Case study meta-analysis in the social sciences. Insights on data quality and reliability from a large-N case survey

Jager, Nicolas W.; Newig, Jens; Challies, Edward; Kochskämper, Elisa; von Wehrden, Henrik

Published in:
Research Synthesis Methods

DOI:
10.1002/jrsm.1514

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Received: 9 December 2020 | Revised: 28 April 2021 | Accepted: 10 July 2021

DOI: 10.1002/jrsm.1514

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Case study meta-analysis in the social sciences. Insights on data quality and reliability from a large-N case survey

Nicolas W. Jager[1] | **Jens Newig**[2] | **Edward Challies**[2,3] | **Elisa Kochskämper**[4] | **Henrik von Wehrden**[5]

[1]Research Group on Ecological Economics, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

[2]Research Group Governance and Sustainability, Leuphana University of Lüneburg, Lüneburg, Germany

[3]Waterways Centre for Freshwater Management, University of Canterbury, Christchurch, New Zealand

[4]IRS Leibniz Institute for Research on Society and Space, Erkner, Germany

[5]Faculty of Sustainability, Leuphana University of Lüneburg, Lüneburg, Germany

Correspondence
Nicolas W. Jager, Research Group on Ecological Economics, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany.
Email: nicolas.jager@uni-oldenburg.de

Funding information
FP7 Ideas: European Research Council, Grant/Award Number: 263859

Abstract

Meta-analytical methods face particular challenges in research fields such as social and political research, where studies often rest primarily on qualitative and case study research. In such contexts, where research findings are less standardized and amenable to structured synthesis, the case survey method has been proposed as a means of data generation and analysis. The method offers a meta-analytical tool to synthesize larger numbers of qualitative case studies, yielding data amenable to large-N analysis. However, resulting data is prone to specific threats to validity, including biases due to publication type, rater behaviour, and variable characteristics, which researchers need to be aware of. While these biases are well known in theory, and typically explored for primary research, their prevalence in case survey meta-analyses remains relatively unexplored. We draw on a case survey of 305 published qualitative case studies of public environmental decision-making, and systematically analyze these biases in the resultant data. Our findings indicate that case surveys can deliver high-quality and reliable results. However, we also find that these biases do indeed occur, albeit to a small degree or under specific conditions of complexity. We identify a number of design choices to mitigate biases that may threaten validity in case survey meta-analysis. Our findings are of importance to those using the case survey method – and to those who might apply insights derived by this method to inform policy and practice.

KEYWORDS

case survey method, evidence-based governance, inter-rater reliability, meta-analysis, publication bias

Highlights

What is already known

- The case survey method synthesizes published qualitative case studies into quantitative data.

- Potential biases: Publication bias; interpretations by raters; variable characteristics.

What is new

- We did not detect significant publication bias in our analysis, but found that it may be advisable to diversify data sources.
- Effects of raters' coding behaviour were only detected to a minimal degree.
- Variables requiring relatively simple coding proved more reliable. Reliable data could also be derived from more complex coding settings.

Potential impact for *Research Synthesis Methods* readers outside the authors' field

- Case survey meta-analysis holds considerable potential to generate wider and more general conclusions from the large body of qualitative case studies
- Potential biases may be mitigated through design choices:
  - How many cases to include;
  - What are the minimum information requirements for primary case studies;
  - Which types of publications to include;
  - How to design variables and measurement scales;
  - How many raters to employ.

1 | INTRODUCTION

Meta-analytical and systematic review methods offer a powerful means to cumulate and synthesize existing knowledge. Such methods are increasingly called for in fields that have traditionally relied primarily on qualitative and case study research.[1–5] However, meta-analytic methods face particular challenges in research settings where case studies and qualitative data predominate.[6,7] This applies, for example, to many fields of political and organizational research, where empirical insights are mainly derived from an ever-growing body of case studies[8]; or in certain areas of public health research, where research focuses on the context and implementation of public health interventions.[7,9] Such research settings pose particular challenges for the cumulation and subsequent analysis of evidence, because while single or small-N comparative case studies provide rich insights,[10] transferability and generalizability of findings are inherently limited.[11]

The case survey method has been designed to bridge this gap between focused, qualitative case narratives on the one hand, and more widely applicable meta-analytic research on the other. Developed over the last four decades,[12–14] the case survey method aims to produce systematic synthesis of existing case-based research, drawing on published case studies. By means of a conceptually grounded coding scheme, qualitative case narratives are translated into quantitative data through coding – typically by multiple raters –, allowing for statistical analysis. The method thus provides for much wider generalization and extrapolation over diverse settings and contexts.[3,15]

While in principle, the case survey method has considerable advantages, its uptake has remained limited.[16–18] Hence, little insight into the potential and limitations of this method is available, especially around methodological threats to data quality and validity. Indeed, given that the method relies on reinterpretation of original narrative case data, it may be prone to specific methodological risks regarding validity. In particular, publications and sources included in analysis may be subject to publication bias; during coding (i.e., transforming narratives into numeric data), the coding behaviour of raters and their experience may introduce bias; and different types of variables may pose different challenges in the coding process. While some of these potential sources of bias, such as publication and rater bias, are known in principle and have been explored in relation to primary research,[19,20] there is little experience with the methodological limits and biases specifically associated with case survey methods, and the validity and quality of the resulting data.

This article contributes new insights into the applicability of the case survey method in particular, and meta analyses more generally, by empirically investigating and quantifying the potential sources of systematic error in application of these methods. Data for this analysis come from what is, to our knowledge, one of the largest and

most comprehensive case surveys in the broader field of political science.[1] The project focuses on the question of 'what works' in participatory environmental governance.[21,22] We analyzed 305 published cases of more and less participatory environmental decision-making processes, using a coding scheme of 250 variables.[22] Each case was coded by three independent raters. While arguably, the case survey we report on here is larger than average in terms of sample size and resources needed, it is precisely these qualities that allow us to draw a number of insights on data quality and reliability, which should be of use for the wider application of the method.

The remainder of this article proceeds as follows. In Section 2, we outline the case survey method in more detail, and discuss the potential biases to which the method is prone. Section 3 specifies our data sources and the method by which we generated our dataset. Section 4 presents the results of our analyses of publication bias, rater influence, and variable characteristics. The results are discussed in Section 5, where we also draw some wider conclusions for design-choices in case study meta-analysis and make recommendations for the application of the case survey method.

