the irresponsible use of bibliometric indicators in assessing research and researchers (such as the *h*-index, Journal Impact Factor and citations). It is a global initiative for all research disciplines (for more information about this, see https://sfdora.org/). NWO implements its principles in all instruments. Therefore it is not allowed to mention (sic) these and similar bibliometric indicators in XS-applications." See https://www.openscience.nl/sites/open_science/files/media-files/application_form_replication_studies_iv_2025_0.docx

References

Abramo, G. (2024) 'The Forced Battle Between Peer-Review and Scientometric Research Assessment: Why the CoARA Initiative is Unsound', *Research Evaluation, 2024*, 00: 1–8. https://doi.org/10.1093/reseval/rvae021

Adriaanse, L. S., and Rensleigh, C. (2011) 'Comparing Web of Science, Scopus and Google Scholar from an Environmental Sciences Perspective', *South African Journal of Libraries and Information Science*, 77: 169–78.

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019) 'Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories', *SAGE Open*, 9: 21582440198. https://doi : 10.1177/2158244019829575

Alonso, S. et al. (2009) 'h-Index: A Review Focused in Its Variants, Computation and Standardization for Different Scientific Fields', *Journal of Informetrics*, 3: 273–89. https://doi.org/10.1016/j.joi.2009.04.001

Barnes, C. S. (2016) 'The Construct Validity of the h-Index', *Journal of Documentation*, 72: 878–95.

Bordignon, F., Chaignon, L., and Egret, D. (2023) 'Promoting Narrative CVs to Improve Research Evaluation? A Review of Opinion Pieces and Experiments', *Research Evaluation*, 32: 313–20.

Bornmann, L., and Daniel, H. D. (2007) 'What Do we Know about the h Index?', *Journal of the American Society for Information Science and Technology*, 58: 1381–5.

Bornmann, L. et al. (2020) 'Disruptive Papers Published in Scientometrics: meaningful Results by Using an Improved Variant of the Disruption Index Originally Proposed by Wu, Wang, and Evans (2019)', *Scientometrics*, 123: 1149–55. https://doi.org/10.1007/s11192-020-03406-8

Campbell, D. T. (1979) 'Assessing the Impact of Planned Social Change', *Evaluation and Program Planning*, 2: 67–90. (https://doi.org/10.1016/0149-7189(79)90048-X

Collins, J., and Simmonds, P. (2024). *Lifting economy through science, tertiary sectors*. https://www.beehive.govt.nz/release/lifting-economy-through-science-tertiary-sectors

Cronbach, L. J., and Meehl, P. E. (1955) 'Construct Validity in Psychological Tests', *Psychological Bulletin*, 52: 281–302.

Dawes, R. M. (1994). *House of Cards. Psychology and Psychotherapy Built on Myth*. New York, NY: The Free Press.

Denissen, J. J. A., and Sijtsma, K. (2022) 'A New Academic Incentive Structure: Does It Fit the Psychology of Human Motives? ', *Personality Science*, 3: 18–21. Article e9227, https://doi.org/10.5964/ps.9227

DORA (2012). *San Francisco declaration on research assessment (DORA)*. https://sfdora.org/. Accessed 27 December, 2023.

Drenth, P. J. D., and Sijtsma, K. (2006) *Testtheorie: Inleiding in de Theorie van de Psychologische Test en zijn Toepassingen*. Bohn Stafleu van Loghum.

Edwards, J. R., and Bagozzi, R. P. (2000) 'On the Nature and Direction of Relationships Between Constructs and Measures', *Psychological Methods*, 5: 155–74. https://doi.org/101037//1082-989X.S.2.155

Espeland, W. N., and Sauder, M. (2007) 'Rankings and Reactivity: How Public Measures Recreate Social Worlds', *American Journal of Sociology*, 113: 1–40.

Espeland, W. N., and Stevens, M. L. (1998) 'Commensuration as a Social Process', *Annual Review of Sociology*, 24: 313–43.

