

RESEARCH ARTICLE

JASIST WILEY

Conjoint analysis of researchers' hidden preferences for bibliometrics, altmetrics, and usage metrics

Steffen Lemke¹  | Athanasios Mazarakis^{1,2}  | Isabella Peters^{1,2} 

¹Web Science Department, ZBW – Leibniz Information Centre for Economics, Kiel, Germany

²Web Science Department, Kiel University, Kiel, Germany

Correspondence

Steffen Lemke, Düsternbrooker Weg 120, Kiel 24105, Germany.
Email: s.lemke@zbw.eu

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: 314727790

Abstract

The amount of annually published scholarly articles is growing steadily, as is the number of indicators through which impact of publications is measured. Little is known about how the increasing variety of available metrics affects researchers' processes of selecting literature to read. We conducted ranking experiments embedded into an online survey with 247 participating researchers, most from social sciences. Participants completed series of tasks in which they were asked to rank fictitious publications regarding their expected relevance, based on their scores regarding six prototypical metrics. Through applying logistic regression, cluster analysis, and manual coding of survey answers, we obtained detailed data on how prominent metrics for research impact influence our participants in decisions about which scientific articles to read. Survey answers revealed a combination of qualitative and quantitative characteristics that researchers consult when selecting literature, while regression analysis showed that among quantitative metrics, citation counts tend to be of highest concern, followed by Journal Impact Factors. Our results suggest a comparatively favorable view of many researchers on bibliometrics and widespread skepticism toward altmetrics. The findings underline the importance of equipping researchers with solid knowledge about specific metrics' limitations, as they seem to play significant roles in researchers' everyday relevance assessments.

1 | INTRODUCTION

In an age of exponentially growing publication rates (Bornmann & Mutz, 2015; Pautasso, 2012; Tian, Wen, & Hong, 2008) and with an abundance of publication-related usage data readily available (Moed, 2018), basing assessment of research outputs on quantitative metrics becomes ever more tempting. No matter whether on publication-, author-, journal-, or institution-level, calculating indicators based on citations is most often faster than engaging in thorough qualitative peer review. And with reliance on online platforms within researchers' workflows increasing

(Kramer & Bosman, 2016), indicators from more sources arise to complement citations—often summarized by umbrella terms like “altmetrics” (e.g., an article's mentions in news outlets, social media, policy documents, gray literature, academic syllabi, or clinical guidelines) or “usage metrics” (e.g., an article's download counts or online view counts). Linked to these alternative indicators is the hope that they might provide means to assess wider forms of research impact, especially for domains like social sciences and humanities, where the applicability of bibliometric assessments is limited due to idiosyncratic publication- and citation norms (Hicks, 2005; Sivertsen &

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Journal of the Association for Information Science and Technology published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

Larsen, 2012) and which lie outside the major citation databases' focus on STEM fields.

1.1 | Background and related work

A large body of work from numerous fields of research has studied the ramifications of using quantitative indicators in academic evaluation and assessment systems (see de Rijcke, Wouters, Rushforth, Franssen, and Hammarfelt (2016) for an overview). Frequently such studies conclude with warnings of inappropriate applications of the indicators, pointing to undesirable effects an extensive reliance on them could have on science. Common concerns for instance include that the use of indicators as auditing instruments could lead to researchers and institutions becoming more market-oriented, and as a consequence unduly focusing on research areas or publication types that have been shown to attain high metrics (Butler, 2005; Willmott, 2011). Such a climate could hinder certain kinds of research proposals, for example, particularly unusual or innovative and therefore risky projects (Butler, 2007). Similarly, a rising importance of indicators could lead to increasing publication pressure for researchers and put fields with lower overall metrics at a disadvantage for funding. Particularly fierce criticism against the use of citation counts for evaluative purposes is voiced by MacRoberts and MacRoberts (2018), who argue that the practice of citation analysis would be based on fundamentally false assumptions. They point to citation analysis' political implications by referencing Sosteric (1999, p. 19), who states that "citation analysis tends to support a particular one-sided, reified, and elitist view of scientific contributions that ignores [...] certain groups of scholars and by doing so justifies the highly stratified nature of the academy where certain groups are privileged over others." It should be noted that the much-debated aspect of evaluation is only one of research metrics' major use cases. Looking at eight types of typical metrics users, the *NISO Alternative Assessment Metrics Project* identified three overarching themes of use cases for metrics: *showcase achievements*, *research evaluation*, and *discovery* (National Information Standards Organization, 2016). Albeit the project's focus was placed specifically on altmetrics, all three themes apply to bibliometrics and usage metrics as well.

1.1.1 | Individuals' perceptions and usage of metrics

Several surveys investigated how indicators are used and perceived by individuals, most of them focusing on researchers. Hammarfelt and Haddow (2018) surveyed humanities scholars from Australia and Sweden about

their publication practice as well as about their knowledge and use of evaluation tools. They found a considerable share of Australian (62%) and a significantly lower share of Swedish (14%) respondents to actively use indicators. In a related study, Haddow and Hammarfelt (2019) found slightly larger but overall similar proportions of indicators users among Swedish and Australian social scientists. Both studies identified citations as the most frequently used indicator, common use cases being CVs, promotion, or grant applications, which in the NISO's use case scheme would fall into *Showcase achievements*. In their survey of Norwegian scientists, Aksnes and Rip (2009) found oftentimes "somewhat cynical" but ambivalent stances on the subject of citation counts. Ma and Ladisch (2016) conducted semi-structured interviews with four researchers about how they are affected by metrics usage. They identified three main themes of metrics usage: *self-monitoring*, *collaboration*, and *choice of journals and research topics*. In their follow-up study, Ma and Ladisch (2019) revealed a discrepancy between interviewed researchers' attitudes toward metrics in practice and in principle: while in principle, most researchers would not trust metrics as objective indicators for quality, in practice they do actively use them for personal or administrative purposes, thereby relying on them as supposedly objective measures. A similar discrepancy was found in a recent qualitative survey across faculty, instructors, and researchers at the University of Minnesota about their attitudes toward research metrics (Bakker, Cooper, Langham-Putrow, & McBurney, 2020). The participants stated several use cases metrics would fulfill for them, for example, in information-seeking activities, as a means of self-assessment, or in the assessment of other individuals. They did however also voice severe concerns linked to indicator use, for instance the feeling that metrics on their own could not serve as robust representations of the participants' own impact.

In an international study combining a long-term interview stage with a large-scale online survey, Nicholas et al. (2020a) specifically inquired about early career researchers' (ECRs) attitudes and practices toward citation-based metrics and altmetrics. Their analysis revealed citation-based indicators to fulfill several purposes for many ECRs, while altmetrics were regarded less favorably—criticism included that altmetrics might primarily reflect curiosity instead of impact, and that they would be vulnerable to being gamed. For the field of bibliometric research on the other hand, Haustein et al. (2014) attested altmetrics potential as a valuable source of impact data, as their analysis had shown large shares of bibliometric literature to be represented on online reference managers and substantial parts of their sample of bibliometricians to be affected by social media

tools in their professional lives. Also concentrating on altmetrics, Aung, Erdt, and Theng (2017) surveyed members of academia regarding their awareness and usage of their different types, differentiating between non-faculty staff and faculty staff among the participants. They found a tendency for non-faculty staff to be more aware of altmetrics than faculty staff. Moreover, mentions and shares on social networks were found to be the most used altmetrics, and usage of most altmetrics was shown to correlate with usage of social media. In the second phase of their study, Aung et al. (2019) investigated more broadly on scholars' familiarity with and usage of both traditional indicators and altmetrics. They found only few indicators to be widely known among scholars and the familiarity with and usage of altmetrics to be particularly low. As potential reasons for altmetrics' low popularity, Aung et al. (2019) mention academics' insecurities regarding altmetrics' added value, missing encouragement of altmetrics- and social media usage from institutions, and privacy concerns, among others. The reoccurring finding of a comparatively low familiarity with altmetrics is in line with multiple other surveys across academic librarians (Miles, Konkiel, & Sutton, 2018) and faculty (Bakker et al., 2019; DeSanto & Nichols, 2017).

