Research Evaluation, 2024, **33**, rvae037
https://doi.org/10.1093/reseval/rvae037
Special Issue Paper

OXFORD

(Dis)harmonic styles of valuation: A study of academic justification across research domains and levels of assessment

Björn Hammarfelt [1,*], Claes-Fredrik Helgesson [2], Gustaf Nelhans [1], Erik Joelsson [1]

[1]Swedish School of Library and Information Science, University of Borås, Allégatan 1, Borås, 50190, Sweden
[2]Centre for Integrated Research on Culture and Society (CIRCUS), Uppsala University, Box 513, Uppsala, 751 20, Sweden
*Corresponding author. Email: bjorn.hammarfelt@hb.se.

Abstract

Disciplines display field-specific ways of valuing research contributions, and these different 'styles of valuation' influence how academic careers are assessed and formed. Yet, differences in how research is evaluated are also prevalent between different levels of assessment: collegial and organizational. Consequently, we employ a multifaceted two-dimensional approach for studying styles of valuation where both horizontal (across domains) and vertical (organization levels) variations in assessment practices are examined. For this purpose, we make use of 16 faculty guidelines and 112 referee reports concerning candidates for becoming 'docent' (Habilitation) from four broad domains: the humanities, the social sciences, medicine and the natural sciences (including technology). By inductively identifying five broad dimensions used when assessing publication merits: (1) *Attribution of work*, (2) *Qualities of content*, (3) *Publication channel*, (4) *Publication impact*, and (5) *Publication volume* we can distinguish specific styles of valuation for each of our four domains. Moreover, by extending the analysis to an organizational level we detect opposing ways in which the evaluations are justified—what we call 'disharmonic styles of valuation'. Thus, when developing insights on 'quality understandings'—and their operationalization through styles of valuation—in academia we need to put less emphasis on their origins and rather focus on how they come to travel between and co-exist within specific evaluative contexts.

Keywords: academic evaluation; research quality; styles of valuation; docent; guidelines; referee reports.

1. Introduction

The evaluation of individual researchers and the entailing consequences that certain valuation practices might have on research has become a topical question in research policy (Lyall 2019; Pardo-Guerra 2022). Assessment procedures in academia, and especially those relying on quantitative measures, have been studied quite intensively in the last decade (de Rijcke et al. 2016), yet the consequences of specific evaluation systems are difficult to study, and distinct causal effects are rarely found (Thomas et al. 2020). A large portion of studies so far has focussed on national and local systems, yet attention has also been devoted to assessment procedures and their consequences for the individual researcher (Pontika et al. 2022). A key insight from previous studies is that evaluation practices tend to differ between research fields, and the way physicists, historians, economists or medical researchers assess projects, papers, and colleagues may vary substantially (Hemlin and Montgomery 1993; Lamont 2009). Overall, such studies have found differences in the weight attached to specific criteria (such as rigour, innovativeness, and style), yet it is also evident that there is a rather strong agreement on important aspects of 'quality' across disciplines. However, when approaching how valuation is performed we find larger disparities. Such differences can for example concern the degree to which metrics or other quality proxies are used, or how authorship is defined and assessed (Hammarfelt 2017; Langfeldt, Reymert and Aksnes 2021; Reymert 2021). Consequently, disciplinary differences in the practice of a what we refer to as 'styles of valuation' may impact how research is conducted, and how academic careers are formed.

In this paper, we employ a multifaceted two-dimensional approach for studying variations in research assessment. Consequently, we ask: *How do domain-specific styles of valuation differ between domains, and how do styles of valuation harmonize with understandings of quality on the field and organization level?* Thus, when studying styles of valuation we aim for both a horizontal analysis (how can styles of valuation be identified across research fields?) as well as a vertical one (how can styles of valuation be used to understand differences in assessment procedures on different levels—organization and collegial—of academia?). Using a scale with five inductively defined dimensions of value we are thus able to compare valuation styles across four domains and two levels of application. Moreover, the vertical dimension, which differentiates between individual peer assessment and institutional guidelines, provides an empirical setting in which the distinction between S- (societal) and F-type (field) understandings of research quality as proposed by Langfeldt et al. (2020) can be critically discussed.

In our analysis, we examine 16 faculty guidelines and 112 evaluation protocols of candidates for becoming 'docent' (Habilitation) across four domains: the humanities (HUM), social sciences (SOC), medicine (MED), and the natural sciences (including technology NAE'). We argue that the evaluations performed in this process share many similarities with observations from previous studies on the evaluation of candidates for academic positions (Hammarfelt and Rushforth 2017; Reymert 2021; Reymert, Jungblut and Borlaug 2021). As our theoretical focus is on how 'research' is assessed we have (like the studies cited above) limited our analysis to

discussions regarding publications, including their impact. Other important merits such as teaching, outreach, administration and leadership have not been included in the current analysis. It is important to note, however, that the docent evaluation has unique features compared to studies focusing on academic hiring in that it focuses more on 'certification' (is the candidate good enough?) rather than 'ranking' (which candidate is best?). The assessments for 'docentship' could be understood as a collegial certification process where both the applicant and the assessor expect a positive outcome (being judged as eligible for the title).

The assessment process tied to the docent title is oriented to what Boltanski and Thévenot (2006: 33) call *justness* (deserving the title) rather than *fitness* (being the most appropriate candidate). Hence, in contrast to many previous studies on disciplinary assessment, we focus on a justification process rather than a competition in which only one or a few candidates may be granted a position or a grant. Hence, the guidelines and referee reports are not only part of a 'gate-keeping machinery' that decides who becomes a docent. Stressing this as a justification process is further appropriate since it emphasizes the public justification of these evaluations which, as Bowen (2020) reminds us, very well might differ from the actual evaluation. As part of public justifications, the guidelines and reports affirm a sense of community within academia where these values are to be shared by stating what counts and should be counted in specific domains and disciplines. This is a particularly salient feature of individual assessments, which provide ample room for re-producing disciplinary styles of valuation.

A subsequent section will outline the assessment process for becoming a docent. Before that, we will introduce styles of valuation and how these relate to the understanding of quality. Then follows a description of the methods and materials used, in which the analysed guidelines and the referee reports are presented. Using an inductive thematic analysis, we proceed by inductively identifying five general dimensions which together comprise a style of valuation—*attribution*, *content*, *channel*, *impact*, and *volume*—for describing how justification and assessment are performed across domains and organizational contexts. We then relate our empirically grounded observations of different styles of valuation to the framework for understanding research quality developed by Langfeldt et al. (2020). The analytical section discusses and illustrates 'styles of valuation' across domains and organizational levels using multidimensional visualization of valuation styles. The paper concludes by discussing a few general implications for research assessment.

2. Disciplinary assessment and styles of valuation

Scholars discussing interdisciplinary research practices have argued that '[d]isciplines have survived for so long in the academic world in part because they serve the very useful function of constraining what the researcher has to think about' (Lyall et al. 2011: 95). These constraints encompass assessment procedures, and peer review studies have analysed how different domains and fields assess quality (for a recent overview, see Forsberg et al. 2022). For example, Lamont (2009) showed how field-specific notions of quality shape the evaluation of grant proposals in multidisciplinary panels. Similar observations are made by Hemlin and Montgomery (1993) in their study of assessment procedures in Swedish academia, in which they found distinct differences between 'soft' sciences, in which writing style and theory were emphasized, in contrast to the 'hard' sciences, where, for example 'international contacts' was deemed necessary for displaying quality. Moreover, assessment practices have changed over time and so has the style in which evaluations, and CVs, are represented (Nilsson 2009; Hamann and Kaltenbrunner 2022).