Garfield, E. (1955) 'Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas', *Science*, 122: 108–11.

Goodhart (1975). 'Problems of Monetary Management: The U.K. Experience', in *Papers in Monetary Economics 1975*, pp. 1–20. Sydney: Reserve Bank of Australia.

Gossink-Melenhorst, K. (2019). Quality over quantity: How the Dutch Research Council is giving researchers the opportunity to showcase diverse types of talent. *DORA (blog), 14 November*.

Hammarfelt, B., Rushforth, A., and de Rijcke, S. (2020) 'Temporality in Academic Evaluation: ‚Trajectoral Thinking' in the Assessment of Biomedical Researchers', *Valuation Studies*, 7: 33. DOI: https://doi.org/10.3384/VS.

Hartwell, C. J., Johnson, C. D., and Posthuma, R. A. (2019) 'Are we Asking the Right Questions? Predictive Validity Comparison of Four Structured Interview Question Types', *Journal of Business Research*, 100: 122–9.

Harzing, A. W. (2023a). *Measuring and Improving Research Impact: Crafting Your Career in Academia*. London, UK: Tarma Software Research Ltd.

Harzing, A. W. (2023b). *Using the Publish or Perish Software: Crafting Your Career in Academia*. London, UK: Tarma Software Research Ltd.

Hicks, D. et al. (2015) 'Bibliometrics: The Leiden Manifesto for Research Metrics', *Nature*, 520: 429–31.

Hirsch, J. E. (2005) 'An Index to Quantify an Individual's Scientific Research Output', *Proceedings of the National Academy of Sciences of the United States of America*, 102: 16569–72. https://doi.org/10.1073/pnas.0507655102

Hirsch, J. E. (2007) 'Does the h-Index Have Predictive Power?', *Proceedings of the National Academy of Sciences*, 104: 19193–8. https://doi.org/10.1073/pnas.0707962104

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: 741–54. DOI https://doi.org/10.1007/s11192-010-0193-9

Hirsch, J. E. (2019) '*hα*: An Index to Quantify an Individual's Scientific Leadership', *Scientometrics*, 118: 673–86. (https://doi.org/10.1007/s11192-018-2994-1

Hönekopp, J., and Khan, J. (2012) 'Future Publication Success in Science is Better Predicted by Traditional Measures than by the h Index', *Scientometrics*, 90: 843–53.

Hutaibat, K. et al. (2021) 'Performance Habitus: Performance Management and Measurement in UK Higher Education', *Measuring Business Excellence*, 25: 171–88.

Ioannidis, J. (2022) 'September 2022 Data-Update for "Updated Science-Wide Author Databases of Standardized Citation Indicators', *Mendeley Data*, V5. https://doi.org/10.17632/btchxktzyw.5https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/5

Ioannidis, J. P. A. et al. (2019) 'A Standardized Citation Metrics Author Database Annotated for Scientific Field', *PLOS Biology*, 17: e3000384. https://doi.org/10.1371/journal.pbio.3000384

Ioannidis, J. P. A., Boyack, K. W., and Baas, J. (2020) 'Updated Science-Wide Author Databases of Standardized Citation Indicators', *PLoS Biology*, 18: e3000918. (https://doi.org/10.1371/journal.pbio.3000918

Ioannidis, J. P., Klavans, R., and Boyack, K. W. (2016) 'Multiple Citation Indicators and Their Composite across Scientific Disciplines', *PLoS Biology*, 14: e1002501. (https://doi.org/10.1371/journal.pbio.1002501 https://doi.org/PMID : 27367269

Kahneman, D. (2011). *Thinking, Fast and Slow*. London, UK: Penguin Books.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise. A Flaw in Human Judgment*. London, UK: William Collins.