2 | CASE SURVEY META-ANALYSIS: APPROACH AND POTENTIAL SOURCES OF BIAS

In contrast to standard meta-analyses, which draw on and aggregate quantitative datasets,[23–25] the case survey method transforms *qualitative case-based narratives into quantitative data*.[12,13] Hence, the method is particularly useful in fields where case studies dominate,[13] where the unit of analysis is at the meso- or macro-social level (i.e., not primarily studying individuals), where a wider range of contextual conditions are of interest, and where experimental research designs are difficult or impossible.[26] For example, as political analysis typically studies institutions, processes, policies, countries and other administrative units, the method is well suited to synthesizing findings and contributing to the systematic cumulation of knowledge in the field.

Conducting a case survey typically follows four main steps[14,26]:

1. *Case study identification and selection*: Starting from a given research question, the 'case' unit, and the universe of possible cases, is defined. Cases are identified via a thorough and structured literature search, ideally triangulating multiple strategies. Here, systematic review methods offer valuable guidance on how to identify and select cases in order to arrive at a case set representative of the wider phenomena under study.[27] Case studies are not excluded on the basis of type of research design, publication status, or time period as is often done in traditional meta-analyses.
2. *Development of a coding scheme for the transformation of qualitative into quantitative data*: The core of the case survey approach is a coding scheme that translates the research questions and phenomena of interest into a set of precisely formulated variables and decision rules.[28] These rules, based on theoretical and conceptual groundwork, instruct raters how to transform the narrative information for each variable into quantifiable data by allocating case information on a numeric scale. When designing these coding rules, researchers face a trade-off between resource-saving, reliable simplicity and information-dense complexity. While, for example, straightforward 'yes-no' questions, variable scales with fewer options (e.g., a 3-point Likert scale), or variables that ask only for easily available information may allow for more reliable results, such an approach may risk losing essential nuance and variety among cases. Hence, it is advisable to start with a more comprehensive coding scheme to elicit maximum information, and to simplify, if needed, until satisfactory reliability is reached.[14,26]
3. *Coding of cases*: Trained raters independently code each of the sampled cases according to the coding scheme. Given the interpretative nature of the method, at least two raters should be assigned to each case in order to be able to establish measurements for inter-rater reliability.
4. *Analysis of coded data*: Conventional procedures of statistical or otherwise structured data analysis can be used.

In light of these methodological steps, the validity and reliability of data generated in case survey approaches may be affected by at least three important, method-specific sources of bias, which we discuss below in more detail.

2.1 | Publication biases

Publication bias occurs 'whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies'.[29] Commonly, representation may be distorted in that cases with specific significant or clear results are overrepresented,[30,31] which would impact on the external validity of the findings of subsequent analysis. Most work on the publication bias, however, focuses on its effect in quantitative studies, whereas the question of potential biases in qualitative and small-N studies remains relatively neglected.[29]

Strategies to mitigate the effects of publication bias in (quantitative) meta-analyses entail the identification and inclusion of unpublished studies and results,[32] as well as publications from less formal outlets, such as grey literature.[33,34] Research may be omitted from academic publishing for various reasons: 'File drawer' research might have been intended for academic publication, but for some reason remained unpublished. In contrast, 'practitioner-generated' studies, such as reports compiled by government entities, consultants, or international organizations or NGOs, may not have been intended for an academic audience.[35] It is assumed that including such publications contributes to a more complete and comprehensive picture of available evidence,[27,33] as they may report on studies with less conclusive or less significant results. As a working hypothesis, we assume *that cases reported in (peer reviewed) journal articles will be more 'successful' than those from other sources, i.e. will display higher values for main dependent and independent variables and larger effect sizes between these.*

Including cases published in grey literature may potentially have a positive impact on the validity of case survey data, but it also raises questions about the reliability and density of information provided in different publication types. The various types of sources through which a given case is described may exhibit considerable differences in information quality, challenging the overall integrity of the results of a meta-analysis.[24] Studies appearing in more formal outlets, such as peer-reviewed journals, have been through some quality control providing a degree of scientific rigor and validity, but usually have only limited space for detailed case descriptions, which may be found in less formal publication sources such as grey literature or books. We assume, therefore, that *publication type has an effect on a case's density of relevant information and therefore the reliability of this information.*

2.2 | Coding biases

The coding process is a second potential source of bias. To arrive at reliable data coding processes should be replicable, meaning that different raters should code the underlying data in the same way.[36] Yet, in reality raters, who bring different personal backgrounds, knowledge and research experience to the coding exercise, may 'package' coding tasks differently,[37,38] potentially leading to systematic errors[39] that distort the data, and undermine the integrity of research findings.[20] Despite prior training, raters may exhibit particular, individual patterns of coding. For example, one rater may tend to interpret coding thresholds rather conservatively resulting in lower codes, while another may lean towards codes at the margins of measurement scales. We hypothesize, therefore, that *raters differ in their individual judgement patterns, which in turn leads to biases in the case survey data.*

Stability, or the extent to which a coding process remains consistent over time, is another essential aspect of data reliability.[36] However, stability may be threatened as individual raters' coding patterns may change over time as they assess more cases. This is sometimes also called 'coder drift'.[40] Such changes in coding behaviour might occur if, for example, raters learn and change their interpretation of specific concepts, or display 'fatigue effects' as case coding becomes more of a routine.[38,40] We therefore assume that *raters' experience (i.e. the number of cases they have already coded) influences the judgement of raters and in turn leads to biases in the case survey data.*

2.3 | Variable types and inferential judgement

Depending on the type of variable, raters may face different requirements for inference during the coding process. For some variables coding decisions are straight-forward, as raters only have to locate the information in the study at hand and record it; codes assigned for these variables can be framed as low-inference codes.[24,41] Such variables, for instance, could describe basic information of a given case, such as its location or the participation of specific actors. However, often raters are required to make relatively high-inference judgements – that is, make informed, individual assessments under constraints of complexity or uncertainty. This may occur where variables aim to measure complex social concepts, and when the information required is not clearly or explicitly reported in the underlying primary case material. An example could be assessing trust and social capital among actors within a specific group, or assessing actors' (hidden) intentions. Given that this is rather common in case-survey meta-analyses, high-inference codes may constitute a special source of bias impacting on data reliability and quality. Hence, we assume that *different types of variables, characterized by different requirements for inferential judgement, can be associated with different degrees of data reliability.*

