Research Evaluation, 28(3), 2019, 241–252
doi: 10.1093/reseval/rvz009
Advance Access Publication Date: 5 May 2019
Article

OXFORD

Construction and validation of a short multidisciplinary research performance questionnaire (SMRPQ)

Martin Daumiller [1,*], Stefan Siegel[2] and Markus Dresel[1]

[1]Department of Psychology and [2]Department of Educational Science, University of Augsburg, Universitätsstr. 10, 86135 Augsburg, Germany

*Corresponding author. Email: martin.daumiller@phil.uni-augsburg.de

Abstract

Research is often specialized and varies in its nature between disciplines, making it difficult to assess and compare the performance of individual researchers. Specific qualitative and quantitative indicators are usually complex and do not work equally well for different research fields. Therefore, the aim of the present study was to develop an economical questionnaire that is valid across disciplines. We constructed a Short Multidisciplinary Research Performance Questionnaire (SMRPQ), with which researchers can briefly report 11 quantitative and qualitative performance aspects from four areas (research quality, facilitation, transfer/exchange, and reputation) in relation to their peer reference groups (fellow researchers with the same status and discipline). To validate this questionnaire, 557 German researchers from Physics, History, and Psychology fields (53% male, 34% post-docs, and 19% full professors) completed it, and for the purpose of convergent and discriminant validation additionally made assessments regarding specific quantitative and qualitative indicators of research performance as well as affective, cognitive, and behavioural aspects of their research activities (perceptions of positive affect, help-seeking, and procrastination). The results attested reliable measurement, endorsed the postulated structure of the newly developed instrument, and confirmed its invariance across the three disciplines. The SMRPQ and the validation measure were strongly positively correlated, and both demonstrated similar associations with affect, cognition, and behaviour at work. Therefore, it can be considered a valid and economical approach for assessing research performance of individual researchers across different disciplines, especially within nomothetic research (e.g. regarding personal antecedents of successful research).

Key words: research; performance; questionnaire; multidisciplinary; assessment

1. Measuring research performance: research gaps and shortcomings of current approaches

Recently, researchers and policy makers have become increasingly interested in the quantification of research activities in higher education and at non-university research institutes—in particular regarding research and its outcomes (Molas-Gallart 2015). This general rise in interest can, for instance, also be seen in the excellence initiatives in several EU countries (Wagner 2007). Next to agreements on budgeting and intensified evaluations, these initiatives also tackle assessments of, and knowledge of research quality (Ochsner, Hug, and Daniel 2016a). For evaluations and funding decisions, there is no doubt that it is necessary to analyse a broad range of research outputs using multiple detailed indicators (see also Gogolin and Stumm 2014). Aside from this, another important application context requiring measures of research quality is research itself, more precisely research on research quality; in particular, to facilitate nomothetic research (i.e. the study of large groups of individuals to derive general laws), so as to advance the understanding of personal processes supporting high research quality (e.g. how motivations, experience, and behaviour at work relate to research success). This

approach does not focus in an idiographic way on the individual diagnosis of the research quality of an individual researcher or an individual research institution, but rather in a nomothetic manner on relations between constructs. To describe them, an economical measurement is necessary. Assessing personal research quality, however, is considered a complex and costly endeavour and its comparability is often not granted, particularly due to the high specialization of researchers within their disciplines and fields (Hicks et al. 2015).

In consequence, economic, homogenous, and comparable performance indicators are needed, which are a function of the much more complex and multifaceted research quality (Print and Hattie 1997). Such indictors of research performance are developed primarily under a measuring rationale on the basis of the theoretical concept of research quality.[1] This multifaceted construct contains not only the production but also the dissemination and transfer of knowledge (Gaston 1970). While the amount of publications is often considered a central indicator (Carnegie Foundation for the Advancement of Teaching 1991), this broad definition implies that other aspects also need to be taken into account; for instance, the acquisition of third-party funding or editorial activities (Creswell 2012). Taken together, the measurement of research quality requires a set of criteria that allows an appropriate evaluation. The choice of these criteria and the use of quantitative or qualitative standards for their assessment often varies greatly, even within individual areas of expertise, leading to distortions and hindering inter-individual comparisons, especially across different fields (Lawrence and Green 1980; Hicks et al. 2015).

Up to now, while there are several indicators and sets of criteria, no multidisciplinary instrument for measuring research performance on a personal level exists (Ochsner, Hug, and Daniel 2016b). In order to derive an adequate set of criteria for the inter-individual and multidisciplinary measurement of research performance of scientists, we will consider different quantitative and qualitative indicators that are typically used, and discuss their advantages and disadvantages. The newly developed instrument is based on the integration of multiple indicators in order to yield a comprehensive contentual coverage of research quality, while combining the indicators' individual strengths and reciprocally compensating indicator-specific weaknesses.

In the next sections, we will therefore review literature regarding the measurement of research performance with quantitative and qualitative indicators and discuss the use of self-report for their assessment. Thereafter, we will describe our research aims and the assumptions and objectives of the Short Multidisciplinary Research Performance Questionnaire (SMRPQ), followed by a methodological section on the construction and testing of this instrument, as well as the results and their discussion.

1.1 Quantitative indicators based on publications and citations

Research outputs can be measured by the (weighted) number of publications (van Raan 2005). These indicators vary depending on the type of publication as well as the number of authors and their position. The weighting of these aspects can be problematic, since they often depend on particular disciplines (cf. for instance the different value of a monograph between Physics and History; see also Braxton and Toombs 1982). Also, if and how co-authorship is taken into account varies and can lead to different results (e.g. Howard, Cole, and Maxwell 1987; Scott and Mitias 1996; Bapna and Marsden 2002; Erkut 2002). Most critically, however, such approaches are often uneconomical, complex, and bound to a particular discipline: due to different publication practices and the above-mentioned characteristics, those indicators do not generally allow for an adequate comparison of research performance of individual researchers from different fields (Moed 2006; Abramo, D'Angelo, and Di Costa 2008; Wilsdon 2016).

In addition to the number of publications, research performance is often measured by citations. Citation indicators are typically based on statistical algorithms that signal how often a researcher is cited by others (Moed 2006). Since citations usually point to important ideas or results that are used to gain new knowledge, it is argued that the number of citations of a scientific work can be regarded as an indication for its reception (Schlinghoff and Backes-Gellner 2002). Apart from the discipline-induced shortcomings discussed above, this approach can also be biased; for instance, publications with already many citations or a methodological focus are cited more frequently than other publications (Voeth, Gawantka, and Chatzopoulou 2006; cf. also MacRoberts and MacRoberts 2018).

On an institutional level, rankings (e.g. World University Ranking THES; Times Higher Education 2015) are often used to compare different universities and non-university research institutions. Although their methodology and their overall quality vary greatly and are controversially discussed, they are still worth considering since they often encompass indicators that could be used or adapted to measure the research quality of individual researchers (Federkeil 2002; European Commission 2009). Apart from the above-mentioned amount of publications and their quality, rankings often contain data on third-party funding, amount of supervised PhD students, and research reputation. The German Council of Science and Humanities (2004, 2013) suggested that research quality indicators can systematically be distinguished in regard to different facets that can be grouped in four superordinate areas: (1) impact/quality, (2) reputation, (3) activities that facilitate research, and (4) transfer to the public. In a series of studies, they also presented first evidence that these areas, and the quantitative aspects that they encompass, can be distinguished across different fields. This is important as research practices and conventions differ considerably between natural sciences, social sciences, and the humanities (Kagan 2009). While their overview was suggested for the overall faculty level, it stands to reason that these aspects also characterize the research quality of individual researchers.

In summary, while quantitative indicators of research quality are considered to be objective, they often require complex algorithms, are limited in their comparability across disciplines, and often strongly depend on the methodology used (e.g. Adler and Harzing 2009). As described above, they often include (weighted) data on publications, third-party funding, citations, talks, conference organization, supervised PhD students, and reputation.