Recent studies on distinct ways of reasoning around metrics show that in fields where indicators and measures are important, evaluators tend to become experts not only in judging quality but also in assessing the usefulness of tools for measuring, such as journal impact factor (in biomedicine) or journal rankings (in economics) (Hammarfelt and Rushforth, 2017; Hylmö, 2018). In contrast, Reymert, Jungblut and Borlaug (2021) found that researchers in physics tend to put less emphasis on the number of publications and instead focus on the broader suitability for the position in terms of the topicality of previous research. In the humanities, on the other hand, evaluators tend to put less emphasis on citations, while being mentioned in a review or receiving prizes can be an indication of academic recognition (Hammarfelt 2017; Söderlind and Geschwind 2020).

Styles of valuation, as applied here, is a broader concept compared to 'notions of quality' or 'assessment criteria' which are used in several of the studies mentioned above. We define styles of valuation as collectively rooted and historically situated ways of judging what is to be valued and how (Lee and Helgesson 2020). In this sense our use of the styles of valuation resonates with a practice-theory perspective where activities are viewed as learned within a specific community; it encompasses how valuation is done and performed in a certain group. To a degree 'styles of evaluation' resonates with the notion of 'evaluation devices' as this concept emphasizes 'how evaluations are achieved in practice by combining various elements …' Brunet and Müller (2022: 488). However, as we here focus on written statements and the public justification (cf. Bowen 2020) of evaluations we find that 'style' is a more appropriate description of our empirical interest. Importantly, we are focusing on documents that are performative in the sense that they provide legitimacy and justification for the assessment made. To paraphrase Kuipers and Franssen (2020: 150) we are interested in how others are persuaded that this particular candidate is a 'good candidate for a docentship' rather than what a 'good candidate' is. In other words, various 'notions of quality' within specific domains may foreground specific styles, yet it is how specific quality dimensions are mobilized and expressed for the purpose of justification that is of essence here.

Moreover, we are interested in other differences beyond the substantive knowledge areas at hand. An example of such differences would be how to justify the assessment of co-authored papers (see Biagioli 2003 on a discussion on large-scale multiauthorships). Notably, different styles of valuation may co-exist within the same site or field of knowledge production, and new 'devices' of evaluation may 'unsettle the status quo' or 'provoke repetitions of well-rehearsed clashes between entrenched styles of valuation' (Lee and Helgesson 2020: 664).

Studies of how different styles, and methods, of evaluation function across organizational levels in academia are rare. One way of discussing variation in the valuation process is to position actors within an evaluative landscape in which actors interact with various values, assessment procedures and measures (Brandtner 2017; Nästesjö 2024). In terms of

evaluating research across institutional and organizational levels Langfeldt et al. (2020) have developed a conceptual apparatus which distinguishes the so-called F-type and S-type understanding of quality. Research quality can in their framing be comprehended through two basic types: field-type (F) and space-type (S). The field-type notions of research quality are those that originate in a collegial disciplinary context, whereas S-type notions are notions that originate in policy and funding spaces (Langfeldt et al. 2020: 119). The former includes matters such as properties of knowledge and are enforced in peer assessment settings. The latter is more established and maintained outside the specificity of research fields. For example, traditional peer review is in this scheme primarily related to an F-type understanding of quality. At the same time, proxies like the Journal Impact Factor are more related to S-type notions. Interestingly, Langfeldt et al. (2020) point to the 'research organisational level', where F-type and S-type views on quality meet and often conflict. So, rather than understanding these notions as fixed, we focus on how S and F-type quality notions are employed and upheld rather than on their origin and initial use. Of particular interest is therefore 'sites where F-type and S-type research quality notions most obviously meet (and quite possibly collide)' (ibid.: 128). The appointment of docents, which could be said to be located at the intersection of collegial and organizational structures, thus appears to be a fruitful site for exploring how S- and F-type quality notions are employed when reviewers justify their assessment.

3. The assessment for 'docent' in the context of Swedish academia

The word 'docent' originates in the Latin phrase 'venia docendi', meaning the right to teach. Today, however, the title 'docent' is mainly associated with research merits rather than pedagogical skills, and it can be seen as a step between the doctorate and full professorship. Internationally, 'docent' can be compared to the German Doctor Habilitatus ('Dr habil.') which customarily is given based on a 'Habilitationsschrift', a second monograph after the dissertation. France ('habilitation à diriger de resechers') and Italy ('Abilitazione Scientifica Nazionale') have similar systems. The tradition of Habilitation is reflected in the formal requirements for docent in Sweden, which is often expressed in terms of having performed research equivalent to an 'additional dissertation' after finishing the PhD.

Guidelines for evaluating docent competence are central because nothing concerning the title is regulated in the Swedish national constitution or agreement. The guidelines for docent assessment regularly span many different disciplines. Consequently, a guideline is negotiated to accommodate and regulate the docent assessment for several disciplines. Hence, an individual referee report is regulated by a guideline that accommodates differences in styles of valuation across the disciplines within its purview. In this sense, guidelines may be seen as a site in which S- and F-notions meet, rather than a context which is completely dominated by either societal or field understandings of quality.

As indicated above, the process for becoming a docent differs slightly between universities and faculties. Generally, the applicant seeks approval at the granting institution that, if awarded, allows the candidate to proceed with a formal application to the faculty board (or equivalent). In some cases, the board makes a first vetting whether the application should be treated (Brommesson and Erlingsson 2012). Given a positive outcome, in the following step expert referees are selected (so-called 'sakkunniga') who are at least docents—but usually full professors—in the same field as the applicant. The referees, usually one or two, are provided with documentation consisting of a CV, a written summary of achievements, a publication list, and in many cases a selection of specimens on which they base their evaluation. These assessments are given in referee reports, where the expert reviews the applicant's merits and recommends whether he or she should be appointed the title docent. Usually, the form, length, and level of detail of the statements vary across disciplinary domains, but a length of a couple of pages is the most typical. Based on the referee reports, if affirmative, the board then recommends the dean (or equivalent) to appoint the applicant as a docent (ibid.). Referee reports as well as the guidelines that regulate the evaluation, are the empirical material analysed in this paper.

4. Studying guidelines and referee reports

The selection of materials was guided by an effort to cover a broad range of fields within all domains: humanities and arts (HUM), social sciences (SOC), medicine (MED), and natural sciences, including technology (NAE). Six institutions with different characteristics were chosen. Chalmers University of Technology (CTH) and Karolinska Institutet (KI) were selected as specialized universities with only one faculty (technology and medicine, respectively), while Lund University (LU), Stockholm University (SU), and University of Gothenburg (GU) were chosen as examples of older and comprehensive universities. In contrast, Linnaeus University (LN) was selected as an example of a 'new' university, established in 2010 through a merger of two university colleges.

Typically, there are specific documents containing guidelines for docent promotion at each HEI (Higher Education Institution). However, slight variations at different faculties and HEIs are common, for example in one case (LN), the guideline was—rather than being a separate document—incorporated into the employment regulations' document for the whole HEI in question. Likewise, guidelines vary in depth and scope; some consist of specific instructions for applicants and/or expert referees, while others contain general information regarding the promotion procedure. As inclusion criteria, we established that each document should state the qualifications for docent promotion, as well as cover the faculties to which the studied referee reports belong, which resulted in a population of 16 guidelines (see Table 1). Generally, the broad, comprehensive universities such as LU and GU have eight guidelines each, roughly corresponding to the number of faculties. In contrast, Linnaeus University, a younger university, only had a single guideline for all faculties. Similarly, specialized universities (KI and CTH) only have a single guideline for the whole university.