3 | DATA: CASE STUDIES ON PUBLIC ENVIRONMENTAL DECISION-MAKING

The case survey data with which we test these assumptions were derived from a multi-year research project, which investigated how, and under what conditions,

different modes of public and stakeholder participation (as opposed to hierarchical modes of governance) affect environmental outcomes of public environmental decision-making processes.[42,43] We conducted a case survey of 305 published cases of participatory environmental governance, drawing on a wide and diverse body of knowledge in meta-analytical and case survey research from various social science disciplines (e.g., political science and governance, [inter-]organizational research).[12–14,16,17,24,26,44,45] Using a comprehensive coding scheme of variables capturing the context, the process, and the outputs and outcomes of decision-making,[22] the study is, to our knowledge, one of the most comprehensive case surveys in the broader field of political science to date. As such, it provides a unique setting within which to systematically and quantitatively test and control for data quality and potential biases and to learn about these for the benefit of other researchers and future applications of the method.

We define a 'case' as a public environmental decision-making process aiming for a collectively binding decision, which is to a lesser or greater extent *participatory* – in the sense of involving stakeholders, including citizens, not typically engaged in such decision-making.[46] In our meta-analysis, one case does not necessarily correspond to one study, but may be described in multiple studies. Similarly, one study may describe multiple cases of public environmental decision-making see also Reference 47.

In the literature, it is widely assumed that different forms of participation, in different contexts, help to produce more environmentally sound decisions and foster their implementation.[21,48,49] However, these claims are contested, and while there is a considerable amount of empirical evidence, this is scattered across many single or small-n comparative case studies, which makes it difficult to generalize. In order to be able to test such hypotheses, we measure in each case (1) the degree of participation and (2) the environmental standard of the output/decision, in a variety of different dimensions, and thus through a multitude of different variables.[2] In addition, we capture a range of data on secondary social outcomes, implementation-related aspects, and context factors.[22]

In conducting the case survey analysis, we followed the process outlined above (Section 2):

1. *Case study identification and selection*: We conducted a thorough search of several online scientific databases and library catalogues to capture research from numerous disciplines. We limited our search to cases from Europe, North America, and Australia and New Zealand, and to texts written in English, German, French or Spanish. We included a wide variety of publication types, such as peer-reviewed journal articles, books, edited collections and chapters therein, theses, working papers, conference papers, reports and other forms of grey literature, so long as these were publicly available. We thus identified over 2000 cases, described in over 3300 individual texts. We then screened these for suitability, assessing whether studies conformed to our case definition (public decision-making on an environmental issue) and contained sufficient information on the actual decision-making process, its outcomes and context. Studies that did not meet these criteria were excluded. From the resulting database of 639 'codeable' cases, described in around 1245 single texts, we randomly sampled 305 cases for coding. These cases were documented in 431 individual publications, with a majority of 198 cases being documented in a single text, and the rest requiring collation of information from up to 6 different texts.
2. *Coding scheme development*: The elaborated coding scheme reflects our conceptualization of participatory decision-making processes, and the hypothesized links between process attributes, outputs and outcomes, implementation, and environmental impacts, as well as context – and breaks these components down into multiple variables.[22] Our coding scheme went through several iterations of testing and adjustment, until we arrived at a final version comprising 259 quantitative (and additional qualitative) variables. Each variable definition specifies the measurement scale and provides detailed instructions for coding. The vast majority of variables are coded on a five-point quantitative scale (from 0 to 4). As suggested by Yin and Heald,[13] each variable is assigned an additional code capturing the reliability of the information (from 0 to 3) upon which the coding decision is based.[3] This gives an indication of the quality of the underlying information, and allows for ex-post information-guided selection of cases and individual variables for analysis, similar to critical appraisal procedures used in systematic reviews.[50]
3. *Case coding*: Each case was independently read and coded by varying teams of three trained raters, most of whom were student assistants. Research suggests that three raters are sufficient to realize the majority of the improvement in data quality that could be achieved by using a very large number of raters.[44] Coding occurred via an online form and database. Raters then met to discuss discrepancies and deviant codes, aiming to address technical errors and explore divergent interpretations, but not to reach consensus. By averaging across the three individual codes, different interpretations of the texts by individual raters were accommodated.[45]

The resulting dataset forms the basis for this analysis. For the subsequent analyses, we use two different sets of variables: The assessment of effects of publication types and rater personalities relies on a set of 186 variables. These are all measured on a [0, 4] or [−4, 4] scale. In defining this sub-set, we aim to preserve comparability between variables and to eliminate potential distortions stemming from alternative measurement scales, such as binary or count scales. For the analysis of variable characteristics, we use a wider set comprising 231 variables.

4 | FINDINGS ON THE QUALITY OF CASE SURVEY DATA: PUBLICATION TYPES, RATERS AND VARIABLES

4.1 | Publication bias: Potential distortions due to publication type

Controlling for publication bias is difficult, as the unbiased 'true' value of a given phenomenon cannot be known. We follow previous studies[51–53] and approach this problem by comparing cases published in different kinds of outlets (e.g., peer-reviewed to grey literature). While this method cannot account for those results that remain in researchers' file drawers without ever being written up and published,[19,54] two factors make this approach the most promising at hand: First, in social science, and in comparative governance research in particular, where a case study usually comprises a number of alternative hypotheses or more exhaustive interpretations of the case study material, a complete rejection of the null hypotheses is highly unlikely. Hence, selective reporting is actually more likely than a complete lack of data,[55] mitigating the 'file drawer' problem. Second, measures suggested for the analysis of publication bias in quantitative meta-analyses, such as funnel plots, trim and fill techniques, and selection models,[31,54] are not transferable as they require the quantitative estimation of the relations of interest. Hence, a comparative analysis of peer-reviewed publications, non-peer reviewed publications, and grey literature likely provides a robust estimate of publication bias in this understudied area of the case survey literature.

Our sample includes various types of information sources. We distinguish between (peer-reviewed) journal articles,[4] books and chapters therein, and grey literature, with several cases comprising a combination of these publication types. According to this classification, 24.6% ($n = 75$) of cases are exclusively from peer-reviewed journals, 34% ($n = 104$) are exclusively described in commercially published outlets like books and non-listed journals, and 18.7% ($n = 57$) are exclusively published in grey literature. The remaining cases ($n = 69$, 22.6%) are covered by multiple records from various of these sources (e.g., one peer-reviewed journal article and one book chapter). In the below analysis, unless noted specifically, these 'mixed' cases are excluded from the study, as they do not permit a clear attribution to one category. As stated above, we assume that cases described in journal articles will display (a) higher values for each main dependent and independent variable, and (b) larger effect sizes between these.