Other research about individuals' use and perception of indicators placed its focus on stakeholders from academic administration. Abbott et al. (2010) combined a poll among researchers with interviews of academic administrators to investigate the former's perception and the latter's use of metrics. They find that researchers tend to overestimate the actual importance of metrics when it comes to how research administration measures their achievements, as most administrators reported not to use metrics to a large extent. Interestingly, when asked what researchers think should be the criteria for evaluating their careers, metrics were chosen very frequently. Regarding the use of indicators by administrations, McKiernan et al. (2019) arrived at different conclusions by analyzing 864 documents used in review, promotion, and tenure processes of North American research institutions. Their data suggests that research-intensive institutions in particular make heavy use of impact factors for evaluative purposes, noting that cautious statements about indicator usage were very rare in the analyzed documents.

In an own previous effort of inquiring about researchers' perceptions, knowledge, and use of various indicators, we conducted a series of group interviews and online surveys with social scientists (Lemke, Mehrazar, Mazarakis, & Peters, 2019). Our study's results revealed that even though many researchers are justifiably wary regarding metrics' reliability as indicators of scientific relevance or quality, they still are inclined to use some of

them—particularly citation-based ones—quite regularly. Frequently metrics serve as filters, for example, during literature research, when deciding which sources to cite, or when planning where to publish own works. Also, while most specific concerns about using metrics for research evaluation voiced by the respondents could apply to all types of metrics (i.e., bibliometrics, altmetrics, usage metrics) in equal measure, altmetrics performed considerably worse than bibliometrics considering perceived usefulness and trustworthiness.

1.1.2 | Previous research on researchers' literature selection processes

Researchers' information-seeking behavior and the criteria with which they decide on literature to read have also been examined in previous studies, although not necessarily with a focus on properties of the documents that are available, like for instance their metrics, but more often on the readers themselves (Tenopir, King, Spencer, & Wu, 2009). Through building regression models on survey data from 2,063 researchers from natural sciences, engineering, and medical sciences, Niu and Hemminger (2012) found numerous factors to affect scientists' information-seeking behavior, proposing a framework that includes demographic, psychological, role-related, and environmental factors. The most important determinants of information-seeking behavior found included the academic position, gender, and discipline. Looking at both demographic and contextual factors, Tenopir et al. (2009) used an online survey to analyze reading patterns of academic staff of universities from the US and Australia. They found subject disciplines and work responsibilities to be important characteristics determining reading behavior—for instance, while medical faculty tend to read more, specifically for current awareness, engineering faculty tend to spend more time per article and more often read for research than others. In a more recent large-scale survey, Tenopir et al. (2016) examined which activities the over 3,600 participating researchers would find most important to determine an article's trustworthiness, finding “checking if the arguments and logic presented in the content are sound,” “checking to see if the data used in the research are credible,” and “reading the abstract” being ranked most highly. Overall, they found participants to quite consistently value content properties more than respective articles' meta-information, for example, author or publisher names. On the other hand, regarding criteria for judging reading trustworthiness, the most highly rated statement was “Peer-reviewed journals are the most trustworthy information source,” indicating the type of publication

venue to be an important aspect in this regard. In another international survey aimed at early career researchers, Nicholas et al. (2020b) found a journal's prestige, rank, and impact factor as well as ease of access to be influential factors for participants when deciding what to read.

1.1.3 | Research problem

While several previous studies examined researchers' overall stance and practices toward different types of indicators for research assessment, little empirical research investigated indicators' role in the concrete practical decisions researchers take virtually every day when selecting literature to read. Furthermore, previous studies in this field are mostly restricted to surveys and interviews, which share some methodological weaknesses, for example, vulnerability to certain response biases. In this study, we approach the question of how researchers make use of indicators for evaluation purposes from a new angle by introducing conjoint analysis to the context of scientometrics. We designed an interactive online experiment in which invited researchers were asked to rank fictitious publications against each other regarding their expected scientific relevance—three publications at a time. Participants had to base their judgments on a set of preselected bibliometrics, altmetrics, and usage metrics, of which individual manifestations were presented for each single article. Through applying regression analysis we could later determine how different types of indicators overall affected the rankings made by participants of our experiment. The experiment was enclosed by a questionnaire that should help to put the results from our conjoint analysis into context.

With this study's results, we try to expand on insights from Lemke et al. (2019) on how indicators influence researchers' decisions during literature selection. By examining whether the researchers' ranking behavior in a fictitious practice situation complies with statements made in previous interviews and surveys, we aim to achieve a more truthful picture of researchers' perceptions of indicators than surveys and interviews alone could possibly provide us with. Moreover, we aim to detect and describe distinguishable types of indicator users by clustering respondents based on similarities in their ranking behavior. This way, the data obtained in our conjoint experiment should allow us to make detailed observations on the comparative values of individual indicators for certain users, which would be extremely difficult to obtain in such granularity in a regular survey. We primarily focus on researchers from social sciences, as our previous experiences with this target group should

help us to better put our findings into perspective. Also, due to the social sciences being subject to discipline-inherent limitations regarding the applicability of citation-based assessments (see for instance Hicks, 2005), it should be particularly interesting to gain further insights into how open researchers from this domain are toward using alternative indicators, which might circumvent some of citations' shortcomings (Wouters & Costas, 2012), in relevance assessments.

We chose the specific use case of utilizing indicators for reading prioritization during literature research as our experiment's core scenario for two reasons: first, we can assume most researchers to regularly be in the situation of encountering individual articles' metrics as possible filter criteria during online literature research and thus being familiar with it. This was also evidenced by responses during our previous interviews with social scientists (Lemke et al., 2019), which showed that even researchers in very early career stages are aware of article-level metrics and sometimes use them as filters in this particular scenario. Second, in much previous literature about researchers' perceptions of indicators, an emphasis often lies on use cases surrounding their own evaluation or their choice of publication channels (see for example Abbott et al., 2010; Haddow & Hammarfelt, 2019; Hammarfelt & Haddow, 2018; Ma & Ladisch, 2016). How researchers perceive and use metrics when evaluating others' work is in our eyes an equally interesting question that seems to have received less attention from the scientific community so far.

1.2 | Conjoint analysis

The term *conjoint analysis* encompasses a variety of decompositional multivariate techniques with the goal of estimating consumers' preference structures (Green & Srinivasan, 1978). In most implementations of conjoint analysis, a sample of the respective target population is asked to evaluate multiple differing alternatives of a product in question in an experimental setting (a so-called *choice- or ranking experiment*). The presented alternatives feature different manifestations of attributes, which are hypothesized to be relevant for the participants' decisions. After the collection of data, statistical techniques (e.g., regression models) are used to model participants' preferences and allow for conclusions about individual attributes' effects on participants' overall choices. One main benefit of conjoint analysis is that it comparatively truthfully emulates participants' real decision-making situations, as participants evaluate a respective product's attributes implicitly while evaluating whole products—as they would most likely have to in the real world.