4.1 Collecting and analysing referee reports

In 2017—the data used here was collected in 2018—there were 488 docent applications at the chosen universities.[1] These documents are the outcome of the distant peer review done by external referees, and a considerable part of the legitimacy of the process of becoming a docent depends on the justification provided by these reports. We wanted a selection of referee reports from each relevant HEI/domain combination. However, since

Table 1. Overview of the 16 examined guidelines, including year of publication

HEI	HUM	MED	NAE	SOC
CTH	n.r.	n.r.	2010	n.r.
GU	2017	2016	2014	Year missing—SOC 2014—EDU
KI	n.r.	2014	n.r.	n.r.
LN	2014 (university-wide)	n.r.	n.r.	2014 (university-wide)
LU	2014—HUM 2017—ART	2014	2017—NAT 2016—ENG	2015
SU	2013	n.r.	2015	2016
No. of guidelines	5	3	5	5

n.r., not relevant for this HEI; ART, area of arts; ENG, engineering; EDU, educational sciences; NAT, natural sciences excluding engineering; HUM, humanities; MED, medicine; NAE, natural sciences & engineering; SOC, social sciences.
Regarding LN, there is a university-wide guideline, which is why it appears in two columns (HUM and SOC). Thus, we consider it as two guidelines.

Table 2. Number of cases and selected referee reports across domains

Domains	Total no of cases in domain	Main fields targeted for report selection	Total selected no of reports examined	Total number of docent cases examined
HUM	73	Languages and Literature	32	23
MED	232	Surgery	17	12
NAE	88	Astronomy, Chemistry, and Physics	23	18
SOC	95	Political Science and Pedagogy/Educational Sciences	40	25

the population was not evenly distributed, we targeted specific fields with enough applications for each HEI instead of using a stratified sample. We randomly selected three applications for each domain/HEI combination within these chosen fields. This gave us a broad and varied material, yet with the drawback that our material does not allow for direct comparison across institutions (see Table 2 for an overview of collected referee reports). Overall, the material provides us with a wide variety in terms of more applied fields (surgery), professional fields (pedagogy), as well as disciplines which traditionally are viewed as more 'pure', like physics and literature. However, a consequence of this selection method is that our material is more suitable for comparing *across* domains rather than examining variations *within domains*. Still, the choice of specific fields for each domain is a limitation that should be considered when reflecting upon the findings. For example, selecting surgery within the medical domain may render somewhat different results than if we had opted for biomedicine or genetics.

Since some applications contained more than one referee report, the complete sample consists of 112 referee reports[2] related to 78 applications for promotion. For the four domains, this means that we have examined between 43% and 7% of all referee reports available in each domain for 2018 (for an overview, see Table 2). Direct quotes from referee reports have been anonymized so that individual applicants or reviewers cannot be identified.

Based on the five dimensions identified (see Table 3), we employed a directed approach for content analysis (see Hsieh and Shannon 2005) of the refeere reports. Furthermore, we describe the degree to which these dimensions occur, where possible centres of gravity lie, how they are specified, and how they can (possibly) be related to each other. We worked with automatic keyword coding in Atlas.ti (a software for qualitative data analysis) and manual coding based on reading the documents. Automatic coding is particularly suitable when we deal with precise keywords such as 'H-index' or 'impact factor'. Notably, refeere reports are not a direct description of how an assessment is performed, but are, as discussed previously, a justification of the assessment made. These documents thus have a performative role, and their public justificatory function can make them differ from how the actual assessment has been done (cf. Bowen 2020). Yet we argue that it is still possible to infer some aspects of the evaluation from the documentation. For example, if a referee mentions citation counts from a specific database we suspect that a search has been performed in the stated database, if a referee refers to the number of total pages produced, they have probably counted, or if referees allude to specific claims in an assessed publication we suspect that they have read, or at least skimmed the text. Furthermore, including such aspects in the report indicates that these are aspects that are legitimate when justifying the evaluation.

5. Reading guidelines: five dimensions for valuing publication merits

What dimensions to describe publication merits are mobilized at one time or another when justifying that a candidate can be awarded the title docent? In this section, we aim to generate a crude inventory of all dimensions observed in our corpus of guidelines and then examine how they are articulated in them.[3] Our method for studying the guidelines can be described as 'conventional content analysis', based on an inductive approach (see Hsieh and Shannon 2005).

Through a detailed reading, we derived and coded key concepts appearing in the guidelines. Concepts were then sorted into emerging themes resulting in five broad dimensions for assessing publication merits: (1) *Attribution of work* (single-authored or co-authored); (2) *Qualities of content* (the use of theory, method, style etc.); (3) *Publication channel* (where something is published: book, high ranked journal etc.); (4) *Publication impact* (citations); and (5) *Publication volume* (number of publications or pages). These five broad

Table 3. Five dimensions for assessing publication merits identified in the 16 guidelines examined

Assessment dimension	Illustrative quote from a guideline (our translations)
A. Attribution of work	'Note that the applicant should be listed as the last author on at least one of the articles written after the dissertation and the first or last author of an additional of said articles. Furthermore, these articles should be done without any of the previous supervisors listed as co-authors'. (GUI-MED-LU)
B. Qualities of content	'The quality of the production should be significantly above the minimum requirements of a PhD dissertation. Criteria for this are theoretical and methodological awareness, capacity to develop fresh ideas and independent scholarly work that has produced new knowledge, and capability to present such results'. (GUI-HUM-SU)
C. Publication channels	'Excellent quality of original work means original publications published in relevant international journals with recognised high quality for the field'. (GUI-MED-KI)
D. Publication impact	'Number of citations for the publications (with database information)'. [An item requested in the application] (GUI-NAE-LU)
E. Publication volume	'The applicant should have deepened and expanded his/her research experience and should, in addition to the PhD dissertation, present scientific publications that at the least correspond to an additional dissertation'. (GUI-SOC-GU)

dimensions are presented in Table 3 below alongside an exemplifying quote from the guidelines.

These dimensions are to be seen as compounded by many emic (Harris 1976) notions to capture different facets for assessing publication merits. Each compounded dimension therefore accommodates several different facets. For instance, this is readily observable in the illustrative quote for the dimension 'qualities of content', which spans both epistemic qualities ('theoretical and methodological awareness'), assessment of novelty ('fresh ideas'), and style ('capability to present such results'). This means that two statements evoking the same dimension in different guidelines might contain subtle differences in what is meant by 'Qualities of content'. The latter should also, given its emic basis, not be confused with a conceptually and rigorously defined notion of 'research quality'. As to links between conceptualizations and our inductively deduced dimensions, the most 'S-type' notion found in the guidelines is the dimension of publication impact (D). However, one could argue that aspects of the publication channel dimension, operationalized into impact factor, would be more of an S-type notion. The mixture of S and F-type notions of research quality found in the guidelines should not be a complete surprise if we follow Langfeldt et al. (2020).

Significantly, dimensions can be compounded in a given guideline, even close to one another. To take one example: The quote provided as an example of the publication channel dimension ('in for the field relevant international journals with recognised high quality') also contains the phrase 'original publications', which in the subsequent paragraph is defined as articles containing new original data in contrast to such things as review articles. Hence, the statement on 'original publications' relates to something that in our reading belongs to the dimension 'qualities of content'. Together, these dimensions are key components of specific ways—here conceptualized as valuation styles—of justifying assessment.

5.1 Guidelines in practice: negotiating criteria and procedures

The referee reports exhibit a reasonably wide range regarding whether and how they relate to the assessment guidelines, from not mentioning them in the statements to making substantial comments regarding them. For example, it is not uncommon to discuss existing criteria or suggest new ones:

> 'For the assessment of quality, in my opinion, in addition to the requirements specified in the Faculty of Social Sciences' guidelines, special emphasis should be placed on such aspects as originality in problem formulation, design and methodology, conceptual precision and argumentation, as well as the degree of consistency in the evaluation of results and conclusions […]' (REF-SOC-GU-070).

So, while guidelines provide an important framing for evaluating candidates for 'docent', the referees greatly influence how these guides should be interpreted. Many external referees seem to consider guidelines as loose recommendations rather than as a 'binding set of rules'. Hence, they could be said to navigate, and negotiate between styles of evaluation as formulated in guidelines, and their own practice of how to perform and justify assessment. This, in turn, results in a greater variety and detail when studying the actual referee reports. The following section will discuss the five dimensions and how they are used in candidates' assessments (e.g. referee reports) across fields.