In our study of the effectiveness of participatory environmental decision-making, the degree of participation serves as the main independent variable, while the environmental standard of the output is considered the key dependent variable. These are complex, multi-dimensional concepts, and are therefore constructed as composites of multiple single variables. For further details on the composition of these variables please consult the supplementary material.

As each case was coded independently by three raters, our design involves three instances of each case. In order to account for this, we analyze the effects of different publication types on our main variables using multi-level modelling (hierarchical linear modelling).[56] Level 1 unit of analysis comprised the three single instances of each case (one coded by each rater), while level 2 unit of analysis comprised the cases as such. We built these models around our variables of interest, considering publication type as fixed effect, and the cases as such as random effect, allowing intercepts to vary among these.[5] Maximum likelihood criteria were used for fitting the models. Analyses were performed with R.[6]

TABLE 1 Pairwise comparison (Tukey contrasts) of publication types for public participation

				95% Confidence intervals	
Pairwise comparisons	**Mean difference**	**Standard error**	***p*-value**	**Lower**	**Upper**
Journal–book	0.05	0.14	0.94	−0.29	0.38
Journal–grey	−0.23	0.17	0.34	−0.62	0.15
Grey–book	0.28	0.016	0.17	−0.08	0.64

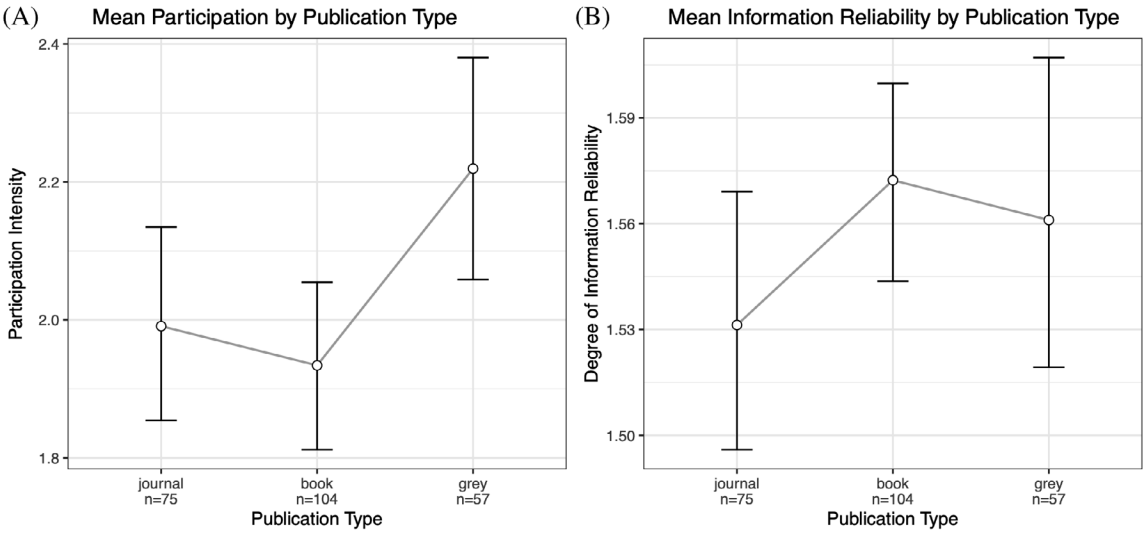

FIGURE 1 (a) Mean degree of participation per publication type. (b) Mean degree of information reliability per publication type. *Note*: Error bars display 95% confidence intervals

Considering the main independent variable, the model detects a slight, yet not significant, effect of publication type on the degree of participation described in the case studies [$\chi^2(2) = 3.40$, $p = 0.18$]. We specify this effect further through a post-hoc test (Tukey's all-pair comparisons) to investigate precisely the differences between publication types (see Table 1), and this supported the results presented in Figure 1(a). While cases described in journal articles and in books or book chapters show very similar values, cases reported in grey literature show a slight effect, which is, however, not significant: The mean level of participation (i.e., the independent variable for cases described in this type of publication) is around 0.28 standard deviations higher than for cases described in book chapters.

Fitting our model in the same way around our main dependent variable, the environmental standard of the output, results pointed in the same direction – grey literature with slightly higher output values – but also did not show any significant effect of publication type [$\chi^2(2) = 2.54$, $p = 0.28$]. Hence, publication type did not prove to be a significantly better predictor for the main dependent variables than the null hypothesis.[7]

As publication bias is assumed to be a function of the magnitude and significance of effect estimates between independent and dependent variables,[57] we also examine the relationships between our main variables. To this end, we calculated correlations between participation and the environmental standard of the output for each publication type and compared the resulting coefficients by means of a Williams test.[58]

Generally, correlation coefficients indicate a moderate to strong effect between participation and environmental output variables, ranging from 0.28 for cases described in journal articles or grey literature to 0.39 for those from books and chapters.[9] However, among the different publication types no clear pattern is apparent. No type stands out, and also the comparison of minimum and maximum coefficients for each output measure proved insignificant [Williams test grey vs. book: $z = 1.18$, $p = 0.24$].

However, a noteworthy effect emerges here, when including the 69 cases from the 'mixed' category. Those cases were reported in multiple studies published in different types of outlets (e.g., peer-reviewed and grey literature). On average, cases in this category are reported in 3.07 texts per case, while cases from the other three publication categories are based on 1.15–1.25 studies per case on average. Within this mixed category, the correlation coefficient between participation and environmental output is 0.48, which is significantly higher than that of grey and peer-reviewed literature (but not book chapters) [Williams test, grey vs. mixed: $z = 2.19$, $p = 0.03$].

Based on these investigations the hypothesis that peer reviewed journals display stronger relations between key independent and dependent variables cannot be supported. Most of our analyses did not find any significant difference between various types of publications. Where effects could be detected, as in the case of the main independent variable of public participation, differences between publication types were small, not significant, and showed higher values in grey literature rather than in more formal published or peer-reviewed publications. One exception is the significant difference among correlation coefficients between cases reported in a mix of publications and those reported in grey and peer-reviewed literature.