1.2 Qualitative indicators

While bibliometric indicators are established and generally more accepted in the natural sciences, the use of such strictly quantitative methods seems to only be possible to a limited extent in the social sciences and humanities and is discussed controversially (Hicks 2004; Guillory 2005; Nederhof 2011). For example, in the social sciences and humanities it is often argued that quantitative indicators based on publications and citations are not transferable due to

discipline-specific publication characteristics (Nederhof 2006). Furthermore, considering only quantitative indicators is regarded as problematic because other, often crucial aspects of research quality, such as originality, are neglected, and dysfunctional effects (e.g. loss of diversity; Fench 2009) may occur (Fisher et al. 2000). Therefore, especially in the social sciences and humanities, there have recently been extensive efforts to develop qualitative indicators for assessing research quality (Ochsner, Hug, and Daniel 2016a).

The results of several studies conducted by Ochsner, Hug, and Daniel (2012, 2014, 2016b) also emphasized that an evaluation of research quality based on quantitative indicators is limited in the social sciences and humanities. The authors argue, however, that an evaluation of research quality by means of qualitative criteria is promising within those disciplines, 'if a broad range of quality criteria are applied' (Ochsner, Hug, and Daniel 2016b: 64). Throughout a series of studies, the following criteria were agreed on by experts in German literature studies, English literature studies, as well as art history: (1) scholarly exchange, (2) innovation/originality, (3) rigour, (4) fostering cultural memory, (5) impact on research community, (6) connection to other research, (7) openness, (8) erudition, (9) enthusiasm, (10) vision, and (11) connection between research and teaching. The authors also developed a first version of a questionnaire to assess these aspects of research quality within the humanities (Ochsner, Hug, and Daniel 2016b). Since such qualitative indicators are considered to be particularly important in the humanities but also appear to be central to research quality in natural sciences, it consequently seems suitable to include them to assess research quality across different disciplines. To this end, it should, however, be noted that some of these criteria from Ochsner, Hug, and Daniel (2016b) do not seem suitable in this respect. Fostering cultural memory is presumably an aspect that matters in literature and art studies but does not appear to be a relevant aspect of research quality in harder sciences. Also, while the connection between research and teaching is probably an important determinant for a scholar's work experiences (cf. Daumiller and Dresel 2018), not all researchers are also teachers. Finally, we consider enthusiasm (like the experience of positive affect) as an important correlate of research quality, but not a central indicator of it.

1.3 Use of self-reports for assessing research performance

In this work, we propose the use of self-reports to economically measure research quality indicators of individual researchers from different disciplines. A viable alternative to this could be the use of peer reviews. In these procedures, scientists belonging to the same discipline use their expertise to evaluate the research of fellow researchers (Royal Netherlands Academy of Arts and Sciences 2011; van den Akker 2016). However, while such procedures often allow for an adequate understanding of the specificity of the field at hand, peer reviews also come with disadvantages: depending on the reviewer and his or her relationship to the researcher being reviewed, the evaluation may be biased (Tomkins, Zhang, and Heavlin 2017). Furthermore, due to the lack of transparency regarding the choice of reviewers and the review processes, intersubjective understanding, i.e. a shared evaluation amongst two or more reviewers is often not provided, making their use in large-scale studies problematic as reliability and validity are not granted (Demetriou, Ozer, and Essau 2014). Additionally, peer reviews are rather time-consuming and costly, and thus not an economical method for assessing research quality (Bornmann 2011; Guthrie, Ghiga, and Wooding 2018).

In the social sciences, self-report questionnaires are a common data source and are used for assessing performance at work (Garcia and Gustavson 1997). For instance, self-reported job performance was reported to be a strong predictor for the development of actual job performance (Abele and Spurk 2009). However, there are discussions about the accuracy and reliability of self-reports and whether they should be replaced by objective measures (e.g. Retelsdorf et al. 2010). Self-reports suffer from specific disadvantages: researchers using self-report questionnaires are dependent on the honesty of their participants.

For self-reports, it is very important to ensure anonymity and affirm potential respondents of this in order to increase their probability of participation and to obtain honest answers. Especially in studies with a small sample size, this can be a delicate issue and must be reflected carefully within the study design and data management. Apart from phenomena such as social desirability or impression management, researchers must also assume that the participants are capable of providing an accurate response, that is, that they are able to understand the questions, introspectively retrieve the requested information, and evaluate it adequately (Rosen, Porter, and Rogers 2017).

The limitations stated above (e.g. difficulties in understanding the items) can be solved or reduced through careful design of the instrument, and by explicitly validating it. Self-reports also have a series of advantages: first, they can equally comprise qualitative and quantitative aspects and allow participants to refer to their specific contexts. Second, they can be regarded as efficient and economical methods as they can easily be implemented for large samples and often constitute the only practicable possibility for assessing certain data. Third, self-report studies can be self-administered and conducted anonymously which may lead to honest responses (Garcia and Gustavson 1997) and enable researchers to reach a broad range of participants, including, for instance, those that might opt out of peer ratings. Finally, since scientists are usually experts in the field of research, it is plausible to assume that they can—to a substantial extent—validly assess their own research performance, and in such situations self-reports are commonly used (e.g. Retelsdorf et al. 2010). Since academic peers are the strongest reference group in academia (Minssen and Wilkesmann 2003), self-reports are expected to work especially well when participants are asked to compare themselves to a peer group. For the research question at hand, this implies that self-reports can be expected to work well when asking the participants to assess their research in relation to researchers from the same field with the same status (Ringelhan et al. 2013). All in all, it thus seems that self-reports, when applied properly, can be useful to economically measure the research quality of individual researchers from different disciplines.

2. Research aims and construction of the SMRPQ

Building on the importance of understanding personal research quality, and the absence of a multidisciplinary measure, we strived to develop and validate an instrument that measures research performance efficiently on the level of researchers and equally well across different disciplines.

To this end, we gave an overview of different aspects of the theoretical construct of research quality and the research performance

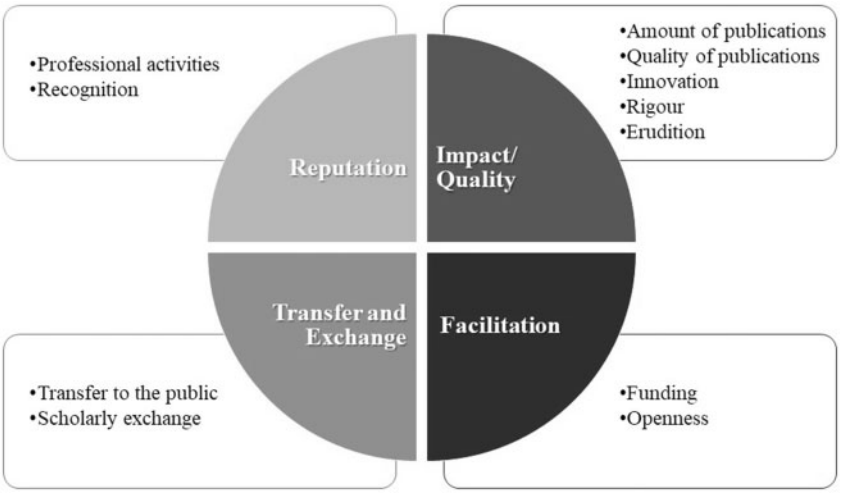

Figure 1. Overview of the distinguished aspects of research quality.

indicators used for its assessment. While specific qualitative and quantitative indicators exist, it is often unclear which of them should be used, especially since they do not work equally well for different disciplines. Moving forward from the above presented criteria and assessment methods, it should be noted that no methodology dominates all alternatives in all properties and therefore their results and interpretations are limited to the method-induced limitations. Consequently, to adequately evaluate research quality across various disciplines, a broad range of qualitative and quantitative criteria should be considered (Ochsner, Hug, and Daniel 2012; Hicks et al. 2015). To this end, based on the definition of what research quality entails, we combined the qualitative criteria of Ochsner, Hug, and Daniel (2013) that also appeared central in natural sciences, with the rather quantitative aspects and their differentiation into four areas proposed by the German Council of Science and Humanities (2010). We summarize these aspects below and propose their assessment through self-reports as an efficient and economical method suitable for (particularly nomothetic) purposes of research on science. For adequately measuring *impact/quality*, the amount of publications as well as their quality are taken into consideration (German Council of Science and Humanities 2010). Further indicators of this aspect encompass the innovation of one's research, as well as the rigour and erudition behind one's work (Ochsner, Hug, and Daniel 2016b). *Research facilitation* encompasses the amount of third-party funding as a rather quantitative aspect, as well as openness as a qualitative aspect (German Council of Science and Humanities 2010; Ochsner, Hug, and Daniel 2016b). In regard to *transfer and exchange*, we followed the suggestions from the German Council of Science and Humanities (2010) and included transfer of knowledge from the scientific to the public sphere (however, we excluded transfer of staff since this aspect is typically only relevant for full professors). In addition, and in line with Ochsner, Hug, and Daniel (2016b), we propose that scholarly exchange (i.e. within the academic community) should also be considered and therefore it was included as a separate aspect next to knowledge transfer. Finally, regarding *research reputation*, we included engagement in professional activities as well as one's recognition in the scientific field (German Council of Science and Humanities 2010; Hug, Ochsner, and Daniel 2013). Figure 1 provides an overview of these aspects.