Koltun, V., and Hafner, D. (2021) 'The h-Index is no Longer an Effective Correlate of Scientific Reputation', *Plos One*, 16: e0253397. (https://doi.org/10.1371/journal.pone.0253397

Krauss, A., Danús, L., and Sales-Pardo, M. (2023) 'Early-Career Factors Largely Determine the Future Impact of Prominent Researchers: Evidence Across Eight Scientific Fields', *Scientific Reports*, 13: 18794.

Lamont, M. (2012) 'Toward a Comparative Sociology of Valuation and Evaluation', *Annual Review of Sociology*, 38: 201–21.

Leising, D. et al. (2022) 'Ten Steps toward a Better Personality Science-How Quality May be Rewarded More in Research Evaluation', *Personality Science*, 3: Article e6029. https://doi.org/10.5964/ps.6029

Lissitz, R. W. (2009). *The Concept of Validity. Revisions, New Directions, and Applications*. Charlotte, NC: Information Age Publishing, Inc.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008) 'Improving the Peer-Review Process for Grant Applications: reliability, Validity, Bias, and Generalizability', *American Psychologist*, 63: 160–8.

Martín-Martín, A. et al. (2018) 'Google Scholar, Web of Science, and Scopus: A Systematic Comparison of Citations in 252

Subject Categories', *Journal of Informetrics*, 12: 1160–77. https://doi.org/10.1016/j.joi.2018.09.002.

McAdams, D. P. (2008). 'Personal Narratives and the Life Story', in O. P. John, R. W. Robins, & L. A. Pervin (eds.), *Handbook of Personality: Theory and Research*, pp. 242–262. New York, NY: Guilford Press.

Meehl, P. E. (1954). *Clinical Versus Statistical Prediction*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1989) 'Law and the Fireside Inductions (with Postscript): Some Reflections of a Clinical Psychologist', *Behavioral Sciences and the Law*, 7: 521–50.

Michels, C., and Schmoch, U. (2014) 'Impact of Bibliometric Studies on the Publication Behaviour of Authors', *Scientometrics*, 98: 369–85. https://doi.org/10.1007/s11192-013-1015-7

Moed, H. F. (2008) 'UK Research Assessment Exercises: Informed Judgments on Research Quality or Quantity?', *Scientometrics*, 74: 153–61. https://doi.org/10.1007/s11192-008-0108-1

Morf, C. C., Horvath, S., and Torchetti, L. (2011). 'Narcissistic Self-Enhancement: Tales of (Successful?) Self-Portrayal', in M. D. Alicke and C. Sedikides (eds.), *Handbook of Self-Enhancement and Self-Protection*, pp. 399–424. New York, NY: Guilford Press.

NWO (2022). *Science Works! NWO Strategy 2023-2026*. The Hague: NWO. https://www.nwo.nl/en/nwo-strategy-2023-2026

O'Reilly, C. A., and Chatman, J. A. (2020) 'Transformational Leader or Narcissist? How Grandiose Narcissists Can Create and Destroy Organizations and Institutions', *California Management Review*, 62: 5–27.

Polanyi, M. (1962) 'The Republic of Science: Its Political and Economic Theory', *Minerva*, 1: 54–73.

Roe, R. A. (1983). *Grondslagen Der Personeelsselectie [Foundations of Personnel Selection].* Assen, The Netherlands: Van Gorcum.

Ruscio, J. et al. (2012) 'Measuring Scholarly Impact Using Modern Citation-Based Indices', *Measurement: Interdisciplinary Research and Perspectives*, 10: 123–46. https://doi.org/10.1080/15366367.2012711147

Ryan, R. M., and Deci, E. L. (2000) 'Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being', *American Psychologist*, 55: 68–78. (https://doi.org/10.1037110003-066X.55.1.68

Schölvinck, A. F. et al. (2024) 'How Qualitative Criteria Can Improve the Assessment Process of Interdisciplinary Research Proposals', *Research Evaluation*, 33: rvae049.