Publication type may not only have an influence on the validity of findings, but also, potentially, on the quality of the underlying case material and the resulting information reliability. To assess the effects of publication type and amount of case material, we established a measure to gauge the information reliability of our data (see above). We aggregated these reliability scores by calculating the arithmetic mean as a single reliability indicator for each case. This indicator incorporates the codes of three raters for 305 cases, and 186 variables. Mean information reliability over all cases and variables is 1.55 ($SD_{\text{info reliability}} = 0.28$).

As with the previous analysis, we tested the effect of publication type on information reliability through a multi-level modelling approach. Additionally, we controlled for the word count of each publication in the analysis as a means to test the effect of the volume of available case material. Figure 1(b) summarizes the distribution of information reliability means together with respective error bars (95% confidence intervals). We find only minor deviations between the different publication types, with book chapters ($mean_{book} = 1.57$) containing slightly more reliable information than the other types ($mean_{journal} = 1.53$, $mean_{grey} = 1.56$). This finding is mirrored by the results of the multilevel models, where no significant effect could be detected.[9]

Overall, analysis of publication bias revealed only small differences between grey literature and other, often more controlled forms of publication (e.g., book chapters and journal articles) concerning our variables and effects. Contrary to our expectations, cases described in grey literature tend to contain processes that scored higher on the independent variable than those described in other publication types. Hence, we could not confirm our initial hypothesis that cases described in journal articles tend to be more 'successful' than those described in other publication types.

These findings differ from previous studies examining the effects of publication bias in quantitative political analyses, which have clearly detected a publication bias effect. This may be due to the different nature of data sources and the inclusion of a broad range of cases in the case survey method,[16] based on the information they contain rather than the type of data. Case studies are often produced in the course of (small-N) comparative research, where case selection aims to contrast different experiences and follows various selection rules – for example, most similar, most different case design, extreme cases for the dependent or independent variable, deviant cases, stratifying and matching (see, for example, Reference 59.) We can assume, therefore, that the universe of published case studies contains a diverse array of experiences. Many of these may be considered 'success stories' with strong effects between the independent and dependent variables of interest, but many others may describe less clear-cut cases, more ambiguous outcomes, and weaker effects, mitigating publication bias.

This effect may be compounded by the heterogeneity of topics and issues covered in the case studies considered.[16] In contrast to quantitative meta-analyses, where the studies included follow the same research aims, we incorporated a wide variety of studies pursuing diverse research questions, often different from our own aims. Hence, 'success', or a positive effect of interest may be defined very differently for the different cases. For example, one publication may define 'success' as a fair decision-making process, others may define 'success' through social learning, or stakeholder acceptance, or – as in the case of our primary research question – benefits for the environment. Overall, therefore, the case survey method may be less prone to publication bias in its classical sense than other meta-analytical techniques.

However, our analysis of correlation effects between our main dependent and independent variables in cases reported in a mix of publications of different types, suggests that there could be an alternative mechanism of bias at play. Correlations were significantly higher within this group than in cases reported in grey or peer-reviewed literature. We can only speculate about the reasons for this effect, but cases in this 'mixed' category rest on around 2.5 times more texts than those in the other categories. This might be an indication for a mechanism of publication or attention bias, where 'successful' cases become overrepresented in the field because they figure prominently in multiples studies and become highly visible. This potential bias mechanism differs from the type of publication bias prevalent in quantitative studies, where inconclusive or non-significant results may go unpublished, as it entails not the exclusion of single publications, but disproportionate attention on specific, successful 'landmark' cases within a field. Hence, focusing mainly on these easily accessible, well-described cases may be a potential source of bias within a case survey meta-analysis. However, our evidence in this respect remains tentative and more research may be required to substantiate this hypothesis.

Finally, concerns regarding the quality of studies from grey vis-à-vis books or peer-reviewed literature and the reliability of information therein were not supported by our analyses.

4.2 | Rater 'personalities' and drift effects

It is good practice in studies relying on the judgement of multiple raters for the assessment and quantification

TABLE 2 Effects of rater 'personalities' and experience on case coding

Model	# of variables/%	# of information reliability scores/%
#0: Null model	44/24%	0/0%
#1: Rater	92/49%	25/13%
#2: Rater and rater experience	33/18%	25/13%
#3: Rater and rater experience (interaction)	17/9%	136/73%

Note: Summary of model comparison for the effects of rater personalities and coding experience on case survey variables ($n = 186$). Numbers indicate absolute and relative abundance of variables, for which the explanatory model performed best.

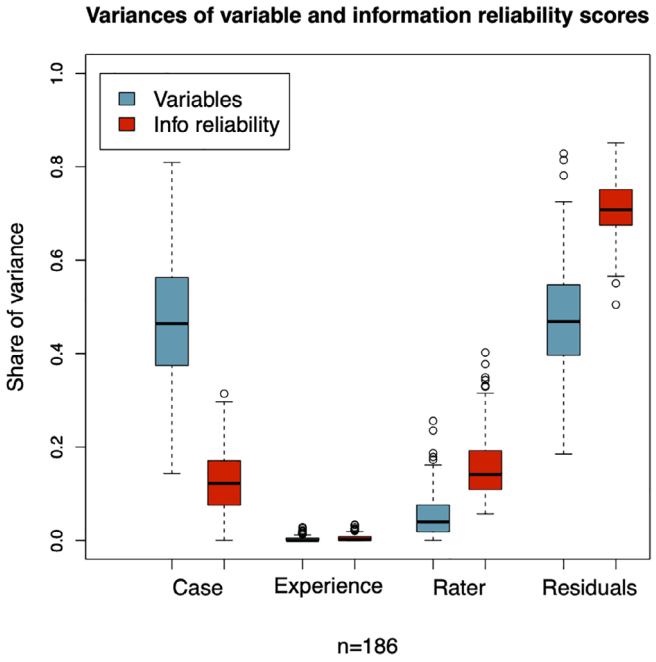

FIGURE 2 Share of variance explained by case, coding experience, raters, and residual variance. Aggregated display for 186 variables (blue) and their information reliability indicators (red) [Colour figure can be viewed at wileyonlinelibrary.com]

of empirical material[60] to calculate measures of inter-rater reliability and agreement in order to evaluate the quality of the data generated. We adopted the widely used measure of r_{WG}[61] as an inter-rater agreement index indicating the degree to which raters actually assign the same values.[62] Additionally, we assessed inter-rater reliability using the measure of $G(q, k)$.[63] This index is a special version of the more commonly employed intra class correlations (ICC) measures, explicitly adapted for situations where changing rater teams are assembled from a larger pool resulting in a design that is neither fully crossed (each rater codes all cases) nor fully nested (different raters for each case). Given the design and methods of the present study, this reliability measure is the most appropriate. Both mean reliability, and agreement across all variables score very similar, with $\text{mean}_{G(q,k)} = 0.71$ ($\text{SD}_{G(q,k)} = 0.11$) and $\text{mean}_{rWG} = 0.72$ ($\text{SD}_{rWG} = 0.07$). These indicators fulfil basic standards of data quality,[64] particularly considering that the case survey method explicitly allows for dissent and different interpretations of the empirical case material.