Research quality is consequently considered a theoretical construct that encompasses a multitude of different facets, which is also in line with an emerging consensus in the fields of bibliometrics and research evaluation (e.g. Moed 2017: v). Only by acknowledging and including the various aspects mentioned above as indicators, a content-valid operationalization seems possible. Constructing a measurement instrument that is well suited for nomothetic studies on researchers puts restrictions on the sources and the amount of indicators to be used. As delineated above, we propose self-report relative to a peer reference group in order to economically measure the research performance of individual researchers from different disciplines. Here, also the duration of data collection also matters as it can be positively associated with expenses and costs and negatively associated with the commitment of participants. Consequently, we operationalized each of the aspects of research quality described above with a single item each in which we asked the participants to rate themselves relative to fellow researchers with the same status and discipline. To ensure high validity, we propose that all of these aspects of the underlying construct (see Figure 1) should be considered equally in a discipline-unspecific way (e.g. speaking generally of 'publication outputs' instead of specific aspects that only apply to some disciplines, such as the number of publications in peer-reviewed journals [with a high impact factor]) in order to allow an adequate conceptualization and an assessment that is not limited to a specific field, but valid across different disciplines, such as the natural sciences (e.g. Physics), humanities (e.g. History), and social sciences (e.g. Psychology). Consequently, the multiple facets of research quality are conflated in one single dimension reflecting a broad and overall estimate of research performance.

It is important to note that we are not measuring research quality directly, but rather *research performance* dependent on the reference group and the researchers' self-assessments on the employed indicators. This has important consequences regarding the interpretations of our findings and the inferences that could be derived from using the instrument. Specifically, while we expect the tasks that researchers have to fulfil in order to get the most out of research to be multidimensional, the research performance measured by our short scale should empirically portray a one-dimensional structure. As such, our findings need to be interpreted in light of the short assessment of research performance that can provide a useful instrument to

analyse the determinants, correlates and consequences of research quality, but do not serve conclusions regarding the internal structure of the more encompassing construct of research quality itself and do not allow for a valid evaluation of a single researcher's performance in an idiographic sense (which would be needed, e.g. in the context of funding decisions). On a general note, this illustrates the importance of being very clear and explicitly addressing different conceptualizations and their meanings—especially in empirical investigations of research quality.[1] Since research quality is a very broad and multifaceted construct, strivings to measure it typically have to curtail (e.g. by limiting the range of assessed facets and the indicators used). Consequently, there are often different concepts of research quality underlying a measure of research performance that frequently stay implicit. This is presumably an important reason as to why measuring research quality has often created problems and controversies (cf. Sonnentag and Frese 2005; Bazeley 2010).

Apart from investigating the measured relative research performance construct itself and its invariance across different disciplines, the validation of a measurement instrument like the SMRPQ should also include an examination of convergent and discriminant validity. Specifically, the newly developed measure should be strongly associated with a validation measure based on established instruments that are, in total, measuring research performance, while both measures should be similarly related to other constructs. To encompass a broad array of aspects, affective, cognitive, and behavioural aspects of research activities should be included. To this end, researchers' experiences of positive affect, attitudes towards help-seeking, and procrastination appear worthwhile, since they are likely associated with researchers' individual performance at work (e.g. Lyubomirsky, King, and Diener 2005; Kim and Seo 2015).

Taken together, the aims of the present study were to validate the newly constructed SMRPQ. For this purpose, we focused on four main points:

a. We expected that the SMRPQ portrays good descriptive psychometric features and can be well answered by researchers (i.e. participants feel confident when making their assessments).
b. We assumed that the latent construct measured with the SMRPQ can best be described with a one-dimensional structure (i.e. alternative models with 2 or 4 factors do not fit the data better).
c. Further, we wanted to test that this structure is invariant across the three different disciplines of Physics, History, and Psychology.
d. Finally, we expected that the SMRPQ is strongly positively correlated with a validation measure encompassing specific aspects of research performance measured with established instruments, and that both measures are similarly related to positive affect and considerations of help-seeking as useful (positive associations) as well as considerations of help-seeking as threatening and procrastination (negative associations).

3. Method

3.1 Procedure and sample

To pursue these research aims, we conducted an online study in which researchers made assessments concerning (1) the novel research performance questionnaire and how confident they felt when giving these answers, (2) a validation measure encompassing a variety of established instruments to assess specific research performance aspects, and (3) affective, cognitive, and behavioural aspects of their research activity (positive affect, attitudes towards help-seeking, and procrastination).[2] The researchers received a small incentive after participating in the study (a 5 € voucher). Regarding Physics, History, and Psychology—disciplines that are considered typical for the natural sciences, humanities, and social sciences (Kagan 2009)—we identified all public universities in Germany where these fields were taught, and randomly selected researchers from each university who were subsequently contacted by email (response rate: 14.7%). The study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists and anonymity was assured. It was not plausible that completing our survey would have any negative effects on the participants.

Altogether, 557 German university researchers (297 male, 260 female; mean age 36.4 years, $SD = 10.1$; 185 without PhD, 189 post-docs, and 106 full professors, disciplines: 184 Physics, 177 History, and 196 Psychology)[3] from 72 public German universities participated.

3.2 Measures

3.2.1 Short Multidisciplinary Research Performance Questionnaire

As described above, we developed one item for the distinguished aspects of research performance proposed by the German Council of Science and Humanities (2010) and Hug, Ochsner, and Daniel (2013). Therefore, the SMRPQ encompasses 11 items in total that serve as distinct indicators. Several researchers (including two psychologists, one historian, and one physicist) were consulted in the formulation, and iterative development, of item pools that subsequently led to the final items. Each item was formulated by systematically considering the delineated theoretical basis and operationalization of related constructs (e.g. German Council of Science and Humanities 2010; Hug, Ochsner, and Daniel 2013).

To answer them, participants were instructed to refer their answers to their research community, containing fellow researchers with the same status and discipline (cf. Ringelhan et al. 2013). Additionally, they were asked to focus their answers on the last 5 years using the item stem 'Over the last 5 years…'. Afterwards, participants were asked to evaluate the 11 indicators of relative research performance (e.g. regarding the quality of their publications: '…I published publications of higher quality than ___% of my fellow researchers with the same status and discipline'). The complete instrument is presented in Table 1.

Immediately after having completed this measure, we asked the participants to rate how confident they felt when making these assessments on a scale from 1 (*not confident at all*) to 8 (*very confident*).

3.2.2 Specific aspects of research performance used as validation measure

To validate the SMRPQ, we gathered assessments regarding specific quantitative and qualitative indicators of research performance using established measures focused on specific aspects of impact/quality, facilitation, transfer/exchange, and reputation that were directly derived from the conceptualization of research performance underlying the present work (cf. section 2). To this end, a wide range of (quantitative and more qualitative) measures has been taken into account. As in the SMRPQ, participants were asked to refer their answers to the last 5 years.