6. Assessing researchers: the five dimensions in practice

The first dimension concerns the attribution of work and focuses on how authorship is discussed across domains. Deliberations on authorship are more present in the examined referee reports from medicine and social sciences (53% and 48%, respectively) than they are in the humanities (16%) and natural sciences (17%). Researchers in the natural sciences and technology (NAE) and medicine (MED) focus on the organizational aspects of authorship as an indication of independence. Of particular importance for understanding how works should be attributed and credited is the authorship order: 'Most of the articles after the PhD are not co-authored with any of his previous supervisors. He is the last author of one article (2016) and the first of two'. (REF-MED-KI-44). Being the last author signals seniority—being the main responsible (principal investigator), while first authorship indicates who did most of the work (experiments). Similarly, the relation to supervisors reflects the degree of independence. Independence is an important dimension in the assessment reports from the social sciences, yet here mentions of independence are more directly attributed to specific works: 'This work is in my view one of the best and of most merit for NN, which is further accentuated by him being the *sole author*'. (REF-SOC-SU-105, our *italics*). Similarly,

qualitative statements about the collected works of the applicant are common:

> 'The text is clearly independent and innovative in that there is not much research on these phenomena in Baltic countries'. (REF-HUM-SU-031a)
>
> and
>
> 'I assess that she clearly […] exhibits "tangible independence, extensive reading, the ability for scientific argumentation and good judgment in theoretical and methodological matters"'. (REF-SOC-GU-076)

It should be noted that all referee reports within NAE, except one, mention independence (REF-NAE-CH-36). This report, which favourably assesses the applicant, focuses more on collaboration and interdisciplinarity of the researcher and states that while in their recent research, the applicant obviously is not the only one to appreciate a specific research agenda, s/he has found a niche of their own (REF-NAE-CH-36). Notably here is thus that the dimension of authorship is important across domains, yet it is assessed and justified rather differently within the specific styles of valuation identified in our study.

6.1 Qualities of content: going deep or broad

To a large extent, the referee reports in the humanities and social sciences focus on the depth and breadth of the research when assessing the specimens of the application. In contrast, the assessment reports in the medical and natural science and technology domains do that to a lesser degree and in an entirely different way. Depth and breadth are mentioned once in a guideline in both the medical sciences (MED-KI) and the humanities and social sciences (SOC-LU and HUM-GU), including the all-encompassing guideline at Linnaeus University (ALL-LN). However, in the referee reports, 'depth' and 'breadth' are almost only found in the humanities and social sciences. The single report that mentions width in the medical domain only does this concerning the applicants' multiple stays abroad: 'Through these stays abroad, [s/he] shows a clear internationalisation in [his/her] way of working, an aspiration to broaden [his/her] research activities and international collaborations'. (REF-MED-KI-21). This quote resonates well with Hemlin and Montgomery's (1993) observation that 'international contacts' is an important quality criterion in the medical sciences.

Examples of broadness as a dimension of value are also evident in the natural sciences, although much less often when compared to the humanities and social sciences, in which discussions regarding breadth and depth, often in conjunction with each other, are prominent. Two distinct quotations can be shown as examples, one from each domain:

> 'On the basis of what I have stated along the way, I would like to conclude that [NN]—especially in the contributions [#1] and [#2]—has shown scientific quality at a high level. Both monographs treat their themes with exhaustive breadth and depth'. (REF-HUM-GU-075).

In a quote from the social sciences, where the assessor first ponders on the applicants' tremendously versatile production in terms of moving both between fields and the use of a wide array of theories and methodological approaches, which is evaluated as a 'strength', the assessor nevertheless asserts that

> '[I]t also means that NN has not yet had a chance to consolidate his research in substantial monographs with university presses. This is a bit of a shame as the article format does not quite allow for the amount of discussion and depth that NN:s work surely deserves'. (REF-SOC-GU-081).

Here, several different evaluation practices are brought to the fore. The view that the format of the article does not allow for depth is only expressed in the humanities and social sciences. At the same time, it shows that in these domains, breadth and depth go together in the oeuvre of the exemplary academic. Moreover, the focus on broadness could, at least in fields such as history, be interpreted in terms of also being able to teach a range of subjects and periods.

We have already seen examples of how referee reports do more than simply conform to the assessment dimensions and criteria stipulated in the guidelines. This is even clearer regarding specific technical terms related to dimensions of publication channels and impact.

6.2 The valuation of publication channels: reputation and ranking

In this section, we are interested in expert opinions highlighting the publication channel as an indicator of quality. Foremost we focus on whether referees discuss 'ranked outlets', with two specific yardsticks in focus: Journal Impact Factor and the 'Norwegian list' (Sivertsen 2016). An important finding is that these terms were frequent in the referee reports despite rarely being mentioned or recommended in the guidelines.

Three of the referee reports in the medical domain explicitly refer to Journal Impact Factor (JIF) as a yardstick for ranked journals: 'Among these are several works published in highly reputable surgical journals with a high impact factor. He also has four published review papers/book chapters'. (REF-MED-KI-39). Similarly, JIF is mentioned in referee reports concerning natural scientists, where the measure is linked to mentions of 'well-regarded' or 'relevant' research or statements suggesting that 'a significant fraction of the publications are in high impact factor journals' (REF-NAE-SU-70). Yet, referee reports may contain more diffuse references to 'ranked', 'well-reputed', or 'high impact' journals. One statement in the humanities (out of a total of two occurrences) makes such claim without qualifying it: 'The publication channels chosen are reputable, and here are several examples of internationally published articles as well as a publication written in English. All in all, these conditions provide a good basis for the continued qualitative assessment'. (REF-HUM-SU-31).

Unlike 'Impact factor', which is found in statements from all scientific fields except the humanities, references to the Norwegian list are found only in the social sciences and the humanities (one occurrence). In the social sciences, all three mentions of the Norwegian list form parts of balanced assessments of the respective researchers' entire output: 'He has also in recent years started publishing in very reputable journals (level 2 according to the "Norwegian list")'. (REF-SOC-SU-106).

Overall, experts use the reputation or ranking of publication channels to measure a specific publication or as a weighted judgement of the applicant's entire production. In addition, experts can relate to a rising trend in the choice of publication channels (e.g. aiming for better-ranked journals). We can only speculate, but one reason for emphasizing the quality of the channel, rather than the impact of individual works that we will examine next, is that it is difficult to measure such impact for ongoing and recently published research. Thus, when evaluating candidates for docent—the application is often made quite close (within ∼4–10 years) after receiving the PhD—at least in some fields, it might be difficult to assess impact.

6.3 Assessing the impact of the research

Impact, here broadly understood as the influence that research has had in academia and society, is more explicitly discussed in the natural sciences and medicine. Societal impact—a typical S-type quality criterion—is rarely discussed. Instead, citations are the primary method for evaluating impact. The natural and social sciences stand out in terms of the experts' consideration of the applicants' citations. In the natural sciences, 35% of the statements studied discussed citation rates (8 out of 23 referee reports). These range from fairly general findings that the applicant's publications have been cited (REF-NAE-CH-55) or well-cited (REF-NAE-CH-56; REF-NAE-GU-61) to more detailed reviews of citation frequencies where experts are impressed by high citation frequencies and h-index: 'He is very well cited for his age with h = 23, over 1400 citations, two papers with over 100 citations, and a steadily growing citations score. This is an excellent publication record showing both good impact and productivity'. (REF-NAE-SU-68). The mention of high citation frequencies often seems to substitute for the need to familiarize themselves with the actual research content.