Moreover, we aim to more precisely understand the possible bias introduced by raters. We identified two factors that may potentially introduce bias during the coding process, namely the 'personality' of the rater, and the degree of coding experience acquired by the rater, which may lead to coder 'drift'. In our case survey, a total of 25 raters participated in coding, with experience ranging from 8 to 94 cases coded per person (median $_{raters} = 30$, $\text{SD}_{raters} = 25.3$). Although all raters received the same pre-coding training, we expect to see some sort of 'human factor' in case coding.

To assess the impact of raters' personalities and experience, we applied a multilevel modelling approach. Level 1, as above, comprises the rater-specific case instances (three per case), and level 2 is constituted by the cases as such. For all 186 variables, we fitted four different models, each representing one particular explanatory factor of interest[10]: The first model (model #0) represents the null-hypothesis, assuming no influence between our explanatory factors and the respective variable values. On the basis of this model, fixed factors were added – first, the rater (model #1) and then, additionally, the number of previously coded cases as a proxy for coding experience (model #2) and, finally, the interaction of these two factors (model #3) as an indicator for rater-specific experience. A comparison of these four models should reveal if raters and coding experience are better predictors than pure chance, and which factor shows the greatest effects for the respective variables. Model comparisons are based on Akaike's information criterion (AIC).

The results for 186 variables are presented in Table 2. Each variable is attributed to the model with lowest AIC value, that is, the model that has the highest relative explanatory value for the respective variable.

The results show that for 445 variables (24% of all variables) the null model has the strongest explanatory value – that is, raters and experience show no considerable effect here. For most variables, the models including rater personalities (and experience) fit best, while only 9% of our variables indicated rater-specific drift.

Beyond simply examining the number of variables subject to some rater effect, the relative size of this effect is of particular interest, and will reveal how much variable variance can be accounted for by raters and

experience. To investigate this question, we again fit a multilevel model for each variable. However, to compare the relative effects, we fit raters' personalities and experience, together with the case ID, as random factors, and then compared the amount of variance attributed to each of the factors.[11] The aim of this procedure is to assess the relative weight of rater personality and experience in relation to the importance of the case itself, which we actually aim to measure.

The result of this analysis is displayed in Figure 2 (blue boxes), summarizing standardized values for each of our 186 variables in an aggregated way. As the boxplots show, most of the variable variance ($mean_{case} = 0.47$, $SD_{case} = 0.13$) can be attributed to the actual case ID – that is, the case characteristics we sought to actually measure. Raters ($mean_{rater} = 0.05$, $SD_{rater} = 0.05$) and experience ($mean_{experience} < 0.01$, $SD_{experience} = 0.01$) account for only a much smaller share of variance, with a fairly large remainder of residual variance ($mean_{residual} = 0.47$, $SD_{residual} = 0.11$).

We repeated this analysis of the influence of rater and experience for the information reliability codes. These codes accompany each variable value assigned by a rater, and give an indication of the information basis for the given variable value. In analyzing this data, we followed the same approach as above, using a model comparison of four different models to assess the effects of rater personalities, time effects and their interaction. As Figure 2 shows, results differ considerably from those for the actual variables. None of the information reliability scores is best reflected by the null model, while 73% show rater-specific experience effects. The rest exhibit single rater (13%) and experience effects (13%). Comparing the extent of variance attributed to each single factor (Figure 2, red boxes), it becomes clear that rater personalities account for slightly more variance ($mean_{rater} = 0.16$, $SD_{rater} = 0.07$) than the case as such ($mean_{case} = 0.13$, $SD_{case} = 0.07$). Most variance, however, is accounted for by the residuals ($mean_{residual} = 0.71$, $SD_{residual} = 0.06$), which also include the rater-experience interaction effects (i.e., individual drift effects). Unfortunately, it was not possible for us to calculate the actual influence of these drift effects by means of post hoc tests, due to limited sample sizes and insufficient statistical power in our dataset. For illustration, a graphical representation of the individual learning effects is included in the supplementary material.

Overall, our analysis of rater and drift effects indicated high data quality. Rater personalities accounted on average for less than 6% of the variable variance and codings appeared to be stable over time. We can assume from these results that raters arrived at a rather coherent understanding of concepts, cases, and the coding procedure. Thorough preparation and training of raters may have played a decisive role in this.[65] Raters all went through the same training and coding of test cases, which spanned a wide range of possible case contexts and trajectories. Therefore, all raters began from a similar starting point, with a good understanding of the relevant social science concepts and of how to translate case narratives into numeric data. Further, this common understanding was regularly reinforced through interaction and exchange with other raters during post-coding discussions, which thereby served as a kind of internal peer review or quality check for each case.

Stronger bias effects were discovered for the coding of information reliability codes. These were shown to be rather strongly influenced by the personality and drift of raters. This does not have an immediate effect on the quality of the data itself, but brings into question the use of this measure to gauge information quality in the case texts. Information reliability can mainly be employed during data analysis to build subsets of cases with strongly reliable information, or to aggregate coding instances by different raters for the same case using a reliability weighted mean.[39]

4.3 | Variable characteristics

The characteristics of the variables themselves may pose specific challenges to the coding process and the reliability and validity of the resulting case survey data. Raters may face situations which require high-inference judgement, as variables aim to elicit information that is not readily available in the underlying primary case material or that may be subject to individual interpretation. These challenges may appear in various forms and constellations and, thus, may be associated with different aspects of data validity and reliability. To identify such different constellations, we undertook a cluster analysis to identify different types of variables with particular characteristics.