Singh, V. K. et al. (2021) 'The Journal Coverage of Web of Science, Scopus and Dimensions: A Comparative Analysis', *Scientometrics*, 126: 5113–42. https://doi.org/10.1007/s11192-021-03948-5

Schönbrodt, F. D. et al. (2025) 'Responsible Research Assessment I: Implementing DORA AND Coara for Hiring and Promotion in Psychology', *Meta-Psychology*, 9.

Schmidt, F. L., and Hunter, J. E. (1998) 'The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings', *Psychological Bulletin*, 124: 262–74.

Schmidt, F. L., Ones, D. S., and Hunter, J. E. (1992) 'Personnel Selection', *Annual Review of Psychology*, 43: 627–70.

Schreiber, M,. (2008) 'A Modification of the *h*-Index: The -Index Accounts for Multi-Authored Manuscripts *hm*', *Journal of Informetrics*, 2: 211–6. https://doi.org/10.1016/j.joi.2008.05.001

Simonton, D. K. (1988). *Scientific Genius: A Psychology of Science*. Cambridge: Cambridge University Press.

Sinatra, R. et al. (2016) 'Quantifying the Evolution of Individual Scientific Impact', *Science*, 354: https://doi.org/10.1126/science.aaf5239

Sijp, W. (2018) *Paper Authorship Goes Hyper: A Single Field is behind the Rise of Thousand-Author Papers*. News Feature, Nature Index. https://www.nature.com/nature-index/news/paper-authorship-goes-hyper

Strinzel, M. et al. (2021) 'Ten Ways to Improve Academic CVs for Fairer Research Assessment', *Humanities and Social Sciences Communications*, 8: 1–4.

Taylor, S. E., and Brown, J. D. (1988) 'Illusion and Well-Being: A Social Psychological Perspective on Mental Health', *Psychological Bulletin*, 103: 193–210.

Taylor, H. C., and Russell, J. T. (1939) 'The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection. Discussion and Tables', *Journal of Applied Psychology*, 23: 565–78.

Torres-Salinas, D., Arroyo-Machado, W., and Robinson-Garcia, N. (2023) 'Bibliometric Denialism', *Scientometrics*, 128: 5357–9. https://doi.org/10.1007/s11192-023-04787-2

Van den Besselaar, P., and Sandström, U. (2015) 'Early Career Grants, Performance, and Careers: A Study on Predictive Validity of Grant Decisions', *Journal of Informetrics*, 9: 826–38.

Van den Besselaar, P., and Sandström, U. (2019) 'Measuring Researcher Independence Using Bibliometric Data: A Proposal for a New Performance Indicator', *PLoS ONE*, 14: e0202712.

van den Besselaar, P., and Sandström, U. (2020) 'Bibliometrically Disciplined Peer Review: On Using Indicators in Research Evaluation', *Scholarly Assessment Reports*, 2: 5. DOI: https://doi.org/10.29024/sar.

Van der Aalst, W. M. P. (2022). *Yet Another View on Citation Scores*. LinkedIn Pulse Article. https://www.linkedin.com/pulse/yet-another-view-citation-scores-wil-van-der-aalst, accessed 27 Dec. 2023.

Van der Aalst, W. M. P., Hinz, O., and Weinhardt, C. (2023) 'Ranking the Ranker: How to Evaluate Institutions, Researchers, and Conferences?', *Business & Information Systems Engineering*, 65: 615–21. https://doi.org/10.1007/s12599-023-00836-5

VSNU, KNAW, NWO (2020). Strategy Evaluation Protocol 2021–2027. The Hague. https://storage.knaw.nl/2022-06/SEP_2021-2027.pdf

Waks, L., and Kramer, E. O. (2023) 'Bibliometrics and Qualitative Assessment: A Pragmatist Approach', *Contemporary Pragmatism*, 20: 150–68.

Waltman, L., and van Eck, N. J. (2012) 'The Inconsistency of the h-Index', *Journal of the American Society for Information Science and Technology*, 63: 406–15.