A more general reflection on the importance of being cited as a researcher considers that a regular citation rate over the years is supposed to carry much weight in the scientific community (REF-NAE-SU-67). Similarly, a referee suggests that an H-index of 20 is impressive for the applicant at his career stage as it signals an even citation rate instead of single publications with many citations (REF-NAE-GU-61). However, we find examples where caution with bibliometric measures is advocated, for instance, where a research area has been significantly hyped due to inflated expectations from the research community rather than actual research results. 'Even if it shows N's ability to understand experimental data and imagine possible original interpretations, I think that its large number of citations (like for all papers on this particular subject) is due to its current popularity within the community (for an effect that in the end disappeared) rather than to its intrinsic quality'. (REF-NAE-CH-52). These findings point to referees who are not only experts in how to evaluate research but also in how specific measures are valued in these fields. Hence, as argued by Hammarfelt and Rushforth (2017), referees become experts not only in assessing quality but in judging which metrics should be used and how.

Regarding academic impact, the social sciences have a more similar style of valuation to the natural sciences and medicine than to the humanities. Likely, this is due to the increasing focus on English language journals as the primary outlet within social sciences, which allows for a more systematic evaluation of research through citation scores. A quarter of the referee reports in the social sciences contain reasoning about citation impact, while only one out of 32 reports in the humanities includes such a rationale.

Generalizable conclusions cannot be drawn based on numbers alone, but our observations correspond well with previous findings showing that impact factors are essential when assessing the merits of medical researchers; journal rankings play a role in the social sciences, while these types of proxies play less of a role in traditional humanities (Rushforth and de Rijcke 2015; Hammarfelt and Haddow 2018; Langfeldt, Reymert and Aksnes 2021). Notably, citation scores and indicators are rarely recommended in guidelines, but these are incorporated on the initiative of external referees themselves. Hence, in this context, they are evoked on the field level (F-type) rather than imposed from above (S-type).

6.4 Volume of research: How much is enough?

The volume of research is the most straightforward dimension identified, and the equivalent of an additional doctoral thesis is an estimate given in almost all guidelines. In contrast, others stipulate twice or three times that volume to qualify as a docent. Depending on the field, it might be translated into a specific number of articles or a monograph. Counting pages may be used to provide a more exact measurement in the humanities and social sciences. The number of book pages written is one of the few quantitative indicators used when evaluating candidates for professorships in history (Hammarfelt 2017). This practice, we suggest, shows that the second monograph (or the Habilitation thesis) is still the golden standard in the humanities. While articles can be a substitute, these are counted by the page rather than by the number of articles:

> 'The general rule regarding volume for attaining the title "docent" is the production of one additional monograph, or research publications equivalent to one monograph. NN has achieved this in the attached articles, which together amount to more than 200 pages'. (REF-HUM-GU-4)

Yet, as with all the dimensions identified in our study, evaluating these criteria can seldom be separated from other aspects of the assessment. Hence, while volume often is seen as perhaps the most central component—without enough publications the evaluation of impact, quality, or authorship cannot take place—some evaluators emphasize that 'quality' should always be a key concern:

> 'In practice, attaining the docent title is mainly a question about "having produced the additional work required" after achieving the PhD. This is also reflected in the guidelines as cited above. I take the liberty to add another aspect, namely if the demonstrated quality of research in exact terms significantly exceeds that of a PhD-thesis'. (REF-SOC-GU-81)

As illustrated by this quote, the five assessment categories we identified are often intermingled and connected. While acknowledging the interrelatedness of dimensions, our ambition is that the classification presented here may be helpful when analysing how styles of valuation relate to disciplinary identity and collegial community. Such a perspective suggests that the issuing of guidelines and the production of the

referee reports are part of the academic practice that reproduces disciplinary styles of valuation. For example, we find that humanist scholars tend to write more extended reports and quite often engage with the epistemic content of the research done by the candidate. In medical and natural sciences and technology, evaluations are shorter and more focused on channels and impact. These differences, we argue, can be attributed to how research is organized in different fields, where factors such as audience, divisions of work tasks, epistemic orientation, and agreement on procedures and practices may influence how evaluation is performed.

The differences across domains in how publication merits are assessed mobilize different dimensions of value. This means that it would be difficult for someone working in, for instance, the substantive area of literature to become docent *if* said scholar had a publication portfolio that at the surface, but not in content, mimicked how it is supposed to be done in medicine. The many multi-authored papers of such a strange scholar, to highlight one likely characteristic of such an imagined portfolio, would not fit how publication merits are assessed within the humanities. The guideline would provide little guidance on distinguishing the nuances of the first and last author, and it would be unfamiliar terrain for the referees. It would be equally valid in the opposite case. In other words, there is more at play when awarding the title of docent than assessing that a candidate is sufficiently epistemically rooted in the given disciplinary area. The candidate's portfolio must also be sufficiently aligned with the prevailing style of valuation in the domain where she or he operates. It is these differences that we have chosen to call differences in the *style of valuation*.

7. Visualizing domain-specific styles of valuation

This section presents a concerted analysis of manifest styles of valuation across both horizontal (across domains) and vertical (across organization level) levels of analysis. More specifically, we aim to find a way to visualize the specific styles of valuation in each domain as they are expressed in both the guidelines and the actual referee reports. The ambition is to quantify and visualize the relative presence of the five dimensions in organizational guidelines and individual referee reports (Figure 1). The primary data here is based on the detailed coding of where these dimensions are discussed in the guidelines ($n = 16$)[4] and the referee reports ($n = 112$).

The visualization highlights the relative weight of different dimensions in the reviewer reports and guidelines. This thus provides an insight into 4×2 different styles of valuation. Each variable is measured as relative frequencies for the population of guidelines and reviews where the most frequent dimension in the respective corpus is 1. For a detailed description of the method, see Supplementary Appendix SI.

The visual depictions of the different styles of valuation are telling when put next to one another. The first level of comparison would be between the four domains. Here we see the already discussed differences in how little or much weight is given to the different dimensions. Take 'attribution', for instance, which in guidelines in medicine is among the three most important dimensions and is largely absent in the humanities and social sciences guidelines. We can further see how the most important dimension in humanities guidelines by far is 'content'.

A second comparison level looks at each of the four guideline-referee report pairs. Here our depictions reveal both some striking congruences and differences. At first glance, there is, for instance, a remarkable congruence (or harmony) in the pairs for the humanities. It is worth noting that nuances in this comparison very well may be affected by the way we selected what referee reports to examine. For medicine, where we focused on surgery, we examined only 7% of the total number of referee reports available within medicine for 2017. For the other three domains, we examined between 26% and 43% of all reports (see Table 2 and Supplementary Appendix SII for additional information). The striking congruence apparent in the humanities is thus based on examining 43% of all reports available for 2017.

There is some congruence for the other three couples. Yet, there are also differences. In the humanities, discussions about 'attribution' are found in a quarter of the 32 referee reports examined. We did not find this dimension discussed at all in any of the three examined guidelines. There are other striking differences *within* each of the pairs. In the social sciences, two-thirds of the referee reports mention 'attribution', and almost all of them discuss 'impact', whereas these two dimensions are almost entirely absent in the guidelines. This we interpret as a disharmony between the valuation styles expressed in guidelines and referee reports. In medicine, all examined referee reports mention 'impact', whereas this dimension is marginal in the guidelines. In natural sciences, finally, there is a striking difference where 'volume' is assessed much more in the guidelines than in the examined referee reports and, conversely, where 'impact' is assessed in a quarter of the referee reports but not at all in the guidelines. Hence, all four domains contain one or a few striking differences in what is given weight between our selection of referee reports and guidelines. These observed differences may well be graphically exaggerated if we happened to select particularly non-compliant disciplines within each domain. Yet, the difference is an indisputable affirmation that referees might systematically employ styles of valuation which differ from those proposed by the relevant guideline.

It should, finally, be stressed that all these notable differences within each guidelines-referee reports pair are essentially different, with one striking exception. The dimension of 'impact' is markedly more prominent in the referee reports than the corresponding guidelines in all areas except the humanities. 'Impact' has made its way into many referee reports in these three areas but is not present in the guidelines.