Originally stemming from the fields of mathematical psychology and psychometrics (Green & Wind, 1975), conjoint analysis has been used prevalently in marketing where it is typically utilized to estimate consumer behavior. However, conjoint analysis can in principle be applied to innumerable scenarios from other domains in which human subjects choose between multiple alternatives with the goal of maximizing fulfillment of their personal preferences. For an example for an application of conjoint analysis in the field of Computer Science see Kirchhoff, Capurro, and Turner (2014), who used it to assess users' "preferences" for different types of errors made by machine translation engines. Tenopir et al. (2011) even utilized conjoint analysis in context of a research question from the sphere of scholarly communication similar to ours: controlling for seven different characteristics of research articles, they found article topic, online accessibility, and peer review status to be the most important factors for researchers when deciding which articles to read.

To the best of our knowledge, conjoint analysis has not been used for assessing preferences of users of research indicators yet. In this study, we combine methods of conjoint analysis with an online survey to investigate factors that influence researchers' decisions about which scientific publications to consume with priority. In particular, we inquire about the role quantitative indicators play in this—how do different types of indicators compare regarding their perceived utility as selection-criteria during literature research? Also, the majority of previous studies on researchers' perceptions and use of indicators focused on bibliometrics. We address this gap by, in addition to bibliometric indicators, also incorporating a selection of altmetrics and usage metrics in our study.

2 | METHODS

The following sections provide information on the planning, implementation, and dissemination of our ranking experiment as well as on the methods used for analysis of the collected data.

2.1 | Definition of attributes and levels

A key decision when planning a conjoint analysis concerns the attributes that should be incorporated in the experiment—meaning the features of the products in question that participants are meant to base their preference judgments upon. As we want to examine how different quantitative indicators influence researchers' choices

when deciding which literature to read first, the *products* in our experiment will be fictitious scientific articles that showed up as results of our participants' hypothetical literature search, while their *attributes* will be a selection of indicators, for example, their individual citation counts, or numbers of mentions on Twitter. Different articles sport different *levels* of those attributes—one article could for example have five citations and five mentions on Twitter, while another article has zero citations but 250 mentions on Twitter.

The decision about how many attributes and how many levels per attribute to include in a conjoint analysis is closely connected to the chosen method for data collection. Two fundamentally different approaches for data collection exist (Green & Srinivasan, 1978; McCullough, 2002): partial-profile and full-profile approaches. In full-profile designs, participants are asked to rank products for which individual data on all of their attributes is visible; in partial-profile approaches, participants can only see a fraction of the attributes for all products during each ranking task. While partial-profile approaches make sense in scenarios with extremely high numbers of relevant attributes (such designs can include up to 50 or more attributes; McCullough (2002)), we decided for a full-profile design, as it provides more realistic and comprehensible tasks if the number of included attributes is fairly low.

To prevent our participants from information overload, we decided to follow the common recommendation for full-profile designs to include a maximum of six attributes (Green & Srinivasan, 1978; McCullough, 2002). An initial list of potential indicators to include as attributes was created through a combination of literature review, inspection of the data sources covered by prominent altmetrics providers, and brainstorming. Due to their extremely diverse sources, especially the different types of altmetrics quickly led this initial list grow to more than 20 entries. When we felt a saturation regarding the incorporation of further relevant indicators to be reached, the list was then iteratively reduced to six indicators we deemed to have particularly high presence and relevance in scientometric literature and practice. We balanced this criterion with the aims of incorporating at least one prototypical indicator from each of several different areas of metrics and picking indicators that as many participants as possible should already feel familiar with:

- the article's *citations* (e.g., on *Google Scholar*) as an article-level bibliometric indicator;
- the publishing journal's *Journal Impact Factor* as a prominent and much-debated journal-level indicator;
- the first author's *h-index* as a widely known author-level indicator;

	Level 1	Level 2	Level 3
Citations (e.g., on Google Scholar)	0	5	250
Journal impact factor	0	5.0	30.0
h-index	0	5	30
Downloads	0	100	5,000
Tweets	0	10	500
Mendeley readers	0	10	500

TABLE 1 Attributes and their levels in the experiment



FIGURE 1 Random example publication based on our choice of attributes and attribute levels [Color figure can be viewed at wileyonlinelibrary.com]

- the article's number of *downloads* as an article-level usage indicator;
- the article's number of mentions in *tweets* as an altmetric drawn from a prominent social media platform targeted at a general audience;
- the article's number of *readers on Mendeley* as a comparatively well-examined altmetric drawn from a social media platform targeted at scholars.

Ideally, for each attribute the same number of levels should exist in the experiment to prevent attributes' numbers of levels from having an effect on individual attributes' estimated importance (McCullough, 2002). Moreover, different levels of the same attribute should be easily perceivable as distinct; the ranges covered by them may be slightly larger than in reality, but not so large as to be unbelievable (Green & Srinivasan, 1978). We decided to include three levels per attribute: level 1 should intuitively translate to *no occurrences of this indicator*, level 2 to *few occurrences of this indicator* and level 3 to a *high number of occurrences of this indicator*. Starting from this, the authors discussed and agreed upon values they deemed to be plausible for the examined domain while being in accordance with the advice from Green and Srinivasan (1978). Table 1 shows the resulting three levels for our six attributes.

An example for how a publication profile based on this selection of attributes and levels finally looked like in our experiment is shown in Figure 1. It should be noted that in the actual experiment the order in which attributes were listed was randomized between participants to reduce a possible influence of their order of

appearance on individual attributes' measured effect. The image of an article's front page on the left was depicted for illustration only and did not provide any bibliographic information, as participants should base their judgments solely on our six pre-selected attributes.

2.2 | Design of tasks

After the included attributes and their levels were defined, decisions had still to be made about how many alternative publications a participant should have to compare during a single task and about how many tasks each participant should be asked to complete during their run of the experiment. Based on pretests with a first prototype of the software used for our experiment (see below), we decided to let participants assess three fictitious articles at a time—this should keep individual tasks short and comprehensible. We planned for every participant to complete a total of 20 tasks, an amount that should be solvable in approximately 15 minutes without degradation of data quality due to participants' fatigue (McCullough, 2002). Following advice by McCullough (2002), we decided to regard the first two tasks completed by every participant as warm-up tasks which would not be considered during analysis. This should account for the fact that it can take a little while for participants' behavior to stabilize, as they might only get a feeling for the experiment's concept and scale after having completed one or two tasks. For the analysis, this would leave us with 18 tasks to evaluate for every participant who completed the full experiment.

To come up with a definite set of 20 tasks to be used in the experiment, we largely followed the guideline for the design of choice experiments using R by Aizaki and Nishimura (2008). This included the creation of a full factorial design based on our predefined attributes and levels and the subsequent generation of a fractional factorial design matching the number of tasks in our experiment. The approach makes use of Federov's exchange algorithm (as implemented in Bob Wheeler's R package *AlgDesign*¹) that, given our restrictions regarding number of attributes, levels, and tasks per participant, creates a

combination of tasks to include in our design to make model estimation as efficient as possible.

Another fundamental question during the planning of conjoint analysis experiments concerns the way in which participants are asked to express their judgments. Two widely used paradigms exist (Louviere, Flynn, & Carson, 2010): the “traditional,” rankings-based conjoint analysis and “discrete choice experiments” (DCEs), in which participants have to choose exactly one option out of the given alternatives in each task. In our experiment we want to emulate the situation of researchers prioritizing between different articles found during literature research—a scenario in which the respective researcher will usually intend to read several of the articles a search has led her to, not only the single most appealing one. We therefore believe a ranking of articles to more realistically represent the kind of decisions in question than a discrete choice. Moreover, literature has shown that rank order data can be expanded into sets of implied discrete choices (Chapman & Staelin, 1982; Louviere et al., 2010; Vermeulen, Goos, & Vandebroek, 2011). For us this means that relying on a ranking-based method of data collection increases the amount of information obtained per task without leading to a significant loss of flexibility regarding data analysis.