Table 1. Item statistics of the SMRPQ

Items (item stem: over the last 5 years. . .)	*M*	*SD*	Min	Max	Missing (%)	Skew	r_{it}
1. . . .I published more. . .	46.5	24.5	1	98	16	−0.01	0.63
2. . . .I published publications of higher quality. . .	46.1	23.0	1	96	18	−0.07	0.67
3. . . .I researched more innovatively and more originally. . .	50.0	22.5	2	99	10	−0.04	0.67
4. . . .I worked methodically more reflectively. . .	53.1	20.9	2	99	9	−0.33	0.60
5. . . .I had richer professional knowledge. . .	45.9	22.0	2	99	11	−0.05	0.66
6. . . .I raised more third-party funding (including Ph.D. scholarships). . .	47.2	27.3	1	99	31	−0.10	0.49
7. . . .I was more open to new things. . .	53.8	19.9	1	99	10	−0.22	0.52
8. . . .I committed myself to more scholarly exchange. . .	45.6	22.3	1	99	10	−0.09	0.70
9. . . .I communicated my research results to the public better. . .	45.4	23.1	1	95	12	−0.04	0.63
10. . . .I was recognized more. . .	41.7	22.6	1	95	18	−0.05	0.79
11. . . .I pursued more professional activities (e.g. editorships, reviews, and positions in professional associations). . .	41.8	23.8	1	99	27	−0.16	0.58

Note: $N = 557$. All items ended with '. . . than ___% of my fellow researchers with the same status and discipline'. Theoretical range: 1–99. r_{it} denotes item-total correlation. Presented are English translations of the original German items, which have not yet been validated in the English-speaking context.

Regarding the *amount of publications* as an aspect of impact/quality, we followed the distinction of the German Council of Science and Humanities (2010) and asked the participants to provide information about the overall quantity, and amount as first author, of (1) monographs/edited books, (2) book chapters, (3) peer-reviewed journal articles, (4) conference contributions (talks/posters), (5) invited talks, and (6) other publications. Alternatively, participants were given the option to upload their current publication list (which was then counted manually and inserted by means of an anonymous ID into the data set; 8% of the participants chose to do this). The answers were then aggregated using discipline-specific weights for the different types of publications. In order to assess further aspects of impact/quality we used the item stem 'Over the last 5 years, to what extent have you . . .', and then presented scales from Hug, Ochsner, and Daniel (2013) to assess *innovation* (2 items, e.g. '. . .created new findings or interpretations', $\omega = 0.64$), *rigour* (2 items, '. . .reflected your methods or your choice of methods carefully', $\omega = 0.94$), and *erudition* (2 items, '. . .gained professional expertise and insights', $\omega = 0.84$). These items were answered on a Likert-type scale ranging from 1 (*very little*) to 8 (*very much*). For reasons of consistency, unless otherwise stated, the same item stem and scale was used for the following aspects of research performance.

As a quantitative aspect of research facilitation, participants were asked to estimate the total sum of *third-party funding* received during the last 5 years (including doctoral scholarships) using a formulation suggested by the German Council of Science and Humanities (2010). As a qualitative aspect, we asked the researchers to make assessments in regard to their research *vision*, using the corresponding subscale from Ochsner, Hug, and Daniel (2012; 2 items, e.g. '. . .stimulated new research perspectives in your community?', $\omega = 0.74$). In addition, participants' *openness to ideas and persons* was assessed using a qualitative subscale from Ochsner, Hug, and Daniel (2012; 3 items, e.g. '. . .recognized other, competing ideas, approaches, theories, and methods', $\omega = 0.88$).

Regarding transfer and exchange, we used two subscales from Ochsner, Hug, and Daniel (2012) assessing *scholarly exchange* (2 items, e.g. '. . .contributed to scientific discourses that are related to your field of research?', $\omega = 0.71$) as well as the *transfer* of findings beyond the scientific community (2 items, e.g. '. . .have you engaged yourself in media and public relations work?', $\omega = 0.69$).

With regard to research reputation, we used a quantitative measure of participants' *professional activities* that was based on the suggestions of the German Council of Science and Humanities (2010). The respondents were asked to provide information on the number of their (1) activities as editors, (2) activities as reviewers, and (3) positions in professional associations. In regard to their *recognition*, participants were asked to state the total number of scientific awards and awards received over the last 5 years and to evaluate the reputation of these rewards on a Likert-type scale from 1 (*very low*) to 5 (*very high*). For the subsequent analyses, these two assessments were weighted (by multiplying them).

Prior to the analyses, all aforementioned single aspects of the validation measure were *z*-standardized within the three status groups and fields, and subsequently averaged ($\omega = 0.86$) to obtain a single validation measure comparable to the SMRPQ. This measure was then used to examine the convergent validity of the SMRPQ.

3.2.3 Affective, cognitive, and behavioural aspects

To measure positive affect when conducting research, we adapted an instrument developed by Keller et al. (2014) to the context of researching. Responding to three items (e.g. 'I very much enjoy doing research'; $\omega = 0.95$) a Likert-type scale ranging from 1 (*do not agree at all*) to 5 (*agree completely*) was used.

In order to assess the researchers' attitudes towards help-seeking, we used two scales of an instrument published by Dickhäuser, Butler, and Tönjes (2007) that we adapted slightly to the research context. With four items each, we assessed perceptions of help-seeking as threatening to the self (e.g. 'Asking for help as a researcher only shows your weaknesses'; $\omega = 0.79$) and perceptions of help-seeking as beneficial for learning (e.g. 'Asking others for help helps me to become a better researcher'; $\omega = 0.85$). The items were answered on Likert-type scales ranging from 1 (*not true at all*) to 5 (*completely true*).

With regard to procrastination, a scale by Klingsieck and Fries (2012) was used. Following the item stem 'In my research. . .', participants were asked to evaluate to what extent typical statements (e.g. '. . .I often find myself doing tasks I actually wanted to do days ago', 5 items, $\omega = 0.89$) applied to them. To this end, a Likert-type scale from 1 (*very untypical*) to 8 (*very typical*) was used.

3.2.4 Missing data

Over all items and participants, there was 12.9% missing data, which was dealt with model-based, using the full information maximum likelihood estimator and the expectation-maximization algorithm for all analyses with Mplus (Peugh and Enders 2004). Acknowledging that there could be certain groups of researchers for which specific items of the SMRPQ, e.g. in regard to acquired funding, might not matter much, we included auxiliary variables for these analyses (field, status, and job position). In regard to the quantitative validation measures focused on specific aspects of the research activity, the participants often did not provide answers to a few of the items used, possibly because the questions were difficult to answer (e.g. amount of conference submissions as first author: 38.4% missing answers) or inapplicable to some of the respondents (e.g. activity as editor; 67.8%). For the manifest analyses with this validation measure, we therefore used a similar procedure as before; by multiply imputing the data using a chained-equation model with predictive mean matching and 100 imputations. Imputed values compared reasonably to observed values and results ignoring the missing data (or setting them to 0 for variables that deemed inapplicable for some of the respondents) were similar, so imputed results are presented (Enders 2010).

3.3 Analyses

To confirm the postulated structure of the underlying research performance construct, confirmatory factor analyses were conducted with Mplus (Muthén and Muthén 2014) using the MLR estimator. χ^2 and SRMR were reported as absolute fit indices, TLI as a relative fit index that also adjusts for parsimony, and RMSEA and CFI as non-centrality-based indices. Due to the construction of the instrument, following the recommendations by Brown (2015), we a priori decided to model correlated errors between the similarly worded items 1 and 2 (focusing on the amount and quality of publications). Latent variables were standardized by setting their means to 0 and variances to 1. We compared more and less parsimonious models. Specifically, we compared (1) a model with one factor against, (2) two-factor models distinguishing between quantitative and qualitative aspects of relative research performance, and (3) a model with the four factors impact/quality, facilitation, transfer/exchange, and reputation. Taking the heterogeneity and breadth of the different items into consideration, it should be noted that a certain amount of model misspecification is automatically induced so that very good model fit indices seem unlikely. Consequently, lenient cut-off values should be used to assess satisfactory model fit (Fan and Sivo 2007). Thus, we used CFI >0.90, TLI >0.90, RMSEA <0.10, and SRMR <0.08 as cut-off values in the present work (cf. Schermelleh-Engel, Moosbrugger, and Müller 2003).