When analysing these findings in relation to the Field-type (F) and Space-type (S) understandings of quality, we find that referee reports more frequently make use of S-type factors (e.g. impact) than their corresponding guidelines. The most pertinent example is some domains' relatively frequent mention of citations and h-index. This relative frequency is surprising since these, according to Langfeldt et al. (2020), originate in policy and funding spaces, not in the collegial context of peer review. If this observation holds, we have thus here an example where such factors seep more into the more field-specific level of individual assessments than into the research organizational level to which the policy space is more adjacent. Consequently, we find that the S- and F-type understandings of quality, as proposed by Langfeldt et al. (2020), are not only dependent on organizational levels, with F-type being more prominent on the group and individual levels, while S-types are more prevalent on the aggregated policy level.

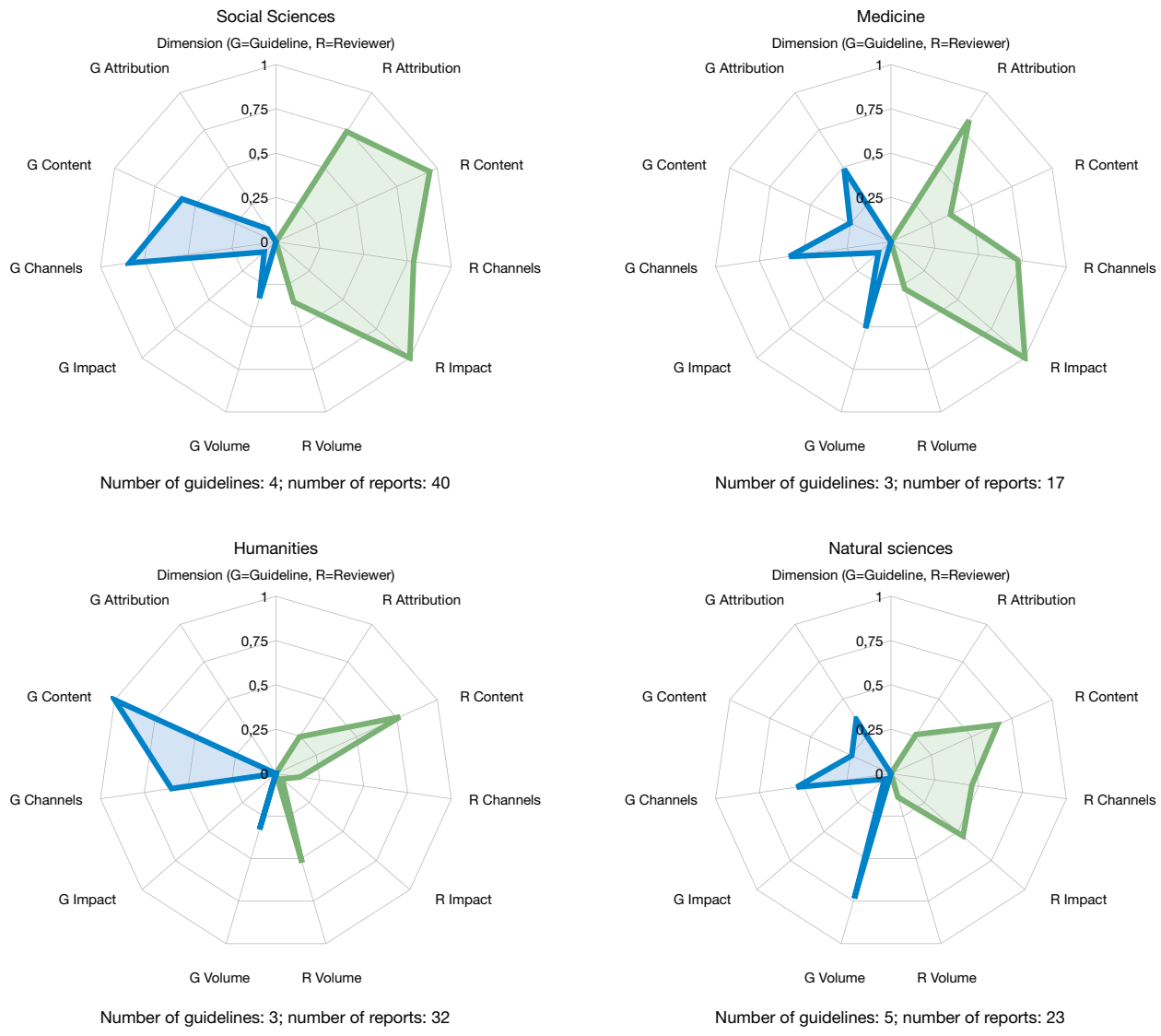

Figure 1. Styles of valuation within guidelines and reviewer reports of five dimensions across four domains (see Supplementary Appendix SII for the underlying data).

Field-specific notions of quality, or what we label as 'styles of valuation', may be more or less inclined to use such measures. Still, it is important to consider that, for example, citations may, on one level—say national comparisons of highly cited papers—be viewed as an S-type notion, yet when used in the context of peer review, this might very well be seen as proxies for collegial recognition (F-type). In fact, one referee even mentions citing a candidate's publication in their own work as a sign of quality. Hence, the distinction between S- and F-type of quality notions is difficult to implement when looking at how specific dimensions are used in justifying the assessment of docents.

What is shown here is that in practice, different justifications of quality—and the proxies attached to them—are always interconnected. Any disentanglement—also the five dimensions of value proposed here—is a simplification of messy and interconnected valuation and justification processes. These visualizations effectively show the degree to which styles of valuation in academia are in accordance with each other. In other words, if the valuation style expressed in guidelines (blue) corresponds with those described in (green). Suppose this is not the case—as, for example, in the social sciences—we have a *disharmonic style of valuation*, a setting in which rivalling public justifications (cf. Bowen 2020) of valuations co-exist within a fixed evaluative context. Such a situation may increase the degree of uncertainty for junior academics (cf. Nästesjö 2021). On the other hand, the presence of *disharmonic styles of valuation* may stimulate more openly reflexive discussions on research quality and assessment criteria by making implicit and taken-for-granted values visible.

8. Conclusions and implications

Our analysis offers an approach for comparing across fields that can be utilizable in a broader discussion on disciplinary styles of valuation and their relation to organizational changes and policy initiatives. The studied domains have many similarities: they all employ the five dimensions (attribution, content, channels, impact, and volume) for the same title, 'docent'. Yet, the emphasis of these different dimensions varies depending on the organization of the field in terms of

aspects such as division of labour, potential audience as well as epistemological characteristics. As our findings indicate, questions of attribution are more important in fields where multiauthorship is common (e.g. medicine), while the type of research conducted in terms of broadness may be significant in a field where teaching competencies are highly valued (e.g. humanities). For example, in the quotes above are monographs mentioned as examples of how broadness, as well as theoretical depth, can be displayed. In a field like medicine however, broadness, might rather be exhibited through international connections. Such observations of differences in styles of valuation are not unique in the literature, and studies have shown how styles of valuation in fields like history, biomedicine, or economics partly are dependent on social and intellectual structure These differences may depend on how densely populated a field is (the more intensive, the more inclined use of quantitative measures), the intended audience (peers or a general public) and the degree to which a distinct disciplinary 'elite' in terms of researchers and journals exists (Hammarfelt 2017). Economics, for example, has a strong orientation towards a distinct core, which has in turn resulted in a few prestige journals being able to dominate the field (Hylmö 2018). In the style of writing refeere reports, we also find different temporal emphasis—economists tend to evaluate in terms of 'future potential' while historians tend to focus on the beginning (doctoral thesis) (Hammarfelt 2017).