2.3 | Implementation

We implemented our experiment as a web application, so that invited participants could access it via a web browser. Our software is open-source and can be obtained from GitHub.²

Figure 2 schematically summarizes the experiment's course a participant would follow after clicking on an invitation link. To collect demographical data as well as free text responses on our participants' thoughts regarding their literature research practices, the use of metrics, and our experiment in general, the experiment both began and ended with questionnaire segments (pages B,

C, G, and H) surrounding the actual 20 tasks a participant was asked to complete (pages F). Pages D and E gave detailed information on the indicators referred to in the experiment and explained how to navigate through the tasks. A file with full screenshots of all pages of the experiment is available in this article's supplementary material.

Figure 3 shows an example for a task in our experiment. Each task started with the same explanatory text seen in Figure 3. Below this text, a set of three fictitious publication profiles from our fractional factorial design was presented as boxes. All boxes could be moved and rearranged per drag-and-drop to achieve a desired order. As soon as all three publications had been allocated to the three slots on the right, a participant could move on to the next task by clicking the continue-button. Just like the order in which attributes were listed on publication profiles was randomized between participants, the order of publication profiles in each task was randomized as well to rule out giving an advantage to certain publications by always listing them first.

The software was tested iteratively with multiple rounds of feedback coming from about a dozen of colleagues from both within and without of our research team, to ensure it being as self-explaining and technically sound as needed. To also make sure that our way of processing input data would later allow us to perform all planned steps of analysis (see below) as intended, we simulated later steps once by automatically generating hypothetical input data for 30 fictitious participants. After successful simulation of data analysis, this software-generated input data was discarded.

2.4 | Dissemination

For the experiment's dissemination, we relied on a subsample of the respondents from a survey among researchers we had conducted in the summer of 2018 (Lemke et al., 2019). Said survey had been sent to authors of social science-related papers found on RePEc and Web of Science, as well as to a mailing list maintained by the ZBW—Leibniz Information Centre for Economics. The latter consisted of about 12,000 email addresses of researchers, with a strong focus on economists from German-speaking parts of Europe. Of the survey's 2,083 respondents, 938 had agreed to be contacted again for further user studies from our project and therefore received invitations to this experiment half a year later. The dissemination of invitation links, which would allow participants to access the experiment's website, was done via email between November 29 and December 10, 2018. Data collection was carried out till January 15, 2019. As

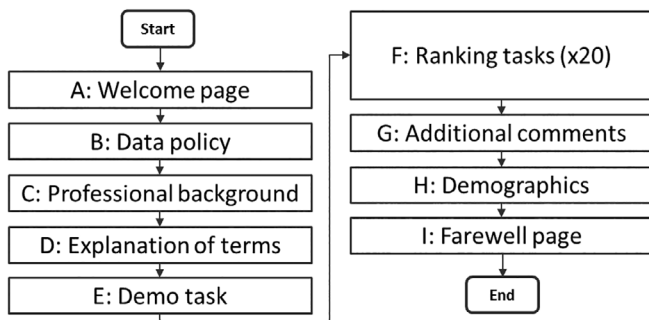









FIGURE 2 Schematic representation of experiment

During the experiment, please imagine the following situation:

You are doing literature research for a topic you are not yet familiar with.
Your query in the scholarly search engine of your choice reveals 3 potentially relevant publications alongside their impact metrics.
Please rank those publications in the order in which you would read them by dragging them to the area on the right.
 The publication you would read first should afterwards be at the top of the list, the publication you would read last at the bottom.

Publication A		Citations (e.g., on GS)	5
		Tweets	500
		Mendeley readers	500
		Downloads	5000
		Journal Impact Factor	0
		h-Index	5

Publication B		Citations (e.g., on GS)	250
		Tweets	10
		Mendeley readers	0
		Downloads	5000
		Journal Impact Factor	5.0
		h-Index	30

Publication C		Citations (e.g., on GS)	5
		Tweets	10
		Mendeley readers	0
		Downloads	100
		Journal Impact Factor	0
		h-Index	30

- Read first -

...

- Read last -

When you are satisfied with your decision, please click 'Continue' to go on with the actual experiment.

Back

Continue

FIGURE 3 Example ranking task from the experiment [Color figure can be viewed at wileyonlinelibrary.com]

an incentive, participants could optionally partake in a random drawing of fifteen 30€-vouchers for Amazon after completing the experiment.

2.5 | Data analysis

Rank orders of publications entered by participants during the experiment were regarded as sequences of choices (Vermeulen et al., 2011). So if in a task participant *P* had ranked the three publications *A*, *B*, and *C* in the order $C > B > A$, for the purposes of our analysis this input would be treated as two parts of information: “between *A*, *B*, and *C*, participant *P* chose *C*” and “between *A* and *B*, participant *P* chose *B*.” The decision data from all participants (barring each participant’s first two tasks, which were regarded as warm-up tasks as described above) was fed into a logit model using *R*, with the six indicators included in the experiment as independent variables and the binary outcome of a publication being ranked above its competitors as dependent variable.

In a subsequent step, we performed a cluster analysis to identify groups of participants with similar ranking

behaviors. To be able to do so, we first for every participant transformed all of their ranking choices to numerical vectors. On these vectors, we then used *k*-means clustering. For every one of the participant groups identified through our cluster analysis we then again computed individual regression models, like we had done before for all participants combined.

Through survey segments right before and after the experiment’s ranking tasks, we collected participants’ demographics as well as further information on their usual strategies during literature research and on their notions about research indicators. Both from previous literature (Nicholas et al., 2020b; Niu & Hemminger, 2012; Tenopir et al., 2009; Tenopir et al., 2011; Tenopir et al., 2016) as well as from personal experience we had reason to believe that in real literature research scenarios the respective researchers would often determine their orders of preference based on more complex heuristics, involving more criteria than the six indicators we could ask for in our experiment. As for instance Tenopir et al. (2011) have shown, there are in particular several qualitative aspects researchers look out for when deciding what to read. To not disregard such aspects,

particular attention during the analysis went into the responses to two free text questions:

- Right **before** the ranking tasks, we asked every participant: “When doing literature research, how do you usually determine which search results to read first? Are there publication features you are looking out for?”
- Right **after** the ranking tasks, we again showed every participant what he or she had responded to the first free text question and asked: “Now after having finished the experiment, would you like to add anything to your previous answer?”

The responses to both free text questions were coded manually by one author (S. L.) and grouped by topics they referred to.

3 | RESULTS

In this section, we present our participants' reported demographics, regression models of the choices they made during our experiment, results from the cluster analysis based on said choices, and an analysis of the free text answers our participants gave.

3.1 | Demographics

In total, 247 of the 938 researchers we had invited participated in our experiment by completing at least its first survey page, meaning a response rate of 26%; 204 participants finished the experiment completely. Figure 4 shows the number of participants that completed a respective page transition.

Table 2 shows the participants' stated demographics concerning their discipline, professional role, country of affiliation, and gender.

The participants' median year of birth is 1980, with a standard deviation of 10.22 years and a range from 1946

to 1993 (a single response of “1900” to the question for year of birth was discarded as implausible). Regarding reported years of academic experience, the median is at 11 years, with a standard deviation of 8.75 years and a range from 2 to 43 years.