To confirm that the questionnaire works equally well for the investigated disciplines (measurement invariance), multi-group confirmatory factor analyses were conducted. We compared hierarchical models with imposing restrictions between the measurement models for the three groups (cf. Gregorich 2007): (1) a model in which the item-factor clusters were set as equivalent for all three groups (*configural invariance*), (2) a model in which we additionally restricted the factor loadings between the three groups (*metric invariance*), (3) a model with additionally restricted item intercepts (*scalar invariance*), and (4) a model with additionally restricted residual variances (*strict invariance*). The corresponding form of invariance can be assumed when a more restricted model does not describe the data worse than the previous, less restricted model. Differences in model fit were evaluated using -2ΔLL rescaled differences in the model log-likelihood values, as well as the differences between CFI and RMSEA (using the commonly suggested cut-off values of ΔCFI = 0.01 and ΔRMSEA = 0.015; Chen 2007).

Finally, we used bivariate correlations as well as regression analyses to assess the relationships between the SMRPQ and the validation measure, as well as the associations between both measures with the affective, cognitive, and behavioural aspects of research.

4. Results

4.1 Descriptive results

Descriptive statistics (Table 1, see also Table 2 for results of the overall scale) revealed means close to the theoretical mean of 50 (i.e. rating that one is average in regard to the indicator in question, with half of the other researchers with the same status and discipline being considered as better, and the other half as worse). At the same time there were substantial inter-individual differences (large variances), while most of the theoretical range was attained. Similarly, the analyses of the skew indicated only slight deviations from nil. Some participants seemed to have trouble answering two of the items (items 6 and 8, as seen in the rather large amount of missing data). These items expressed aspects that might not be applicable for all researchers (e.g. researchers who have not raised funding). All items yielded good item-total correlations that were clearly above the commonly suggested cut-off criterion of 0.30. Also, the internal

Table 2. Descriptive statistics and bivariate correlations

	M	*SD*	Min	Max	Skew	*1*	*2*	Δr
[1] SMRPQ	46.26	16.74	4.50	93.50	−0.20			
[2] Validation measure	−0.67	0.65	−2.82	1.67	0.09	0.75		
Positive affect	7.04	1.21	1.00	8.00	−1.94	0.32	0.39	0.07*
Help-seeking: Beneficial	7.05	0.94	2.00	8.00	−1.29	0.11	0.12	0.01
Help-seeking: Threat	2.34	1.19	1.00	7.00	1.06	−0.11	−0.17	0.05
Procrastination	4.70	1.59	1.00	8.00	−0.07	−0.15	−0.17	0.02
Age	36.35	10.13	23.00	78.00	1.31	0.28	0.37	0.09**
Gender (0 = male, 1 = female)	0.55	0.50	0.00	1.00	0.13	−0.27	−0.23	0.04
Hours per week researching	29.52	13.43	3.00	80.00	0.20	0.19	0.16	0.03

Note: $N = 557$. All correlations are significant at $P < 0.01$. Differences between correlations were tested using the Meng, Rosenthall, and Rubin's (1992) *t*-test. Columns headed 1 and 2 represent the correlations with the SMRPQ and with the validation measure., $^*P < 0.05$., $^{**}P < 0.01$.

Table 3. Results of measurement invariance testing

Model	*df*	χ^2	χ^2/df	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA	TRd	Δ*df*	*P*
Configural invariance	131	291.50	2.22	0.92	0.90	0.08	0.06					
Metric invariance	151	317.93	2.10	0.91	0.91	0.08	0.07	0.004	0.004	22.11	20	0.33
Scalar invariance	171	341.39	1.99	0.91	0.92	0.08	0.07	0.001	0.004	20.97	20	0.40
Strict invariance	191	365.09	1.91	0.91	0.92	0.07	0.08	0.002	0.003	22.29	20	0.34

Note: n(Physics) = 184, n(History) = 177, and n(Psychology) = 196.

consistency of the overall scale was very good (ω = 0.90) and an analysis of the split-half reliability yielded good results (Spearman–Brown = 0.88; Guttman = 0.87).

Finally, an analysis of the subsequent question of their confidence in their assessments showed that the participants felt, on average, rather confident with their answers ($M = 5.70$, $SD = 1.90$, skew = −0.10, attained range: 2–8, on a scale from 1 to 8) indicating that the SMRPQ can be well answered by researchers (cf. first research aim).

4.2 Structure of relative research performance

Regarding the structure of the measured relative research performance construct (cf. second research aim), confirmatory factor analyses indicated that a one-factor model described the data well ($df = 43$, $\chi^2 = 196.90$, CFI = 0.92, TLI = 0.90, RMSEA = 0.08, SRMR = 0.05). A two-factor model that distinguished between quantitative and qualitative aspects of research performance did not describe the data significantly better ($df = 42$, $\chi^2 = 187.37$, CFI = 0.92, TLI = 0.90, RMSEA = 0.08, SRMR = 0.05, ΔCFI = 0.005, ΔRMSEA = 0.001; Chen 2007). Estimation of a four-factor model that distinguished between the four main areas of relative research performance terminated unsuccessfully due to a non-positive latent variable covariance matrix that could be traced back to high correlations among the specified factors (as did a model in which we additionally included one second order factor). Therefore, we additionally conducted an Exploratory Factor Analysis. A parallel analysis as well as an investigation of the scree plot clearly pointed to a one-factorial solution (explaining 51.42% of the overall variance).

4.3 Measurement invariance across disciplines

The analyses of measurement invariance for discipline (cf. third research aim) showed that the more restrictive models assuming invariance across disciplines did not fit the data worse than the models allowing for variations between the three disciplines (Table 3). The −2ΔLL difference tests were not statistically significant and ΔCFIand ΔRMSEA were under the commonly suggested cut-off values of ΔCFI = 0.02 and ΔRMSEA = 0.015 (Chen 2007), indicating strict measurement invariance across the three fields.

4.4 Relationship with validation measure and affect, cognition, and behaviour

The SMRPQ was strongly correlated with the validation measure (corrected for attenuation: $r = 0.86$), indicating that the economical, self-reported answers in the SMRPQ yield similar results as the aggregated, specific indicators from the validation measure (cf. fourth research aim). At the same time, both measures had small associations with the other investigated aspects of research practise. Here, the correlations were similar for both measures, and portrayed only minor (and mostly statistically insignificant) differences, with the correlations for the SMRPQ being smaller than for the validation measure (Table 2).

5. Discussion

The overall aim of the present work was to develop and validate an economical, multidisciplinary questionnaire for assessing research performance relative to fellow researchers. To this end, we proposed an 11-item self-report instrument, the SMRPQ and tested its answerability, psychometric properties, structure, measurement invariance, and associations with a validation instrument and external criteria. Strengths of the study include its innovative focus on multiple disciplines, the broad sample, the analyses on the latent level, and the inclusion of extensive validation measures. Taken together, our results point to the objectivity, reliability, and validity of the newly developed measure and its usefulness with regard to its economy and multidisciplinary applicability and comparability, which may serve well for analysing the link with personal aspects at work.