The cluster analysis started from 230 case survey variables. We extended the previous data set of 186 variables (which were limited by type of measurement scale), as here we do not consider differences in measurement scales as potential distortions, but rather embrace these differences as potential factors to explore the characteristics of variables. For each variable, we considered five important characteristics reflecting their measurement complexity and uncertainty as well as some measures of data quality, as detailed in Table 3: Inter-rater reliability, inter-rater agreement, information reliability, the degree of interpretative judgement required for coding, and the rater variance as measurement for the added value of the third rater over just two raters.

TABLE 3 Description of factors for cluster analysis

Factor	Indicator/description	Measurement scale	Mean	SD
Inter-rater reliability	$G(q, k)$	[0; 1]	0.75	0.13
Inter-rater agreement	r_{WG}	[0; 1]	0.76	0.11
Information reliability	Mean information reliability, aggregated over 230 variables	[0; 3]	1.57	0.23
Interpretative judgement	Degree to which a variable relies on subjective interpretation of case material, beyond basic given facts. To construct this index, five raters independently assessed all 230 variables as to their 'subjectivity', of which we then calculated the mean for each variable	[0; 4]	2.07	0.78
Rater variance	A proxy for the added value of the third rater over only two raters. This measure is derived from a simulation (1000 times) of randomly choosing only two raters per case (rather than three), and then computing inter-rater reliability for this reduced data set. To arrive at a proxy for variable sensitivity to rater elimination we calculated the mean standard deviation of the simulated inter-rater reliability values	[0; 1]	0.03	0.02

TABLE 4 Strength of the presence of each factor in the different clusters, measured by cluster-specific means and indicator values (in brackets) of factors included in the cluster analysis

	Cluster 1 ($n = 47$)	Cluster 2 ($n = 68$)	Cluster 3 ($n = 115$)
Inter-rater reliability ($G(q, k)$)	0.89 (0.39)	0.63 (0.28)	0.76 (0.33)
Inter-rater agreement (r_{WG})	0.90 (0.38)	0.77 (0.33)	0.69 (0.29)
Information reliability	1.75 (0.37)	1.39 (0.29)	1.59 (0.34)
Interpretative judgement	1.08 (0.19)	2.27 (0.40)	2.35 (0.41)
Rater variance	0.014 (0.16)	0.050 (0.58)	0.023 (0.27)
Cluster characteristics	• Often on binary or 3-point scales (62% of variables) • Elicit readily available information	• Measured on [0, 4] or [−4, 4] scale, only 7% on simpler scale • Variables assess difficult social and technical concepts, or specific, detailed information • Rarely explicitly described within case, circumstantial evidence	• All variables measured on [0, 4] or [−4, 4] scale • Require highest level of abstraction on part of the raters • Elicit information central to research interest, hence, good information basis (controlled for during case screening)
Example variables from our data base on participatory environmental governance	• Location of the case • Case initiators • Checklist on content of the output	• Social capital (e.g., shared norms) • Cultural values (e.g., 'green culture') • Information management during process • External transparency	• Interests and intentions of actors • Major process characteristics (e.g., communication flows) • Assessment of environmental and social outputs

We conducted an agglomerative hierarchical cluster analysis using Euclidean distance measures.[12] We used Ward's method of agglomeration as it supports our goal of arriving at homogenous groups and tends to produce readily interpretable and widely understood results. To facilitate interpretation and make characteristics of the resulting clusters more transparent, we calculated group-specific means as well as indicator values for each cluster see Reference 66, both highlighting the association of each indicator to a given cluster.

Based on the analysis, we identified three different clusters, the characteristics of which are summarized in

Table 4.[13] Variables in cluster 1 ($n = 47$ or 20% of all variables) exhibit particularly high values for inter-rater agreement and reliability, high quality of information given in case descriptions, and require only minimal judgement on the part of raters. By contrast, variables in cluster 2 ($n = 68/30\%$) score especially low for inter-rater reliability, and have the lowest information reliability while at the same time requiring a high degree of interpretative judgement by raters around the concepts underpinning these variables. Here, the added value of a third rater can be considered greatest, as reflected in our rater variance index being more than three times higher than for cluster 1. Cluster 3 comprises 115 variables (50%), which exhibit moderately high inter-rater reliability and lower agreement values, based on rather good information quality. Also, while the interpretative judgment required to code these variables is high, the added value of a third rater is rather low, less than half in comparison to cluster 2.

Following this analysis, we can identify three structural challenges to high-inference coding, in different combinations co-occurring with low values for inter-rater reliability and agreement: The reliability of information given in the case descriptions; the intricacy of concepts that are measured by the variables; and the measurement scale.

The less complex variables in cluster 1 (20% of all variables) can be considered low-inferential judgements, and also exhibit the highest inter-rater reliability and agreement scores. They elicit basic, readily available, information about decision-making processes and process outputs. In many cases, the variables in this cluster rely on binary or other simple measurement scales.

The remaining two clusters all display more complex variables that pose some challenges for coding. Inter-rater reliability is lowest for cluster 2, while agreement scores reasonably high. This cluster contains variables that operationalize relatively difficult social or technical concepts (e.g., social capital, cultural values, potential spillovers), or those that target very specific, detailed information about the process, the external context and impacts of the case at hand (e.g., about information elicitation and aggregation, external transparency of a decision-making process, broader public attention). Such details are rarely described explicitly in the case descriptions, as the low information reliability value also highlights. Under these conditions, the contribution of a third rater contributes greatly to stabilizing inter-rater reliability. As such, the 68 variables in this cluster may be considered the highest-inference for coding.

In cluster 3, on the other hand, reliability of information in the case texts is rather high, while the coding of variables and the concepts they measure requires even higher degrees of interpretative judgement as compared to cluster 2. Many of these 115 variables are central to our research interest. They seek to capture detailed information about actors and political decision-making processes, and to make some assessment of outputs. Inter-rater reliability is relatively high for these variables, while agreement has the lowest value of all clusters. This cluster highlights that we can arrive at consistent and robust codes even in situations where raters are faced with coding difficulties. Inter-rater reliability even appears to be rather robust to the absence of a third rater. Comparing this situation to cluster 2 indicates that information reliability may be one important contributing factor. The lower inter-rater agreement may be due to the inter-subjective, interpretative nature of the case survey method, allowing for different readings of the case material. These varying interpretations are particularly prevalent where decisions had to be taken about the orientations and stance of the actors involved in decision-making processes, which make up almost half of the variables in this cluster. Nonetheless, raters remained consistent with their coding across cases, as the considerably high reliability scores indicate. This indicates that the interpretative characteristics of the case survey method do not necessarily work against data quality and reliability.