3.2 | First regression model

In total, 4,222 comparison tasks were completed by our respondents, up to 20 by each one of them. As we discard the first two tasks solved by every respondent as warm-up tasks, 3,774 evaluable tasks remain. Because every task completion actually consists of two separate choices as described earlier, the amount of evaluable choices to analyze is 7,548.

Table 3 shows the parameters of the logit model created on basis of those 7,548 choices. All estimated coefficients are standardized to respective attributes' standard deviations to facilitate comparisons despite the attributes' differing absolute ranges.

We see that for every one of the six indicators an increase also leads to an increase in the respective article's likelihood of being ranked higher than its competitors. The strongest effect per standard deviation have *citations*—increasing an article's citations by one standard deviation increases its log odds of being preferred to competitors by 0.607. The second-highest effect has the *Journal Impact Factor* (0.468), followed after a considerable gap by *downloads* (0.247). The remaining three indicators all perform similar to each other with coefficients close to 0.160.

3.3 | Most helpful indicator

Right after completing the 20 ranking tasks, participants were asked which one of the six indicators they would find “most helpful as a tool for deciding which publications to read.” Figure 5 shows which shares of participants chose which indicator.

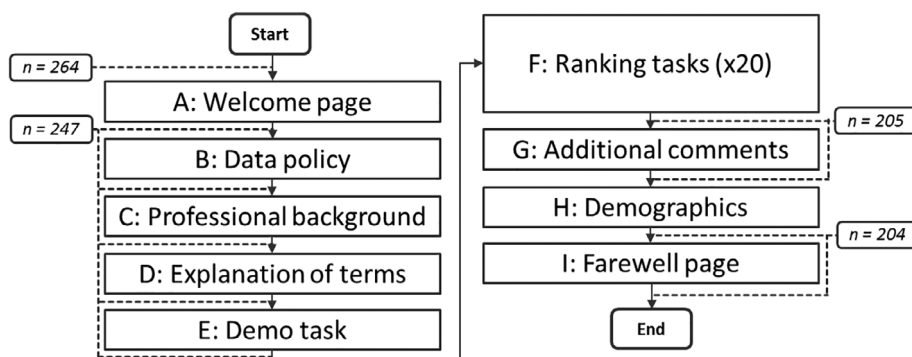


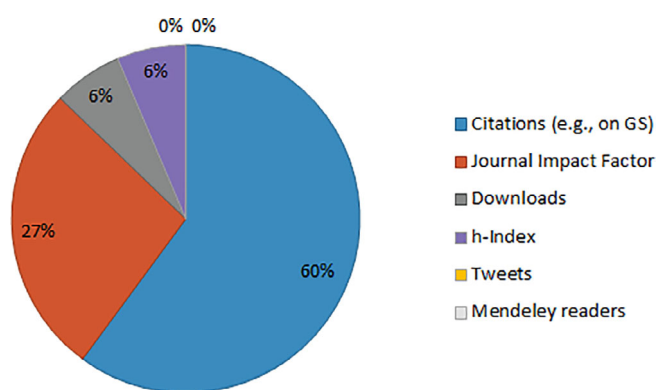
FIGURE 4 Schematic representation of experiment with numbers of completing participants

TABLE 2 Participants' demographics

Discipline	Percentage	Professional role	Percentage
Economics	61%	PostDoc/senior researcher	22%
Social sciences	21%	Professor	18%
Other	9%	Assistant professor	16%
Engineering/technology	3%	Associate professor	16%
Life sciences	3%	PhD student	9%
Arts/humanities	2%	Research assistant + PhD student	9%
Law	1%	Other	7%
Medicine	<1%	Research assistant	2%
		Student/student assistant	1%
Country of affiliation	Percentage	Gender	Percentage
Germany	34%	Male	68%
United States	11%	Female	29%
Italy	8%	I prefer not to answer	3%
UK	6%		
Spain	3%		
France	3%		
Poland	3%		
Other (number of other countries = 31)	32%		

Variable	Estimate	SE	p
(intercept)	−0.419	0.016	<.001
Citations (e.g., on Google Scholar)	0.607	0.017	<.001
Journal impact factor	0.468	0.016	<.001
h-index	0.160	0.017	<.001
Downloads	0.247	0.016	<.001
Tweets	0.159	0.016	<.001
Mendeley readers	0.157	0.017	<.001

TABLE 3 Coefficients for logit model of all participants' ranking data

FIGURE 5 Indicators from the experiment selected as most useful for deciding which publications to read ($n = 203$) [Color figure can be viewed at wileyonlinelibrary.com]

The rank order of indicators' perceived usefulness achieved this way largely confirms the results from the regression, with most participants assessing citations as being most useful, followed by the Journal Impact factor. Equal numbers of participants chose either downloads or h-index as the most useful indicator. Remarkably, not a single participant chose Mendeley readers or Tweets as preferred indicator, despite our earlier observation that both of these indicators have similarly strong effects on an article's likelihood of being preferred over its competitors as the h-index. A possible explanation for this could be that the two social media-based indicators are only ever used when consulting the preferred indicators does not lead to a clear decision—in these cases the altmetrics could serve as “tiebreakers.”

TABLE 4 Coefficients for cluster-wise logit models

Variable	Estimates			
	C1	C2	C3	C4
<i>n</i>	65	67	20	53
Citations	0.853***	0.302***	0.332***	0.935***
Journal impact factor	0.486***	0.808***	0.146**	0.248***
h-index	0.227***	0.167***	0.086	0.171***
Downloads	0.052	0.188***	0.603***	0.455***
Tweets	0.097**	0.171***	0.148**	0.286***
Mendeley readers	0.083**	0.274***	0.104*	0.167***
Cluster description	“Bibliometrics-believers”	“IF-fixated”	“Usage evidence enthusiasts”	“Open-minded citation users”

*** $p < .001$; ** $p < .01$; * $p < .05$.

3.4 | Clusters of participants and their regression models

To be able to see whether there are groups of participants that can be distinguished by their ranking strategies, we performed a cluster analysis based on our dataset of ranking choices. Data from participants who had not completed all 20 tasks of the experiment was discarded for the cluster analysis, leaving us with 205 respondents to include. We started our *k*-means clustering with a *k*-value of 2, which we gradually increased until the addition of further clusters would lead to some very small groups including less than 20 respondents per cluster. We this way found $k = 4$ to be a satisfying configuration for getting distinct clusters of significant sizes. The resulting clusters C1 to C4 consisted of 65, 67, 20, and 53 respondents respectively. Detailed tables of each cluster's demographics can be found in this article's supplementary material. The descriptive statistics indicate that the demographic differences between clusters are overall low.

Table 4 summarizes the coefficients of the logit models we estimated for the four individual clusters, just as we had done before for all participants combined.

In the following we briefly characterize the four clusters' ranking behaviors:

Cluster 1: “bibliometrics-believers”—this fairly large group of participants trusts in citations and their derivatives, but seems very skeptical of all offered usage- and altmetrics.

Cluster 2: “Impact Factor-fixated”—this equally large group assigns very high value to the Journal Impact Factor. Apparently, when having to rank based on article- or author-level metrics, no indicator is rejected completely; also noteworthy is this group's comparatively high acceptance of Mendeley counts.

Cluster 3: “usage evidence enthusiasts”—the smallest of all clusters, these participants seem to trust in usage

metrics (downloads) and traditional citations as indicators for relevance. Also, they are the group with the lowest interest in non-article-level indicators.

Cluster 4: “open-minded citation users”—like the participants from cluster 3, these cluster's members also have a stronger focus on article-level indicators (and particularly citations) than most other participants, although they do not seem to reject any one indicator completely.