Wiesner, W. H., and Cronshaw, S. F. (1988) 'A Meta-Analytic Investigation of the Impact of Interview Format and Degree of Structure on the Validity of the Employment Interview', *Journal of Occupational Psychology*, 61: 275–90.

Wilsdon, J., et al. (2015). The metric tide. *Report of the independent review of the role of metrics in research assessment and management*. https://doi.org/10.13140/RG.2.1.4929.1363

Wiggins, J. S. (1973). *Personality and Prediction: Principles of Personality Assessment*. Reading, MA: Addison-Wesley.

Woolston, C. (2021) 'University Drops Impact Factor', *Nature*, 595: 462.

Appendix A

Case study describing how Dutch research grant supplier NWO shifted from structured performance listing to narrative overview between 2015 and 2025

Introduction

The Netherlands has a relatively sophisticated research infrastructure, and its research universities typically perform relatively well in international rankings (Jongbloed, 2018). After Dutch academia was rocked by the Stapel scandal, which was widely seen as resulting predominantly from an excessive focus on quantitative performance, the Netherlands put itself at the forefront of changes in research evaluation based on multiple performance dimensions and the qualitative evaluation of such dimensions (van der Meer, 2023). According to one of the major policy officers of the Netherlands Organization for Scientific Research (NWO, in Dutch) at the time (Hoogstraat, 2022), an important motivation of the change was the desire to counter the dominance of prolific academics who (due to various structural advantages) would produce more than their peers, and therefore also dominate quantitative metrics, like the *h*-index.

Much of Dutch research funding comes from NWO, which invests approximately one billion euro in supporting various types of grants, research infrastructure, and awards (NWO, 2025). The grants NWO supports include so-called personal grants, which are intended to reward the most innovative and impactful (broadly defined) researchers. The recipients receive a grant of between €320.000 (Veni, ≤ 3 years post PhD) and €1.5 million euro (Vici, 8-15 years post PhD). In our analysis, we focused on the Vici scheme, but the same shift in emphasis has been applied to other NWO talent schemes, and actually the narrative CV was first piloted in the Veni scheme.

Method

To illustrate the movement away from quantitative indicators and towards more qualitative ones, we requested documents from the Dutch NWO funder's Vici scheme. For each submission year, we obtained three documents: 1) the pre-application form, 2) the full application form, and 3) the explanatory notes. Because the ultimate decision depends on the full form, we extracted questions and explanations that were used to describe the profile of the submitting applicant, including a description of the focus and quality and their research. We then added the relevant information from the explanatory notes to the corresponding questions, using a different color. This allowed a full and nuanced analysis of all instructions and requirements for applicants to describe their quality as a researcher. The resulting document on which the following results are based can be found in the online supplementary materials, with permission from NWO.

Results

Description of material

Based on content analysis by the first author, it became clear that there are four stages of NWO's evolution of the Vici form: The 2015-2019 stage, the 2020-2022 narrative CV, the 2023 transition year, and the 2024-2025 evidence-based CV. In the following, each of these four stages is described in some detail.

The 2015-2019 forms structured applicants' information into fixed categories, resulting in open-ended descriptions for each category. In the "curriculum vitae" section, applicants would list input on "academic staff supervised", "international activities", "other academic activities", as well as "scholarships, grants and prizes". The latter category required users to sum up grant money totals. It also required users to list work experience and calculate the "months of research", to facilitate a fair comparison between researchers. Of note, this section also features a narrative account of the "the content of your research over the last five years", with a 250-word limit.

The forms also featured an "output" section that is limited to two full pages, where applicants could list five top publications related to the proposal and provide a rationale for selecting them. They could also provide additional output. The form accommodated the use of impact factors and even (until 2019) required applicants who added impact factors to their publication output to list the median impact factor for their field. This latter question was skipped from the 2019 form, although this form did include an option to disclose one's Clarivate Researcher ID in the form.