Yet, here, we are also able to discuss individual assessments in relation to the guidelines which are supposed to direct individual referees in their evaluation. What becomes evident then is the loose connection between guiding documents and actual practices, something that may be characterized as *disharmonic styles of valuation*. Hence, for some domains—like the social sciences in our case—there might be a disconnect between the recommended style of valuation as expressed in the guidelines, and the actual style used by individual assessors when justifying their assessments. Such disharmony between styles of valuation on different levels could signal that understandings of quality are uncertain and under negotiation. However, it may also be the case that disharmony between guidelines and assessment is due to great variation among the fields included within the domain of the social sciences and that they comprise many very diverse communities. So, in interpreting the domain specific styles of valuation one should consider that our selection of specific fields may have influenced how different dimensions are valued. Hence, selecting other research field for representing the domain—say biomedicine instead of surgery, or economics rather than political science—could result in that some dimensions are further emphasized, while others are less accentuated.

With these field specific caveats in mind the vertical dimension—as opposed to the horizontal comparison of fields—allows us to discuss and visualize how dimensions of value travel between an organization level and a more collegial (and disciplinary) context. Hence, our findings suggest that guidelines, even if issued on the level of faculties, are interpreted and operationalized differently when used by particular referees when justifying their assessment to perhaps more specific communities. There may thus be, although this is a question not addressed in this study, systematic differences between different disciplines operating under the same effective guideline. Indeed, a recurrent discussion in research evaluation is if more elaborated guidelines, including distinct achievement measures, are advantageous for groups, such as women, whose merits often have been undervalued in assessment procedures. But as our study shows, even if guidelines are comprehensive, they may be employed somewhat differently depending on how the external referee interprets them in the actual assessment. Hence, rather than proposing a more standardized and steered procedure, it would probably be better to strive for a more open and reflexive process in which taken-for-granted 'styles of valuation' are made explicit and, if deemed discriminating, questioned and revised (cf. Helgesson and Sjögren 2019).

On a more theoretical level, our findings are valuable when reflecting on the idea of S-type (policy-level) and F-type (disciplinary) understandings of quality. We observe that understandings of quality—such as impact factors or citations—which are viewed as primarily S-type in the categorization proposed by Langfeldt and colleagues, are relatively prevalent in assessing docents, especially in the natural sciences and medicine. Hence, JIF may be seen as a legitimate indicator to justify an assessment in one domain (medicine) while not in another (humanities). Notably, they are employed by referees due to their own judgement of what is a legitimate evaluation method and not due to instructions from guidelines (cf. Hammarfelt and Rushforth 2017). In fact, we argue that the origins of quality notions and criteria are, in this context, not very interesting, as its first inception may not explain its current use. For example, the Norwegian list was first initiated as a policy initiative but has, over time—especially in the domains of humanities and social sciences where other metrics have been lacking—also been incorporated on the (collegial) field level. Another example of such use is the ERA list which was used briefly as a formal nationwide evaluation in Australia. Yet, researchers may still incorporate it in assessing themselves or others long after it was formally abandoned (Hammarfelt and Haddow 2018).

Similarly, measures that may today be seen as integrated into a more formal apparatus of evaluation—such as the h-index—were invented by an individual scientist within the field-specific context of physics. Another famous measure, the journal impact factor, was given its role as an essential marker of quality by individual researchers in STEM disciplines that made use of it. Hence, we argue that the legitimacy of the JIF as an S-type notion of quality builds upon its status and use as an F-type notion in certain high-prestige fields. Moreover, our findings indicate that there are reason to re-consider any notion that S-type quality indicators might evenly percolate via guidelines to actual referee practices. Indeed, such indicators appear, in our cases at least, to be more readily used in referee practices than in the organizational process that shapes official assessment guidelines. Overall, then the S/F distinction may serve as a useful typology for reflecting upon quality understandings on a abstract level, yet when confronted with how assessments are justified in actual practices this dichotomy is of less use. Thus, when developing insights on 'quality understandings'—and their operationalization through styles of valuation—in academia we need to put less emphasis on their origins and rather focus on how they come to travel between and co-exist within specific evaluative contexts.

Nonetheless, our analysis supports the observation that assessments of individual researchers often have field-specific elements, and this line of reasoning can fruitfully be applied to the assessment of docents as a rather traditionalist, F-type-based evaluation procedure. In fact, its role as a certification procedure—is the candidate qualified? rather than a comparative effort of rankings: which candidate is best for the position?—ties

the assessment procedure tightly to notions of belonging to a particular field or discipline. Hence, we argue that the evaluation of docents is governed to a higher degree than many other evaluations in academia to ideas of how an ideal senior member should be. Being a process of *certification* rather than *competition* also means that the style of valuation takes other forms compared to, for example, the evaluation of candidates for professorships (Langfeldt, Reymert and Aksnes 2021). Indeed, detailed metrics, such as citation scores, which may be advantageous in a situation when evaluators have to distinguish between many qualified candidates, while the need for such 'judgment devices' is less obvious when the fitness of a specific candidate is assessed (Hammarfelt and Rushforth 2017). While various metrics are used in the referee reports studied here they take on the distinctive character of 'testimony' which means that they are not employed in order to make comparison possible across cases. This is one of the ways in which styles of valuation concerning 'docent' distinguish itself from related assessment procedures in academia.

How reviewers conform to or deviate from guidelines in the evaluation of researchers is however a question that could be further elaborated upon by interviewing or studying reviewers' actual work in more detail. Here we have only been able to retrospectively analyse written documentation which serves, as we have argued, an appropriate view on how assessments are justified. Yet, a more ethnographical approach may provide further details into how guidelines actually come to shape individual assessments and possible differences between how assessments are made and how they are justified (cf. Bowen 2020).

Another distinct feature of our case is that the assessment is performed in a relatively limited collegial sphere, lacking the 'stakeholder heterogeneity' that Franssen (2022) highlights as necessary for promoting research that tackles significant societal challenges. Moreover, research positions, such as professorships, today often target broad areas of inquiry, meaning interdisciplinary competencies are necessary. Becoming a docent in some fields also means that the candidate should have broadened their research focus, yet this widening of interest should generally take place *within* a distinct discipline. Hence, a lingering question is how interdisciplinary settings may relate to different, sometimes disharmonic styles of valuation. Which styles of valuation prevail in a context where established practices are lacking or disputed? Such a context may possibly be more inclined to follow organizationally imposed guidelines compared to established and well-bounded disciplines. Consequently, that valuation styles are so tightly intertwined with disciplinary identity and boundary-keeping poses a problem for increasing interdisciplinary efforts. However, we do not argue that disciplinary-specific styles of valuation should be abandoned, or synchronized across domains. Rather, recognizing disciplinary variety in valuation styles, and the degree to which they (dis) harmonize with policy understandings of research quality may help develop broader, more inclusive, and flexible systems for evaluating academic research.

Acknowledgements

We thank all participants for the valuable feedback received at the 'Unsettling Research Quality' workshop arranged by R-Quest in Oslo 2022. Moreover, we are grateful for the encouragement and valuable advice provided by the editors of the special issue, Siri Borlaug, Thomas Franssen, and Liv Langfeldt. The funding for data collection and initial analysis was provided by The National Library of Sweden as part of an inquiry into the presence or absence of open access publishing as part of merit assessments in Sweden.

Supplementary data

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

Conflict of interest statement. None declared.

Notes

1. The material used here was originally collected for a report commissioned by the National Library of Sweden (Joelsson, Nelhans and Helgesson 2020).
2. Notably, the style in which the referee reports are written differs between different fields of science. Both the humanities and the social sciences use evaluators who write in Swedish and English as well as in Danish and Norwegian. In contrast, a considerable proportion of statements in the Natural sciences are written in English. However, all statements are written in Swedish within surgery (MED), probably due to the selection of a more applied speciality within the medical domain.
3. Notably, open access/open science was not mentioned in any of the guidelines or referee reports. Hence, making research accessible as open science was not seen as a merit in the context of becoming a docent. Similarly, gender issues, such as reflections about inequality between men and women, were not mentioned in any of the guidelines. Though in the preprinted assessment form used by Chalmers University of Technology, the following statement was given (italics in original):

   '*At Chalmers, we strive to attain equality between men and women in our recruitment and promotion processes. Hence, we are sensitive to inequalities with respect to gender that may occur in the scientific assessment of the candidates. Such inequalities include the risk of judging female candidates less independent than male candidates, the risk of attributing a higher development potential and a larger originality to male candidates, as well as the tendency of writing longer assessment reports on male candidates*'. (NAE-CH-(all))

4. N.B. Due to its university-wide scope, one guideline (LN) was excluded from the analysis resulting in 15 guidelines being used for calculating the figures.