3.5 | Free text responses

A substantial number of participants also provided answers to two free text questions we asked before and after the ranking task-part of the experiment. Not counting non-topical answers (such as “no comment”), 205 participants responded to the question “*When doing literature research, how do you usually determine which results to read first? [...]*”, which participants had been asked right before the start of the ranking tasks. Another 132 topical responses were given to the question “*Now after having finished the experiment, would you like to add anything to your previous answer?*”, which was shown immediately after a participant had finished their 20 ranking tasks, alongside the response that person had given to the first question if applicable. All responses were coded regarding individual selection criteria referred to in them. Table 5 shows these criteria, together with the numbers of responses they have been mentioned in before and after the ranking tasks.

The free text responses confirm that there is a variety of qualitative features researchers tend to look at when deciding about an article's relevance for themselves, for example, its title, abstract, authors, topics, and date of publication. But also quantitative indicators seem to play an important part: citation counts were mentioned as used by many participants both before and after the

TABLE 5 Selection criteria used during literature research as mentioned in free text responses

Criteria	Mentioned before	Mentioned after
Journal (prestige/ranking/impact factor)	86	39
Title	50	1
Citation counts	47	61
Abstract	47	1
Authors	42	9
Date of publication/recency	42	10
Topical proximity	40	1
Keywords	25	-
Other	20	8
Reference-relations (e.g., "cited by"-criteria)	16	3
Publisher/online source	11	-
Order of appearance in search engine	11	-
Availability/access	11	-
Content properties (e.g., article length, methodology)	10	1
Publication type	6	-
Downloads	-	32
h-index	-	24
Mendeley readership counts	-	2
Tweets	-	4

experiment, download counts and an author's h-index were confirmed to be of interest by several researchers after they had finished the ranking tasks. The most frequently mentioned selection criteria of all was the journal an article is published in, although it is often not possible to tell from the responses whether respective respondents base their judgments about a journal's appeal on quantitative indicators like the Journal Impact Factor or on qualitative criteria.

Apart from mentioning selection criteria for articles, some respondents used their free text answer after the experiment to express criticism of specific indicators. Table 6 shows the numbers of occurrences of such comments. The comparatively high frequency of arguments against altmetrics is in line with our previous observations of participants' reluctant use of them, although citations are the only indicator that did not provoke any specific criticism at all.

As a rough check of whether respondents' choices during ranking tasks had been in accordance with the criteria they had explicitly reported to look out for, we

TABLE 6 Arguments against indicators stated in free text responses after the experiment

Comment	Mentioned after
Argument against tweets	11
Argument against downloads	5
Argument against Mendeley readership counts	3
Argument against h-index	2
Argument against journal impact factor	1

also examined free text responses for each of the four respondent clusters (Table 4) individually. Comparing the percentages of respondents from clusters that explicitly mentioned certain metrics as relevant selection criteria with the respective clusters' regression coefficients as reported in Table 4 revealed an overall high degree of congruence between ranking behavior and free text responses. More detailed results of this consistency analysis can be found in the supplementary material.

4 | DISCUSSION

We conducted an interactive experiment to investigate researchers' usage and preferences regarding quantitative indicators when assessing literature's relevance on a micro-level. The regression models based on participants' ranking choices as well as the survey answers revealed clear preferences for bibliometric indicators, first and foremost citation counts, followed by the Journal Impact Factor and usage metrics in form of download counts. While the author-based h-index and the two altmetrics included in the experiment exhibited similar effect sizes in our main regression analysis, both selection- and free text-based survey responses suggested a particularly widespread wariness toward the use of altmetrics. Our clustering of participants based on their ranking data indicated that several groups of indicators users that follow different strategies regarding their use of indicators for relevance assessment can be distinguished. The analysis of free text responses suggested that in practice researchers inspect both quantitative and qualitative properties of research articles to decide which publications to read, in line with previous studies on researchers' reading decisions (Nicholas et al., 2020b; Tenopir et al., 2011).

Comparing our results to the conjoint analysis of article characteristics that researchers value by Tenopir et al. (2011), we can see that every qualitative characteristic they included in their model also in some form came up in the free text responses to our study, confirming

these characteristics' general relevance. If we compare the rankings Tenopir et al. (2011) obtained through their conjoint analysis with the frequencies with which individual characteristics came up in our free text responses, we can see differences: while the respondents of Tenopir et al. (2011) put particular emphases on articles' topics and matters of online accessibility, for our respondents an article's publication venue seems to be of importance. An explanation for our respondents' lower focus on aspects of availability could be the high share of economists in our sample, who due to their prevalent reliance on openly accessible working papers might less regularly experience difficulties regarding article availability.

The free text responses we obtained evince that quantitative metrics for research assessment do play a role in researchers' everyday decisions. Moreover, our study shows that many researchers are considerably more open to the use of bibliometrics and in some cases usage metrics as indicators of scientific relevance than to the use of altmetrics. These findings are in line with observations made in previous surveys and interviews that revealed a critical stance many researchers have regarding both altmetrics as relevance indicators and social media platforms as channels for scholarly communication (Aung et al., 2019; Lemke et al., 2019; Nicholas et al., 2020a). The rank order *bibliometrics* > *usage metrics* > *altmetrics* also coincides with findings by Miles et al. (2018) about academic librarians' familiarity with different types of research impact indicators. It stands to reason that also for researchers their reluctance to use web-based indicators can at least in part be explained by their lesser familiarity with them compared to citation-based metrics (Aung et al., 2019; Bakker et al., 2019; DeSanto & Nichols, 2017; Lemke et al., 2019). Also, citation-based indicators are the type of indicator that (still) counts the most in the academic reward system.

Clearly, a certain amount of caution against basing judgments about individual articles' scientific relevance on any quantitative indicator is advisable—in recent years several high-profile statements have been publicized by experts warning of purely quantitative micro-level assessments as an exemplary form of indicators' misuse (Cagan, 2013; Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Wilsdon et al., 2015). It might therefore be considered unfortunate that the free text responses our participants gave about how they usually determine an article's relevance strongly indicate that their widespread skepticism against altmetrics does not translate to citations and their derivatives. Also, as Ma and Ladisch (2019) have shown, does even a lack of trust in an indicator's objectivity not prevent researchers from using it for certain assessments. This underlines the importance of establishing a level of "metrics literacy" among researchers

regardless of discipline that allows them to gauge various impact metrics' respective scopes, strengths, and limitations and avoid misinterpretations (Ma & Ladisch, 2019; see also Rousseau and Rousseau (2017)). For advocates of altmetrics, who aim to broaden the arsenal of tools used for impact measurement in an effort of mitigating the power given to one single indicator, our results show that there is still a lot of work ahead. Additional efforts of informing researchers about and familiarizing them with alternative indicators will be necessary to enable them to make use of the various web-based complements to bibliometrics that exist today.

5 | LIMITATIONS OF THE STUDY

A limitation of our study lies in its sample, which had a strong focus on social sciences and in particular economics. We assume that many researchers' acceptance of certain metrics as relevance indicators will be affected considerably by their discipline, given the well-documented substantial differences regarding certain metrics' applicability to different fields (Hicks et al., 2015; Thelwall, 2018). Results from this study can therefore not be generalized to other disciplines. Also, potential self-selection bias and the experimental setting might have led to an overemphasis on comparatively tech-savvy participants as well as on those with high interest in the topics of research assessment and/or impact indicators.