Starting in 2020, the 2020-2022 forms introduced the so-called narrative CV, a semantic shift that involved many items

of the previous "curriculum vitae" section to be moved to a separate and novel section on "administrative details", which now contained aspects of the previous "listed CV" of the 2015-2019 version, namely "personal details", "master's degree", "doctorate", and "work experience since completing your PhD". However, it no longer contained a requirement to list information categories, such as academic staff supervised, international activities, other academic activities, or scholarships, grants and prizes. That said, these information categories were listed as examples that participants could mention in the narrative CV.

In the narrative CV, participants were asked to "address your research focus, research agenda and vision and focus on how you achieved this." Using 1200 words in total, users could list any activity they wanted, with the instructions providing a number of examples, including prizes, awards and grants. It also included as an option for the user to reflect on the "motivation for doing research in general and this project in particular". Applicants were explicitly instructed to refrain from references to "publication metrics, or expected/future output and […] not [to] mention (total) numbers of publications."

The form of 2023 appeared as transitory but radical. There was no longer a specific "CV section", and only a short "personal information" section requiring participants to provide their name and contact information. Tabulated information about the accumulated work experience was no longer required and the only way to describe the profile of the researcher occurred in the section with the "description of the proposed research" (max. 14 pages). There, a subsection was available that required the applicant to reflect on the "alignment between research proposal and expertise", so the focus was on fit rather than a decontextualized measure of quality. This form no longer provided any instruction how to craft the narrative. Moreover, it was specified that "CV information" (sic), which is explained to mean aggregated bibliometrics, was explicitly prohibited.

Finally, the forms of 2024-2025 presented a deviation from both the 2020-2022 and 2023 forms. Compared to the 2023 form, the term CV was re-introduced, but it was now called an "evidence-based CV", thus departing from the term "narrative CV" that was used in the 2020-2023 forms. The explanatory notes greatly expanded, perhaps to help users familiarize themselves with the new format. In spite of the name change, narrative accounts still played a role in a section on academic profile, with a 1200 words limit. This part did include narrative accounts of a "general academic profile", and "leadership and mentorship".

This section to describe the applicant's "academic profile" adopted an agnostic approach towards "listing": It was permitted to structure the narrative, but not necessary: "You are free to shape your narrative in any way to suit your profile. You may for example choose to simply describe your academic profile in running text, add highlights by using bold or italic, choose to add structure via subheadings, list achievements point by point followed by an explanation, etc." Candidates were asked to "provide context and *evidence* of how the elements you choose to include show your academic qualities." Aggregated information (even number of papers, volume of research money) was explicitly forbidden in these sections, as were labels that linked to reputation (e.g. "leading"). At the end of the form, there was also a section in which candidates provide information on "work experience since completing your PhD" and "net academic research time".

In the output section of the 2024-2025 form, participants were allowed to choose up to 10 outputs. Each of these outputs could be described in terms of a maximum of three indicators, and explain them in writing under "motivation". The available indicators were 37 in number, and roughly an equal number of scholarly and societal/outreach indicators were available, in addition to "personal development" and "other, please describe". Interestingly, five different types of citations were allowed: Next to the "total number" of citations, users could provide a "sentiment analysis". In addition, three field-normalized citation indicators were possible: article field weighted citation impact, percentile benchmark, and relative citation ratio.

Assessment of psychometric criteria

Beginning with the 2015-2019 form, we rated it as medium on our Fairness Criterion. It is clear that the form focuses on scholarly impact, but applicants can also submit a "societal impact" section. It is positive that the form takes research time into account, although it does not do so in a formal way (e.g., by calculating output per year, or correcting the *h*-index for academic age). Regarding the Standardization Criterion, the 2015-2019 form scores well: It has all applicants list relevant dimensions of information, including academic staff supervised, and international activities. The "brief summary of your research over the last five years" has a clear 250-word limit. Regarding the Objectivity criterion, we rated the form relatively well, because impact factors and citations are objective, data-driven indicators. The Field Comparability Criterion appears relatively well covered in that participants are required to provide median impact factors for their field. However, if *h*-indices are provided, this index is not field-normalized. Finally, we submit that the form scores relatively good on the Predictive Validity Criterion, because citations (in whatever aggregation mode) are known to predict future impact and recognition. That said, it is correct that the impact factor is indeed only a proxy of expected citations for a particular contribution, and that the *h*-factor has ceased to be predictive of recognition in recent years due to strategic behavior by researchers.