All in all, our results attested that the newly developed questionnaire has good descriptive psychometric features, while the participants' confidence in their ratings indicated that the instrument can be answered as intended—and that the (unusual) response format (i.e. percentages in relation to other researchers) did not pose great difficulties for the participants (first research aim). However, our results also suggested that some of the assessed aspects may not be applicable for all respondents (e.g. those who have not raised funding) and need scrutiny within the further validation of the scale. In addition, it might be possible that inexperienced researchers generally could have difficulties answering some items, particularly if they have not yet attended conferences or conducted exchanges with their scientific community. In this regard, the scale format used in the SMRPQ might benefit from the inclusion of a 'not applicable' or a 'not sure'-category so that reasons for missing answers could be further quantified (Krosnick et al. 2002). Despite this, such missing values can be handled well within nomothetically oriented research, since they can imputed (including background variables such as subject and academic status) and the answers regarding the other indicators can still be used. Overall, however, the high degree of standardization of the questionnaire (instructions, examples, and items) can be taken as an indication for its objectivity of application, while the detailed description of the used instrument, the normed reference values, and the clearly defined meaning of the answers (percentage in regard to other researchers) indicates high objectivity in regard to its analysis and interpretation (Ary et al. 2018). Apart from that, the means of all assessed aspects were close to the theoretical mean of 50, which is to be expected if the surveyed researchers are representative concerning their research performance, and compare their research performance to others, i.e. they answer the scale format as intended (albeit some of them might have different

reference groups in mind when answering the items—with the random effects thereof being evened out throughout the sample). This indicates that systematic distortions seem unlikely. Also, taking the broadness of the investigated sample (in which the share of academic status, gender, and fields was very similar to the overall population) into consideration, this speaks for the validity of the scale format used. Especially the rather large variances observed and the use of nearly the complete range of the scale can be interpreted in a similar light, attesting that the measurement instrument is capable of describing inter-individual differences between researchers. Finally, we observed good item-total correlations for all of the assessed aspects, indicating homogeneity of the items, which was also seen in the split-half reliability and the internal consistency. Since we presented the participants with a self-report measure that primarily consists of evaluative questions, these aspects are decisive for the reliability of the scale (DeVellis 2016). Nevertheless, future research might additionally investigate other aspects of reliability using longitudinal designs (e.g. test–retest reliability). Taken together, the psychometric features and the design of the SMRPQ support its objectivity and reliability as well as the validity of the scale format.

Regarding the construct validity, our results confirmed the assumed one-dimensional structure of the research performance measure (second research aim). Discussing these results, it should be kept in mind that we strived for the development of an economical instrument. We did not construct the SMRPQ for diagnostic investigations of particular cases, but as a short scale to enable a fast measurement, specifically for purposes of research on science. In doing so, research performance measure was constructed as a one-dimensional broad and overall measurement. This does not necessarily contradict the assumed multifaceted nature of research quality, encompassing all tasks that researchers have to fulfil for high research outputs—which should not and cannot be investigated with the SMRPQ (however, we expect that using more items, e.g. three items per research performance aspect, would have supported models with multiple factors). Also the homogeneity of the items and their strong associations support the assumed one-dimensional structure of the measure, encompassing impact/quality, facilitation, transfer/exchange, and reputation (German Council of Science and Humanities 2013). Since the instruments' final items were selected systematically along these four areas based on previous literature (e.g. German Council of Science and Humanities 2010; Hug, Ochsner, and Daniel 2013) and their format (evaluation in percent relative to the relevant comparison group) was clearly defined (Borg and Shye 1995), this also speaks for the content validity of the instrument (Ary et al. 2018). The critical discussion with experts and their evaluation of the instrument as suitable (face validity) additionally reaffirms this.

Since the results indicated complete measurement invariance, it can be concluded that for different disciplines, the identical structure holds, and the identical construct is measured (third research aim). Even though the standards and research practises in Physics, History, and Psychology are quite different, these findings imply that the SMRPQ works equally well in these fields, expressing an aspect of external validity of the instrument. Since the three investigated fields are typical for natural sciences, humanities, and social sciences, we consider the stability of our findings across them as an indication that our results can also be transferred to other research fields. Taken together, these results strongly emphasize that the newly developed instrument can be used well in different disciplines and that the measured research performance can be compared adequately across them. Apart from its economy, this multidisciplinary applicability and comparability constitutes a main strength of the newly developed questionnaire.

In this context, an interesting aspect pertains to the statistical specification of the research performance construct. It should be noted that, following Bollen and Diamtopoulos (2017), changes in the latent research performance variable may not necessarily entail a simultaneous change in the value of its indicators. Indeed, it is plausible that researchers might have a very rich professional knowledge but not necessarily publish more or better than others.[4] This issue should be borne in mind for analyses on the latent level. In the present manuscript our latent analyses focused on group (or model) comparisons; however, future research focused on (statistical) effects of research performance on a latent level might consider specifying it in a formative way (Bollen and Diamtopoulos 2017). Especially when being interested in the outcome of scientific activities, it stands to reason that researchers who are doing well in the aspects of research performance summarized in Figure 1 (i.e. publishing a lot, working methodically very reflectively, researching innovatively and originally, etc.) should, as a consequence, also have a higher research outcome. Modelling the direction of this causality in statistical procedures by specifying it in a formative way therefore appears to be a useful research direction for such works.

Generally, our investigations also emphasize the need to make the theoretical conceptualizations (here: research quality) and their empirical measurements (here: research performance) of the constructs in question and what they entail explicit. In order to move the field forward and to resolve controversies, this should rigorously be taken up in investigations of research quality. It is obvious, that only by clear definitions and labels, can the assessed constructs be adequately understood and put in relation to one another.

Furthermore, we tested the instrument's convergent validity (fourth research aim). There was a remarkably strong correlation—documenting that similar constructs were measured—between the SMRPQ and the validation measure that was based on various established instruments encompassing specific aspects of research quality. Additionally, both measures had similar associations with the experience of positive affect, attitudes towards help-seeking, as well as procrastination, which can be considered as evidence for the convergent validity. These associations were in line with our expectations, and allow first results as to which affect, cognition, and behaviour accompanies good research quality. While there is an increasing interest in understanding the drivers behind research quality, most research to this end was conducted on an institutional and not a personal level (Ito and Brotheridge 2007). Using an approach on the individual level contributes to the little research that considers personal factors behind research quality. Here, our results especially pointed to meaningful associations with the experience of positive affect. These findings are in line with positive relations found between the experience of positive affect and job performance in general (Lyubomirsky, King, and Diener 2005), and are specifically in line with first results found for researchers: for instance, Dilger, Lütkenhöner, and Müller (2015) reported associations between the self-reported happiness of 49 German researchers attending a conference, and the amount and quality of their publications. Our results expand this line of research by indicating that the experience

of positive affect might be associated with the other aspects of research quality as well. We also found statistically significant associations with procrastination and researchers' attitudes towards help-seeking that were in line with our expectations and in a comparable small magnitude as results typically found in regard to these variables and job performance (e.g. Kim and Seo 2015). An interesting direction for future research on personal antecedents of research quality might therefore be the analysis of the underling motivation of researchers. In analogy to the importance of motivation (e.g. in the form of their personal goals or their self-efficacy beliefs) of university instructors for their teaching quality as rated by students (e.g. Daumiller et al. 2016) or themselves (e.g. Daumiller, Dickhäuser, and Dresel 2019), it seems likely that the motivation of researchers could be a crucial personal aspect behind differences in their research quality.

Although the SMRPQ, as well as the study at hand, has many strengths, there are also limitations that need to be borne in mind. First, we only investigated German researchers. Since the measurement invariance analyses indicated the robustness of our results, and the selected aspects of research quality are based on an international literature review, we are confident that our results can also be generalized to other countries. However, future research should specifically test this, since we cannot rule out that translations of the items used could be understood differently in other languages. Second, the SMRPQ does not allow for a status and discipline-unspecific ranking of research performance, because its respondents are asked to narrow their answers down to relevant reference groups. Since the purpose of the SMRPQ is to facilitate research on a nomothetic level by explaining inter-individual variance, and not to render global rankings (which might not be very informative anyway, e.g. since professors typically have higher research quality than graduate students), we do not consider this problematic; however, this limitation should be kept in mind when interpreting the results. As a last limitation, the self-report nature of the SMRPQ should to be mentioned. We specifically chose a self-report approach, since it enables to equally comprise qualitative and quantitative aspects of relative research quality, while allowing for discipline and status group specific assessments in an economical and efficient way; however, this also comes with a series of disadvantages since it can, in principle, be biased. While we consider our results as first evidence that researchers nevertheless make accurate and reliable assessments in the SMRPQ, we cannot rule out such distortions (e.g. answers being affected by participants' self-concept and social desirability). Since we did not construct the SMRPQ for diagnostic purposes with regard to particular cases, it should not be used in evaluative settings with high stakes (e.g. promotion of tenure). While the SMRPQ might be a valuable asset in formative settings, we see its main use for investigating the individual processes associated with conducting research. To this end, future research might benefit from additionally including social-desirability scales while ensuring full anonymity to participants.

Although we suggest further validation studies for the SMRPQ, the newly developed instrument can, in summary, be considered as a viable and economical, objective, reliable, and valid approach for assessing relative research performance of individual researchers across different disciplines, and seems well suited to facilitate future research, e.g. on personal antecedents.