Moreover, although we aimed to base our choices of attributes and levels on established guidelines for conducting conjoint experiments where it seemed feasible, these choices remain arbitrary to a degree. One implication of this is that we might have missed indicators of particular value for our target group. Especially for fields with a high reliance on non-journal article publication formats, as is the case in many social sciences and humanities, the future inspection of perceived values of altmetrics based on formats like gray literature, books, or syllabi might be insightful. Another limitation concerns the choice of levels, as despite our measures taken to confront our participants with a balanced experimental design, we cannot rule out that the attributes' individual level ranges affected our outcomes. For instance, an attribute manifestation perceived as disproportionately valuable could lead participants to ignore other attributes when exposed to it. The perhaps most risky example in our experiment was the highest Journal Impact Factor level of 30—while most researchers should be able to envision examples for respective articles by thinking about multidisciplinary mega-journals, for monodisciplinary journals of many fields this value would constitute extreme outliers. Our use of a fractional factorial

design should however reduce the overall potential for one outstanding attribute manifestation to distort the results. Another approach would have been to base attribute levels on real articles, although this might come at the cost of making differences between articles harder to distinguish for the participants (see also the advice by Green and Srinivasan (1978) on defining levels in conjoint experiments).

Another aspect that is not accounted for in an experiment like ours is the comparative difficulty with which certain information is obtainable in reality. For instance, on a platform like *Google Scholar* citation counts and even a first author's h-index might be accessible with comparative ease, while the less common altmetrics might be considerably harder to obtain. So while the experiment's assumption, all indicators would be readily available during search, has for instance led us to find that h-index and tweet mentions are overall valued to similar degrees, in reality the h-index might still exert a stronger influence on reading decisions due to its easier availability.

Finally, in our experiment's survey part we did not explicitly state that participants can assume topical relevance for all hypothetical literature finds to be given. While we do not expect this fact to have meaningful implications for the findings of regression models or cluster analysis, it might have influenced the free text responses, as they included several properties that inform about topical relevance (e.g., title, abstract, topical proximity). For future applications of this study's methodology we would recommend to clarify this aspect in the questionnaire.

6 | CONCLUSIONS AND FUTURE WORK

Our study's findings indicate that quantitative indicators are a part of many researchers' practices when initially assessing literature for its relevance, albeit next to a variety of qualitative aspects like for instance topical relevance or accessibility. We've seen distinct groups of users to value different indicators to different degrees, although overall, traditional citation-based indicators overshadowed altmetrics and usage metrics in this regard. Our results inform various stakeholders interested in providing their users with helpful information for deciding on literature, for example, providers of literature search engines, publication databases, or scholarly publishers.

As noted above, additional work should go into the analysis of disciplinary differences regarding acceptance and use of different indicators. Also, a bottleneck in an approach like ours is imposed by the limited number of product characteristics that can be observed at a time.

Although we hope to have covered the most relevant six, there obviously are more quantitative indicators that would be interesting to analyze regarding their influence on researchers' consumption behavior. And even for qualitative characteristics, like those collected in the survey part of our study, utilities for potential readers could be estimated, as Tenopir et al. (2011) demonstrated.

Furthermore, the NISO use cases (National Information Standards Organization, 2016) offer several starting points for expansions to this study. It would for instance be interesting to see whether researchers' acceptance and use of different types of indicators changes when the task at hand is not about evaluation of search results, but about selecting metrics to showcase their own achievements.

Our study introduced new methods to the field of bibliometrics and provides empirical evidence on how indicators guide ranking processes of readers. A conjoint analysis as performed here is an elaborate endeavor that initially takes a lot of preparation. We hope to considerably facilitate various steps of conceptualization and data collection for future studies by making our software, which should be easily adaptable to a multitude of settings and research questions, openly available. In our own continuation of the study presented here, we next will use this approach and software to pursue the question of how certain prevalent methods of visualizing metrics data affect users' perception of research products' relevance.

ACKNOWLEDGMENTS

We thank our colleagues in the *metrics-project and at ZBW for their helpful feedback on early versions of the software used in the experiments, as well as the DFG for funding our project (grant number 314727790). Also, we wish to thank all researchers who helped us by participating in our online experiment. Open Access funding enabled and organized by ProjektDEAL. WOA Institution: DEUTSCHE ZENTRALBIBLIOTHEK FÜR WIRTSCHAFTSWISSENSCHAFTEN LEIBNIZ-INFORMATION SZENTRUM WIRTSCHAFT. Blended DEAL : ProjektDEAL.

DATA AVAILABILITY STATEMENT

The data collected during experiments and its accompanying survey is openly available on <https://zenodo.org/record/3560886>. The source code of the software used in this experiment can be found on <https://github.com/stlemke/metrics-conjoint/>.

ORCID

Steffen Lemke  <https://orcid.org/0000-0002-3506-7083>

Athanasios Mazarakis  <https://orcid.org/0000-0001-9943-0382>

Isabella Peters  <https://orcid.org/0000-0001-5840-0806>

ENDNOTES

¹ <https://CRAN.R-project.org/package=AlgDesign>

² <https://github.com/stlemke/metrics-conjoint/>

REFERENCES

- Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: Do metrics matter? *Nature News*, 465(7300), 860–862. <https://doi.org/10.1038/465860a>
- Aizaki, H., & Nishimura, K. (2008). Design and analysis of choice experiments using R: A brief introduction. *Agricultural Information Research*, 17(2), 86–94. <https://doi.org/10.3173/air.17.86>
- Aksnes, D. W., & Rip, A. (2009). Researchers' perceptions of citations. *Research Policy*, 38(6), 895–905. <https://doi.org/10.1016/j.respol.2009.02.001>
- Aung, H. H., Erdt, M., & Theng, Y.-L. (2017). Awareness and usage of altmetrics: A user survey. *Proceedings of the Association for Information Science and Technology*, 54(1), 18–26. <https://doi.org/10.1002/prai.2017.14505401003>
- Aung, H. H., Zheng, H., Erdt, M., Aw, A. S., Sin, S.-C. J., & Theng, Y.-L. (2019). Investigating familiarity and usage of traditional metrics and altmetrics. *Journal of the Association for Information Science and Technology*, 70(8), 872–887. <https://doi.org/10.1002/asi.24162>
- Bakker, C., Bull, J., Courtney, N., DeSanto, D., Langham-Putrow, A., McBurney, J., & Nichols, A. (2019). *How faculty demonstrate impact: A multi-institutional study of faculty understandings, perceptions, and strategies regarding impact metrics*. Association of College and Research Libraries (ACRL) Conference. Retrieved from https://scholar.valpo.edu/ccsls_fac_presentations/20
- Bakker, C., Cooper, K., Langham-Putrow, A., & McBurney, J. (2020). Qualitative analysis of faculty opinions on and perceptions of research impact metrics. *College & Research Libraries*, 81(6), 896–912. <https://doi.org/10.5860/crl.81.6.896>
- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. <https://doi.org/10.1002/asi.23329>
- Butler, L. (2005). What happens when funding is linked to publication counts? In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T Systems* (pp. 389–405). The Netherlands: Springer. https://doi.org/10.1007/1-4020-2755-9_18
- Butler, L. (2007). Assessing university research: A plea for a balanced approach. *Science and Public Policy*, 34(8), 565–574. <https://doi.org/10.3152/030234207X254404>
- Cagan, R. (2013). The San Francisco declaration on research assessment. *Disease Models & Mechanisms*, 6(4), 869–870. <https://doi.org/10.1242/dmm.012955>
- Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3), 288–301. <https://doi.org/10.2307/3151563>
- DeSanto, D., & Nichols, A. (2017). Scholarly metrics baseline: A survey of faculty knowledge, use, and opinion about scholarly metrics. *College & Research Libraries*, 78(2), 150. <https://doi.org/10.5860/crl.78.2.150>
- Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5(2), 103. <https://doi.org/10.1086/208721>
- Green, P. E., & Wind, Y. (1975). New way to measure Consumers' judgments. *Harvard Business Review*, 53, 107–117. <https://hbr.org/1975/07/new-way-to-measure-consumers-judgments>
- Haddow, G., & Hammarfelt, B. (2019). Quality, impact, and quantification: Indicators and metrics use by social scientists. *Journal of the Association for Information Science and Technology*, 70(1), 16–26. <https://doi.org/10.1002/asi.24097>
- Hammarfelt, B., & 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(7), 924–935. <https://doi.org/10.1002/asi.24043>
- Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 101(2), 1145–1163. <https://doi.org/10.1007/s11192-013-1221-3>
- Hicks, D. (2005). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T Systems* (pp. 473–496). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/1-4020-2755-9_22
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature News*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>
- Kirchhoff, K., Capurro, D., & Turner, A. M. (2014). A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, 28(1), 1–17. <https://doi.org/10.1007/s10590-013-9140-x>
- Kramer, B., & Bosman, J. (2016). Innovations in scholarly communication—Global survey on research tool usage. *F1000Research*, 5, 692. <https://doi.org/10.12688/f1000research.8414.1>
- Lemke, S., Mehrazar, M., Mazarakis, A., & Peters, I. (2019). “When you use social media you are not working”: Barriers for the use of metrics in social sciences. *Frontiers in Research Metrics and Analytics*, 3, 39. <https://doi.org/10.3389/frma.2018.00039>
- Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, 3(3), 57–72. [https://doi.org/10.1016/S1755-5345\(13\)70014-9](https://doi.org/10.1016/S1755-5345(13)70014-9)
- Ma, L., & Ladisch, M. (2019). Evaluation complacency or evaluation inertia? A study of evaluative metrics and research practices in Irish universities. *Research Evaluation*, 28(3), 209–217. <https://doi.org/10.1093/reseval/rvz008>
- Ma, L., & Ladisch, M. (2016). Scholarly communication and practices in the world of metrics: An exploratory study. *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, 132, 1–4 Retrieved from <http://dl.acm.org/citation.cfm?id=3017447.3017579>
- MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 69(3), 474–482. <https://doi.org/10.1002/asi.23970>
- McCullough, D. (2002). A user's guide to conjoint analysis. *Marketing Research*, 14(2), 18–23.