We also assessed the 2020-2022 form on our five criteria in what follows. To begin with the Fairness Criterion, it still contained a section requiring participants to list their work experiences and compute the amount of time spent on research, which allows for a correction of some kind – yet what and how to correct is no longer obvious given the lack of quantitative information in the other sections. Removing the use of metrics is introduced as improving fairness (because metrics are biased to some degree), yet abandoning metrics (instead of trying to fix bias) also introduces problems concerning objectivity. One could argue that removing the requirement to list academic staff supervised, international activities, etc. would allow a more diverse group of applicants to make a convincing case, although it comes with disadvantages on our Standardization

Criterion. That is, participants are no longer required to list these activities (although they can voluntarily bring them up in the narrative CV), making the forms less comparable. Third, the form seems to deteriorate on our Objectivity Criterion, because objective metrics can no longer be used, and participants are not required to list any evidence of impact. In our view, the fourth Field Comparability Criterion suffers the most from the 2020-2022 changes in format. By removing a section that requires participants to provide field-specific benchmarks and relying only on narrative information, it becomes more difficult to compare across fields. The same is true for the use of absolute citation numbers per chosen output item, which can vary very much across fields. Finally, we think that the fifth Predictive Validity Criterion deteriorated, because metrics (especially when they are corrected for bias) are good predictors of scientific performance but are no longer relied upon. The 2020-2022 forms also no longer required users to reflect on what seems to be information that is predictive of success, like experience with supervision.

In our view, the 2023 form scores particularly bad in terms of psychometric quality, and it is not surprising that NWO abandoned it after one year. In terms of the Fairness Criterion, one could argue that it removes any barriers to receive funding for reasons that are unrelated to the submitted proposal, by focusing on the (narrative) "alignment between research proposal and expertise". However, as discussed in the main paper, the danger is that this introduces other types of bias, for example one that favors self-aggrandizement. It also does not include the calculation of work experience, which could control for bias. Second, the Standardization Criterion is compromised as the form is completely unstandardized and requires no specific information about the profile of the candidate, only a reflection about the fit. Third, the Objectivity Criterion is not met because the form is not evidence-based in any sense, which is clear from the absence of the requirement to provide proof of fit. Fourth, there is a complete lack of comparability of fields and so the form fails to meet the Field Comparability Criterion. Finally, the Predictive Validity Criterion is not met. Although person-work fit is indeed discussed as potential criterion of performance, the evidence is mixed and many studies have found null results (Van Vianen, 2018). Aspects that do predict scientific performance (e.g., metrics, past experience), on the other hand, are not considered.

Finally, the 2024-2025 form represents an improvement by bringing back useful elements of the old forms. In terms of the Fairness Criterion, there are some improvements. For example, applicants again calculate their net academic research time to potentially correct for bias, although it is again not clear how this correction needs to be implemented. The focus on individual output items seemingly provides a level playing field between high-quantity and low-quantity researchers, although it should be noted that high-quantity researchers would have a great pool of items to select from, which likely re-introduces the bias that NWO wants to eliminate. Second, the 2024-2025 form scores better on the Standardization Criterion, both in the academic profile (where it requires reflection on the general profile as well as leadership and mentorship) as well as the output section, which includes a drop-down menu with fixed impact indicators. There is a lack of standardization, however, in that participants likely report on very different output indicators (even within one category, as is the case with citations), and can selectively leave out information. Third, regarding the Objectivity Criterion, the 2024-2025 form is indeed more evidence-based (hence its name) because it requires applicants to provide specific reasons for claims of impact. However, some of the indicators are rather subjective and difficult to back up by evidence at all, such as "originality/novelty" or "personal development". Fourth, the 2024-2025 form performs unevenly in terms of our Field Comparability Criterion. Applicants can choose sophisticated citation metrics that do take account of field differences, but they can also select the raw number of citations, which does not. Finally, the fifth Predictive Validity Criterion is only partially met. Sophisticated individual-level citation metrics (that can be controlled for bias) are not used, and the scientific validity of many of the output indicators (e.g., social media activity) is uncertain.