Notes

1. We thank an anonymous reviewer for pointing us to the important distinction between the theoretical, more encompassing research quality and the practical, empirically measured research performance.
2. Specifically, the present study was conducted in the scope of a larger study conducted by authors (Daumiller 2018). Having completed the questionnaire of this larger study, the participants were asked if they were willing to take part in another short questionnaire regarding their research performance.
3. Twenty-four participants did not provide answers about their academic status.
4. We also investigated the prevalence of possible profiles in the present study. Latent profile analyses indicated three profiles that only differed in the overall level of the means and differences between Cronbach's Alpha (0.89) and McDonald's Omega (0.92) were negligible, indicating that for our data, hidden structures in the research performance construct seem unlikely.

Conflict of interest statement. None declared.

References

Abele, A., and Spurk, D. (2009) 'How Do Objective and Subjective Career Success Interrelate over Time?', *Journal of Occupational and Organizational Psychology*, 82/4: 803–24.

Abramo, G., D'Angelo, C. A., and Di Costa, F. (2008) 'Assessment of Sectoral Aggregation Distortion in Research Productivity Measurements', *Research Evaluation*, 17/2: 111–21.

Adler, N. J., and Harzing, A.-W. (2009) 'When Knowledge Wins: Transcending the Sense and Nonsense of Academic Rankings', *Academy of Management Learning & Education*, 8/1: 72–95.

Ary, D. et al. (2018) *Introduction to Research in Education*. Boston, USA: Cengage.

Bapna, R., and Marsden, J. R. (2002) 'The Paper Chase', *Or/MS Today*, 29/6: 34–9.

Bazeley, P. (2010) 'Conceptualising Research Performance', *Studies in Higher Education*, 35/8: 889–903.

Bollen, K. A., and Diamtopoulos, A. (2017) 'In Defense of Causal-Formative Indicators: A Minority Report', *Psychological Methods*, 22/3: 581–96.

Borg, I., and Shye, S. (1995) *Facet Theory: Form and Content*. Thousand Oaks, USA: Sage.

Bornmann, L. (2011) 'Scientific Peer Review', *Annual Review of Information Science and Technology*, 45/1: 197–245.

Braxton, J. M., and Toombs, W. (1982) 'Faculty Uses of Doctoral Training: Consideration of a Technique for the Differentiation of Scholarly Effort from Research Activity', *Research in Higher Education*, 16/3: 265–82.

Brown, T. A. (2015) *Confirmatory Factor Analysis for Applied Research*. New York City, USA: Guilford.

Carnegie Foundation for the Advancement of Teaching (1991) 'Change: Trendlines: The Payoff for Publication Leaders', *Change: The Magazine of Higher Learning*, 23/2: 27–30.

Chen, F. (2007) 'Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance', *Structural Equation Modeling*, 14/3: 464–504.

Creswell, J. W. (2012) *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Thousand Oaks, USA: Sage.

Daumiller, M. et al. (2016) 'Structure and Relationships of University Instructors' Achievement Goals', *Frontiers in Psychology*, 7/375: 1–14.

Daumiller, M. (2018) *Motivation von Wissenschaftlern in Lehre und Forschung: Struktur, Eigenschaften, Bedingungen und Auswirkungen selbstbezogener Ziele [Motivation of University Scholars for Teaching and*

Research: Structure, Attributes, Antecedents, and Consequences of Achievement Goals]. Wiebaden, Germany: Springer.

Daumiller, M., and Dresel, M. (2018) 'Subjective Perceptions of the Teaching-Research Nexus and Occupational Stress at Universities', *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 50/3: 126–38.

Daumiller, M., Dickhäuser, O., and Dresel, M. (2019) 'University Instructors' Achievement Goals for Teaching', *Journal of Educational Psychology*, 111: 131–48.

Demetriou, C., Ozer, B. U., and Essau, C. A. (2014) 'Self-Report Questionnaires', in Cautin, R. L. and Lilienfeld, S. O. (eds) *The Encyclopedia of Clinical Psychology*, pp. 1–6. Hoboken, USA: John Wiley & Sons.

DeVellis, R. F. (2016) *Scale Development: Theory and Applications*. Thousand Oaks, USA: Sage.

Dickhäuser, O., Butler, R., and Tönjes, B. (2007) 'Das Zeigt Doch Nur, Dass Ich's Nicht Kann [That Just Shows I Can't Do It]', *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie/Journal for Educational Psychology and Pedagogical Psychology*, 39/3: 120–6.

Dilger, A., Lütkenhöner, L., and Müller, H. (2015) 'Scholars' Physical Appearance, Research Performance, and Feelings of Happiness', *Scientometrics*, 104/2: 555–73.

Erkut, E. (2002) 'Measuring Canadian Business School Research Output and Impact', *Canadian Journal of Administrative Sciences/Revue Canadienne Des Sciences de L'Administration*, 19/2: 97–123.

Enders, C. (2010). *Applied Missing Data Analysis*. New York, USA: Guilford.

European Commission (2009) Report on Progress in Quality Assurance in Higher Education (Report from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions n. 487). Brussels, Belgium: COM.

Fan, X., and Sivo, S. (2007) 'Sensitivity of Fit Indices to Model Misspecification and Model Types', *Multivariate Behavioral Research*, 42/3: 509–29.

Federkeil, G. (2002) 'Some Aspects of Ranking Methodology—the CHE-Ranking of German Universities', *Higher Education in Europe*, 27/4: 389–97.

Fench, S. (2009) 'Journals under Threat: A Joint Response from HSTM', *Metascience*, 18/1: 1–4.

Fisher, D. et al. (2000) *Performance Indicators and the Social Sciences and Humanities*. Vancouver, Canada: Centre for Policy Studies in Higher Education and Training, University of British Columbia.

Garcia, J., and Gustavson, A. R. (1997) 'The Science of Self-Report', *Observer*, 10/1: 1–10.

Gaston, J. (1970) 'The Reward System in British Science', *American Sociological Review*, 35/4: 718–32.

German Council of Science and Humanities (2004) *Recommendations for Rankings in the System of Higher Education and Research*. Hamburg, Germany: German Council of Science and Humanities.

German Council of Science and Humanities (2010) *Recommendations on the Differentiation of Higher Education Institutions*. Lübeck, Germany: German Council of Science and Humanities.

German Council of Science and Humanities (2013) *Empfehlungen zur Zukunft des Forschungsratings [Recommendations on the Future of Research Ratings]*. Mainz, Germany: German Council of Science and Humanities.

Gogolin, I., and Stumm, V. (2014) 'The EERQI Peer Review Questionnaire. From the Development of 'Intrinsic Indicators' to a Tested Instrument', in Gogolin, I., Aström, F., and Hansen, A. (eds) *Assessing Quality in European Educational Research. Indicators and Approaches*, pp. 107–20. Wiesbaden, Germany: Springer VS.

Gregorich, S. (2007) 'Do Self-Report Instruments Allow Meaningful Comparisons across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework', *Medical Care*, 44/11: 78–94.

Guillory, J. (2005) 'Valuing the Humanities, Evaluating Scholarship', *Profession*, 11/1: 28–38.

Guthrie, S., Ghiga, I., and Wooding, S. (2018) 'What Do we Know about Grant Peer Review in the Health Sciences?', *F1000Research*, 6: 1335.

Hicks, D. (2004) 'The Four Literatures of Social Science', in Moed, H. F., Glänzel, W., and Schmoch, U. (eds) *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S & T Systems*, pp. 473–96. Dordrecht, Netherlands: Kluwer Academic.

Hicks, D. et al. (2015) 'The Leiden Manifesto for Research Metrics', *Nature*, 520/7548: 429–31.

Howard, G. S., Cole, D. A., and Maxwell, S. E. (1987) 'Research Productivity in Psychology Based on Publication in the Journals of the American Psychological Association', *American Psychologist*, 42/11: 975–86.

Hug, S. E., Ochsner, M., and Daniel, H.-D. (2013) 'Criteria for Assessing Research Performance in the Humanities: A Delphi Study among Scholars of English Literature, German Literature and Art History', *Research Evaluation*, 22/5: 1–17.