- McKiernan Erin, C., Schimanski Lesley, A., Muñoz Nieves, C., Matthias, L., Niles Meredith, T., & Alperin Juan, P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife*, 8. <http://dx.doi.org/10.7554/elife.47338>.
- Miles, R., Konkiel, S., & Sutton, S. (2018). Scholarly communication Librarians' relationship with research impact indicators: An analysis of a National Survey of academic librarians in the United States. *Journal of Librarianship and Scholarly Communication*, 6(1), eP2212. <https://doi.org/10.7710/2162-3309.2212>
- Moed, H. F. (2018). Assessment and support of emerging research groups. *FEMS Microbiology Letters*, 365(17), fny189. <https://doi.org/10.1093/femsle/fny189>
- National Information Standards Organization (NISO). (2016). *NISO RP-25-2016, outputs of the NISO alternative assessment project* [output report]. NISO. p. 86. Retrieved from http://www.niso.org/apps/group_public/document.php?document_id=17091&wg_abbrev=altmetrics
- Nicholas, D., Herman, E., Jamali, H. R., Abrizah, A., Boukacem-Zeghmouri, C., Xu, J., ... Świgon, M. (2020). Millennial researchers in a metric-driven scholarly world: An international study. *Research Evaluation*, 29(3), 263–274. <https://doi.org/10.1093/reseval/rvaa004>
- Nicholas, D., Jamali, H. R., Herman, E., Watkinson, A., Abrizah, A., Rodríguez-Bravo, B., ... Polezhaeva, T. (2020). A global questionnaire survey of the scholarly communication attitudes and behaviours of early career researchers. *Learned Publishing*, 33(3), 198–211. <https://doi.org/10.1002/leap.1286>
- Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, 63(2), 336–353. <https://doi.org/10.1002/asi.21669>
- Pautasso, M. (2012). Publication growth in biological sub-fields: Patterns, predictability and sustainability. *Sustainability*, 4(12), 3234–3247. <https://doi.org/10.3390/su4123234>
- de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—A literature review. *Research Evaluation*, 25(2), 161–169. <https://doi.org/10.1093/reseval/rvv038>
- Rousseau, S., & Rousseau, R. (2017). Being metric-wise: Heterogeneity in bibliometric knowledge. *El Profesional de la Información*, 26(3), 480–487. <https://doi.org/10.3145/epi.2017.may.14>
- Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, 91(2), 567–575. <https://doi.org/10.1007/s11192-011-0615-3>
- Sosteric, M. (1999). Endowing mediocrity: Neoliberalism, information technology, and the decline of radical pedagogy. *Radical Pedagogy*, 1(1), 1–41.
- Tenopir, C., King, D. W., Spencer, J., & Wu, L. (2009). Variations in article seeking and reading patterns of academics: What makes a difference? *Library & Information Science Research*, 31(3), 139–148. <https://doi.org/10.1016/j.lisr.2009.02.002>
- Tenopir, C., Allard, S., Bates, B. J., Levine, K. J., King, D. W., Birch, B., ... Caldwell, C. (2011). Perceived value of scholarly articles. *Learned Publishing*, 24(2), 123–132. <https://doi.org/10.1087/20110207>
- Tenopir, C., Levine, K., Allard, S., Christian, L., Volentine, R., Boehm, R., ... Watkinson, A. (2016). Trustworthiness and authority of scholarly information in a digital age: Results of an international questionnaire. *Journal of the Association for Information Science and Technology*, 67(10), 2344–2361. <https://doi.org/10.1002/asi.23598>
- Thelwall, M. (2018). Altmetric prevalence in the social sciences, arts and humanities: Where are the online discussions? *Journal of Altmetrics*, 1(1), 4. <https://doi.org/10.29024/joa.6>
- Tian, Y., Wen, C., & Hong, S. (2008). Global scientific production on GIS research by bibliometric analysis from 1997 to 2006. *Journal of Informetrics*, 2(1), 65–74. <https://doi.org/10.1016/j.joi.2007.10.001>
- Vermeulen, B., Goos, P., & Vandebroek, M. (2011). Rank-order choice-based conjoint experiments: Efficiency and design. *Journal of Statistical Planning and Inference*, 141(8), 2519–2531. <https://doi.org/10.1016/j.jspi.2011.01.019>
- Willmott, H. C. (2011). *Journal List Fetishism and the Perversion of Scholarship: Reactivity and the ABS List* (SSRN scholarly paper ID 1753627). Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=1753627>
- Wilsdon, J. R., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R. A. L., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (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>. Retrieved from <http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/>
- Wouters, P., & Costas, R. (2012). *Users, narcissism and the control: Tracking the impact of scholarly publications in the 21st century*. Utrecht, the Netherlands: Stichting Surf Retrieved from <http://research-acumen.eu/wp-content/uploads/Users-narcissism-and-control.pdf>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lemke S, Mazarakis A, Peters I. Conjoint analysis of researchers' hidden preferences for bibliometrics, altmetrics, and usage metrics. *J Assoc Inf Sci Technol*. 2021;1–16. <https://doi.org/10.1002/asi.24445>