Conclusions and recommendations

When applying our psychometric criteria to analyze the (evolving) content of the NWO Vici forms, it becomes clear that the 2015-2019 forms performed relatively well across criteria. It seems to us that it would have been possible to incrementally improve this form, for example by introducing more sophisticated corrections for academic field differences, corrections for available research time, and the statistical separation of research quantity and quality (see Sinatra et al., 2016). Instead, NWO decided for a more radical approach, which arguably compromised the psychometric quality of the assessment process (especially in 2023). Since then, however, significant improvements in psychometric quality have been made.

Our psychometric criteria can be used to derive recommendations for further improvement. Although this is to some extent subjective and requires the incorporation of additional (i.e., non-psychometric) considerations, our framework points to certain areas for improvement. Regarding the Objectivity Criterion, we think it would be wise to re-introduce personal metrics (i.e., individual-level indicators) because these are anchored in traceable data, but to correct them for important biases (e.g., academic age, authorship inflation) so as not to endanger the Fairness Criterion. A promising avenue is to use the work of Sinatra et al. (2016) to disentangle quantity and quality of research output, because both would seem to be important predictors of a candidate's ability to use the grant money efficiently. For example, one of the most promising indicators, which is age-independent and field-specific is the propensity of an author to produce papers that belong to the 10% most widely cited in their field. Regarding the Standardization Criterion, it would be important to rigorously implement standardization both in the form's output section, for example by requiring all participants to list both social and academic impact descriptions, with fixed indicators for each (in contrast to the current flexibility to pick different indicators of the same construct, such as the different citation indicators listed in the description of the 2024-2025 form. In particular, for our Field

Comparability Criterion it is important that as many indicators as possible should be corrected for field differences.

For the narrative parts of the form, it is important to introduce fixed input categories that are linked to scientifically validated predictors of scientific performance (improving the score on the Predictive Validity Criterion) in a standardized way (e.g., with fixed word limits, thus improving the Standardization Criterion). Also other parts of the assessment procedure should be focused on scientifically established predictors and avoid formats that would seem to more bias-prone, such as using unstructured interviews to assess quality.

In general, and in line with the main tenet of our paper, we would encourage formats that stimulate a dialogue between qualitative and quantitative indicators, which NWO also seems to aspire to in the introduction of the evidence-backed CV. However, the integration can also flow from quantitative to qualitative information, such as using bibliometric information for a first screening of applicants which can then be evaluated qualitatively, or to let authors and reviewers contextualize performance on metric indicators in narrative terms.

References

Hoogstraat, R. (2022). *On narrative CV*. recognitionrewardsmagazine.nl/2022/narrative-cv

Jongbloed, B. (2018). *Overview of the Dutch science system [CHEPS Working Paper 04/2018].* Enschede: Center for Higher Education Policy Studies.

NWO (2025). *Jaarverslag 2024*. Den Haag: NWO.

van der Meer, M. (2023). How 'Recognition and Rewards' in Dutch Academia Turned Metrics into Incentives. https://blog.trialanderror.org/recognition-rewards

Van Vianen, A. E. (2018). Person–Environment Fit: A Review of its Basic Tenets. *Annual Review of Organizational Psychology and Organizational Behavior*, 5:75-101.

Research Evaluation, 2026, 35, 1–19
https://doi.org/10.1093/reseval/rvag023
Article