Ito, J. K., and Brotheridge, C. M. (2007) 'Predicting Individual Research Productivity: More than a Question of Time', *The Canadian Journal of Higher Education*, 37/1: 1–25.

Kagan, J. (2009) *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge, UK: Cambridge University.

Keller, M. M. et al. (2014) 'Feeling and Showing: A New Conceptualization of Dispositional Teacher Enthusiasm and Its Relation to Students' Interest', *Learning and Instruction*, 33/1: 29–38.

Kim, K. R., and Seo, E. H. (2015) 'The Relationship between Procrastination and Academic Performance: A Meta-Analysis', *Personality and Individual Differences*, 82/1: 26–33.

Klingsieck, K. B., and Fries, S. (2012) 'Allgemeine Prokrastination: Entwicklung Und Validierung Einer Deutschsprachigen Kurzskala Der General Procrastination Scale (Lay, 1986) [Procrastination: Development and Validation of the German Short Scale of the General Procrastination Scale (Lay, 1986)]', *Diagnostica*, 58/4: 182–93.

Krosnick, J. A. et al. (2002) 'The Impact of 'No Opinion' Response Options on Data Quality. Non-Attitude Reduction or an Invitation to Satisfice?', *Public Opinion Quarterly*, 66/3: 371–403.

Lawrence, J. K., and Green, K. C. (1980) *A Question of Quality: The Higher Education Rating Game*. Washington, USA: American Association for Higher Education.

Lyubomirsky, S., King, L., and Diener, E. (2005) 'The Benefits of Frequent Positive Affect: Does Happiness Lead to Success?', *Psychological Bulletin*, 131/6: 803–55.

MacRoberts, M. H., and MacRoberts, B. R. (2018) 'The Mismeasure of Science: Citation Analysis', *Journal of the Association for Information Science and Technology*, 69/3: 474–82.

Meng, X. L., Rosenthal, R., and Rubin, D. B. (1992) 'Comparing Correlated Correlation Coefficients', *Psychological Bulletin*, 111/1: 172–5.

Minssen, H., and Wilkesmann, U. (2003) 'Lassen Hochschulen Sich Steuern? [Can Universities Be Controlled?]', *Soziale Welt*, 54/2: 123–44.

Moed, H. F. (2006) *Citation Analysis in Research Evaluation*. Dordrecht, Netherlands: Springer Science & Business Media.

Moed, H. F. (2017) *Applied Evaluative Informetrics*. Cham, Switzerland: Springer.

Molas-Gallart, J. (2015) 'Research Evaluation and the Assessment of Public Value', *Arts and Humanities in Higher Education*, 14/1: 111–26.

Muthén, L., and Muthén, B. (2014) *Mplus (Version 7.3) [Computer Software]*. Los Angeles, USA: Muthén & Muthén.

Nederhof, A. J. (2006) 'Bibliometric Monitoring of Research Performance in the Social Sciences and the Humanities: A Review', *Scientometrics*, 66/1: 81–100.

Nederhof, A. J. (2011) 'A Bibliometric Study of Productivity and Impact of Modern Language and Literature Research', *Research Evaluation*, 20/2: 117–29.

Ochsner, M., Hug, S. E., and Daniel, H.-D. (2012) 'Indicators for Research Performance in the Humanities: Opportunities and Limitations', *Bibliometrie—Praxis Und Forschung/Scientometrics—Practice and Research*, 1/1: 1–17.

Ochsner, M., Hug, S. E., and Daniel, H.-D. (2013) 'Four Types of Research in the Humanities: Setting the Stage for Research Performance Criteria in the Humanities', *Research Evaluation*, 22/2: 79–92.

Ochsner, M., Hug, S. E., and Daniel, H.-D. (2014) 'Setting the Stage for the Assessment of Research Performance in the Humanities. Consolidating the Results of Four Empirical Studies', *Zeitschrift Für Erziehungswissenschaft*, 17/6: 111–32.

Ochsner, M., Hug, S. E., and Daniel, H.-D, eds (2016a) *Research Assessment in the Humanities. Towards Criteria and Procedures*. Cham, Switzerland: Springer.

Ochsner, M., Hug, S. E., and Daniel, H.-D. (2016b) 'Humanities Scholars' Conceptions of Research Quality' in Ochsner, M., Hug, S. E., and Daniel, H.-D. (eds) *Research Assessment in the Humanities. Towards Criteria and Procedures*, pp. 43–69. Cham, Switzerland: Springer.

Peugh, J. L., and Enders, C. K. (2004) 'Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement', *Review of Educational Research*, 74/4: 525–56.

Print, M., and Hattie, J. (1997) 'Measuring Quality in Universities: An Approach to Weighting Research Activity', *Higher Education*, 33/4: 453–69.

Retelsdorf, J. et al. (2010) 'Teachers' Goal Orientations for Teaching: Associations with Instructional Practices, Interest in Teaching, and Burnout', *Learning and Instruction*, 20/11: 30–46.

Ringelhan, S. et al. (2013) 'Work Motivation and Job Satisfaction as Antecedents of Research Performance: Investigation of Different Mediation Models', *Performance Management im Hochschulbereich/Performance Management in Higher Education*, 22/3: 7–38.

Rosen, J. A., Porter, S. R., and Rogers, J. (2017) 'Understanding Student Self-Reports of Academic Performance and Course-Taking Behavior', *AERA Open*, 3/2: 1–14.

Royal Netherlands Academy of Arts and Sciences (2011) *Quality Indicators for Research in the Humanities*. Amsterdam, Netherlands: Royal Netherlands Academy of Arts and Sciences.

Scott, L. C., and Mitias, P. M. (1996) 'Trends in Rankings of Economics Departments in the US: An Update', *Economic Inquiry*, 34/2: 378–400.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003) 'Evaluating the Fit of Structural Equation Models', *Methods of Psychological Research Online*, 8/8: 23–74.

Schlinghoff, A., and Backes-Gellner, U. (2002) 'Publikationsindikatoren Und Die Stabilität Wirtschaftswissenschaftlicher Zeitschriftenrankings [Publication Indicators and the Stability of Economic Journal Rankings]', *Schmalenbachs Zeitschrift Für Betriebswirtschaftliche Forschung/ Schmalenbach Business Review*, 54/4: 343–62.

Sonnentag, S., and Frese, M. (2005) 'Performance Concepts and Performance Theory', in Sonnentag, S. (ed.) *Psychological Management of Individual Performance*, pp. 1–25. Hoboken, USA: Wiley.

Times Higher Education (2015) *Times Higher Education World University Ranking 2014–2015*. <http://www.timeshighereducation.co.uk/world-university-rankings/>.

Tomkins, A., Zhang, M., and Heavlin, W. D. (2017) 'Reviewer Bias in Single- versus Double-Blind Peer Review', *Proceedings of the National Academy of Sciences of the United States of America*, 114/48: 12708–13.

van den Akker, W. (2016) 'Yes We Should; Research Assessment in the Humanities', in Ochsner, M., Hug, S. E., and Daniel, H.-D. (eds) *Research Assessment in the Humanities. Towards Criteria and Procedures*, pp. 23–29. Berlin, Germany: Springer.

van Raan, A. F. (2005) 'Measuring Science', in Moed, H. F., Glänzel, W., and Schmoch, U. (eds) *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patentstatistics in Studies of S&T Systems*, pp. 19–50. Berlin, Germany: Springer.

Voeth, M., Gawantka, A., and Chatzopoulou, G. (2006) 'Impact Auf Die Deutschsprachige Marketingforschung: Ergebnisse Einer Zitationsanalyse Der Deutschsprachigen Marketing ZFP-Jahrgänge 1979 Bis 2004 [Impact on the German-Speaking Marketing Research: Results of a Citation Analysis of German-Speaking Marketing Journal ZFP from 1979 to 2004]', *Marketing: ZFP*, 28/1: 7–20.

Wagner, G. (2007) 'Does Excellence Matter?', *Soziologie/Sociology*, 36/1: 7–20.

Wilsdon, J. (2016) *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. London, UK: Sage.