References

Biagioli, M. (2003) 'Rights or Rewards? Changing Frameworks of Scientific Authorship', in M. Biagoli, and P. Galison (eds) *Scientific Authorships: Credit and Intellectual Property in Science*, pp. 253–80. Routledge.

Boltanski, L., and Thévenot, L. (2006) *On Justification: Economies of Worth*. Princeton University Press.

Bowen, J. R. (2020) 'Justification', in J. R. Bowen, N. Dodier, J. W. Duyvendak, and A. Hardon (eds) *Pragmatic Inquiry: Critical Concepts for Social Sciences*, pp. 113–27. Routledge.

Brandtner, C. (2017) 'Putting the World in Orders: Plurality in Organizational Evaluation', *Sociological Theory*, 35: 200–27.

Brommesson, D., and Erlingsson, G. Ó. (2012) 'Vad Krävs i Praktiken För Att Bli Docent?', *Ekonomisk Debatt*, 40: 5–18.

Brunet, L., and Müller, R. (2022) 'Making the Cut: How Panel Reviewers Use Evaluation Devices to Select Applications at the European Research Council', *Research Evaluation*, 31: 486–97.

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., and Hammarfelt, B. (2016) 'Evaluation Practices and Effects of

Indicator Use—A Literature Review', *Research Evaluation*, 25: 161–9.
Forsberg, E., Geschwind, L., Levander, S., and Wermke, W., eds (2022) *Peer Review in an Era of Evaluation: Understanding the Practice of Gatekeeping in Academia*. Palgrave Macmillan.
Franssen, T. (2022) 'Enriching Research Quality: A Proposition for Stakeholder Heterogeneity', *Research Evaluation*, 31: 311–20.
Hamann, J., and Kaltenbrunner, W. (2022) 'Biographical Representation, from Narrative to List: The Evolution of Curricula Vitae in the Humanities, 1950 to 2010', *Research Evaluation*, 31: 438–51.
Hammarfelt, B. (2017) 'Recognition and Reward in the Academy: Valuing Publication Oeuvres in Biomedicine, Economics and History', *Aslib Journal of Information Management*, 69: 607–23.
Hammarfelt, B., and Haddow, G. (2018) 'Conflicting Measures and Values: How Humanities Scholars in Australia and Sweden Use and React to Bibliometric Indicators', *Journal of the Association for Information Science and Technology*, 69: 924–35.
Hammarfelt, B., and Rushforth, A. D. (2017) 'Indicators as Judgment Devices: An Empirical Study of Citizen Bibliometrics in Research Evaluation', *Research Evaluation*, 26: 169–80.
Harris, M. (1976) 'History and Significance of the Emic/Etic Distinction', *Annual Review of Anthropology*, 5: 329–50.
Helgesson, K. S., and Sjögren, E. (2019) 'No Finish Line: How Formalization of Academic Assessment Can Undermine Clarity and Increase Secrecy', *Gender, Work and Organization*, 26: 558–81.
Hemlin, S., and Montgomery, H. (1993) 'Peer Judgements of Scientific Quality: A Cross-Disciplinary Document Analysis of Professorship Candidates', *Science Studies*, 6: 19–27.
Hsieh, H. F., and Shannon, S. E. (2005) 'Three Approaches to Qualitative Content Analysis', *Qualitative Health Research*, 15: 1277–88.
Hylmö, A. (2018) 'Disciplined Reasoning: Styles of Reasoning and the Mainstream-Heterodoxy Divide in Swedish Economics', Doctoral dissertation, Lund University.
Joelsson, E., Nelhans, G., and Helgesson, C.-F. (2020) *Hur Värderas Publiceringsmeriter i Det Svenska Akademiska Systemet? En Undersökning av Värderingen av Befordran till Docent Med Särskilt Fokus På Betydelsen av Öppen Tillgång*. Stockholm: Kungliga biblioteket.
Kuipers, G., and Franssen, T. (2020) 'Qualification', in J. R. Bowen, N. Dodier, J. W. Duyvendak, and A. Hardon (eds) *Pragmatic Inquiry: Critical Concepts for Social Sciences*, pp. 143–68. Routledge.
Lamont, M. (2009) *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.
Langfeldt, L., Nedeva, M., Sörlin, S., and Thomas, D. A. (2020) 'Co-Existing Notions of Research Quality: A Framework to Study Context-Specific Understandings of Good Research', *Minerva*, 58: 115–37.
Langfeldt, L., Reymert, I., and Aksnes, D. W. (2021) 'The Role of Metrics in Peer Assessments', *Research Evaluation*, 30: 112–26.
Lee, F., and Helgesson, C.-F. (2020) 'Styles of Valuation: Algorithms and Agency in High-Throughput Bioscience', *Science, Technology, and Human Values*, 45: 659–85.
Lyall, C. (2019) *Being an Interdisciplinary Academic: How Institutions Shape University Careers*. Cham: Palgrave Macmillan.
Lyall, C., Bruce, A., Tait, J., and Meagher, L. (2011) *Interdisciplinary Research Journeys: Practical Strategies for Capturing Creativity*. London: Bloomsbury Academic.
Nästesjö, J. (2021) 'Navigating Uncertainty: Early Career Academics and Practices of Appraisal Devices', *Minerva*, 59: 237–59.
Nästesjö, J. (2024) 'Uncertainty, Worth, Identity: How Early Career Academics Navigate Evaluative Landscapes', Diss, Lund University.
Nilsson, R. (2009) *God Vetenskap. Hur Forskares Vetenskapsuppfattningar Uttryckta i Sakkunnigutlåtanden Förändras i Tre Skilda Discipliner*. Göteborg: Acta Universitatis Gothoburgensis.
Pardo-Guerra, J. P. (2022) *The Quantified Scholar: How Research Evaluations Transformed the British Social Sciences*. Columbia University Press.
Pontika, N., Klebel, T., Correia, A., Metzler, H., Knoth, P., and Ross-Hellauer, T. (2022) 'Indicators of Research Quality, Quantity, Openness, and Responsibility in Institutional Review, Promotion, and Tenure Policies across Seven Countries', *Quantitative Science Studies*, 3: 888–911.
Reymert, I. (2021) 'Bibliometrics in Academic Recruitment: A Screening Tool Rather than a Game Changer', *Minerva*, 59: 53–78.
Reymert, I., Jungblut, J., and Borlaug, S. B. (2021) 'Are Evaluative Cultures National or Global? A Cross-National Study on Evaluative Cultures in Academic Recruitment Processes in Europe', *Higher Education*, 82: 823–43.
Rushforth, A., and de Rijcke, S. (2015) 'Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in The Netherlands', *Minerva*, 53: 117–39.
Sivertsen, G. (2016) 'Publication-Based Funding: The Norwegian Model', in M. Ochsner, S. E. Hug, and H. D. Daniel (eds) *Research Assessment in the Humanities: Towards Criteria and Procedures*, pp. 79–90. Springer.
Söderlind, J., and Geschwind, L. (2020) 'Disciplinary Differences in Academics' Perceptions of Performance Measurement at Nordic Universities', *Higher Education Governance and Policy*, 1: 18–31.
Thomas, D. A., Nedeva, M., Tirado, M. M., and Jacob, M. (2020) 'Changing Research on Research Evaluation: A Critical Literature Review to Revisit the Agenda', *Research Evaluation*, 29: 275–88.

Research Evaluation, 2024, 33, 1–12
https://doi.org/10.1093/reseval/rvae037
Special Issue Paper