5 | LESSONS FOR DESIGN CHOICES IN CASE STUDY META-ANALYSIS

Data derived from our case survey proved largely robust, displaying high degrees of inter-rater reliability and agreement, and only limited effects of the distorting factors tested for. Based on these findings, we identify a number of critical design choices will influence the quality and reliability of case survey data.

1. *Trade-off between numerous and reliable cases*: Selecting cases for coding poses a trade-off between broad inclusiveness on the one hand, and stricter case selection with clear criteria for case features and information requirements on the other hand. The former can result in a large data set with very mixed data quality, as information reliability has shown to be one of the main mitigating factors for coding complexity; while the latter can produce a relatively small data set with good data quality but representing only a fraction of cases available. This trade-off is to a large extent determined by the research aims, data requirements, and the complexity of concepts to be measured, but there remains leeway to select a design appropriate to the research question at hand.

2. *Search strategies and publication types*: Limiting the study to peer-reviewed journal articles, books, and book chapters may be convenient for selecting publications for analysis, as these media are easy to locate in databases and through library searches. Indeed, for the main variables in our data set, we did not find significant differences as a result of the publication type that studies were reported in, indicating that – at least in our case – a good overview could be reached by these means. However, our study also suggested that there may be merit in employing broader search strategies and including a range of publication types. While the search effort might be rather high, cases without broad academic coverage, or those described in grey literature, may provide a more diverse and complete picture of the phenomenon under study.
3. *Design of variables*: The design of variables is an important determinant of the complexity of coding and the quality of resulting data. Much research is available on this topic (including from other methodological contexts, such as survey or questionnaire design).[67] For the case survey, the intricacy of the concept to be measured and the measurement scale are the main design choices. Our analysis has indicated that simple variable design with few response categories and straightforward coding tasks foster the quality and reliability of data. Yet, this strategy risks over-simplifying complex social phenomena to a point where it fails to grasp the diversity of cases and variables. In line with previous studies,[14,26,68] an iterative process in designing variables and a coding scheme is advisable. We began with a complex coding scheme with many variables and response categories, which we pre-tested in a variety of cases and with various raters. As we were familiar with some of the cases in our database, we intentionally chose cases for this pre-test that covered a wide diversity of potential situations; where this prior knowledge is not available, random selection of pilot cases is also possible. After assessing the reliability of data produced, and considering challenges encountered in coding, we simplified variables and coding scales to achieve a balance between reliability and maximum information gain. However, our analyses further highlight that challenges of more 'intricate' variable designs can be mitigated by sound case information basis (see 1), thorough training of raters, and by use of multiple raters.
4. *Selection and number of raters*: The coding of many cases by multiple raters is a resource-intensive process. Hence, the selection of raters is one major design choice in conducting a case survey. We worked mainly with student assistants that underwent a structured, comprehensive training routine, and employed three raters for the coding of each case. A subsequent post-coding discussion functioned as a quality check. While this constellation led to considerably high data quality, amid high resource costs other constellations are also possible. For instance, generally fewer raters could be used, or the number of raters could be adjusted depending on the complexity of variables, such that additional raters could be employed for only the most complex variables (as described in clusters 2 and 3), or for a selection of cases to establish a measurement of reliability see Reference 69. Alternative roles may be possible, such as an observer for post-coding discussions or a 'proof-reader' for coded cases, who has read and understood the cases, but may not have gone through the whole coding procedure herself. However, we strongly recommend having at least two raters, as this is the minimum number allowing for different interpretations and for the calculation of indices of inter-rater reliability and agreement, and ultimately for the evaluation of the resulting data.
5. *Avoiding rater drift over time*: Our analysis indicates significant rater-specific drift only for 6% of all variables (but also for the majority of reliability scores). Even if these effects are not a major threat to overall data quality, it is worthwhile reflecting on how to best avoid rater drift. First, we assume that the detailed and iteratively developed, tested, and refined codebook, as well as the intensive initial rater training, account for the small size of the observed drift effect. Further, post-coding discussions with ever-changing compositions of raters may have proven effective in balancing one-sided interpretations and counteracting drift effects. Literature further suggest spot checking throughout the coding process.[68] Finally, our initial training focused on coding variables, but less so on coding reliability. This might explain the far greater drift effect regarding reliability scores, because raters had to develop their own way of coding reliability. Here, more intensive training, or more specific guidelines, may help.

It should be noted that these results provide only preliminary insights from one case survey study. However, given the multi-faceted nature of the research objective (i.e., assessing the performance of participatory environmental governance), it does reflect many aspects of most research fields in policy sciences and governance and beyond, which are characterized by complex social phenomena and have few tried-and-tested measurement scales available. Our findings may also provide insights of value to other fields, such as health research and the analysis of the context, implementation and impact of medical interventions, which are increasingly advocated and demanded.[9] Despite our aspirations for comprehensiveness and rigor, more studies will be needed to

consolidate the methodological knowledge basis for this kind of synthesis research. However, given the ever-increasing abundance of empirical studies, we believe that a rigorous meta-analytical approach to case studies and text narratives provides a viable way forward, and one that can yield new and innovative insights.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

ORCID

Nicolas W. Jager https://orcid.org/0000-0002-6706-6595

ENDNOTES

[1] EDGE' (Evaluating the Delivery of Environmental Governance using an Evidence-based Research Design), ERC Starting Grant awarded to JN. Project timeframe 2011-2016.

[2] See supplementary material for detailed variable descriptions.

[3] *Information reliability* captures the degree of information available to raters to render their coding judgment. This is not to be confused with *interrater reliability* discussed below.

[4] As listed in Scopus.

[5] See supplementary material 2 for more information on model specification.

[6] Packages used for the analyses in this article: nlme,[70] multcomp,[71] ggplot 2,[72] ltm,[73] lme4,[74] multilevel,[75] cowplot,[76] cluster,[77] labdsv,[78] NbClust.[79]

[7] Model details can be found in the online supplementary material.

[8] Spearman's correlation coefficient, all estimates are significant on the $p < 0.001$ level.

[9] Publication type: $\chi^2(2) = 2.49$, $p = 0.29$; Word count: $\chi^2(1) = 0.17$, $p = 0.68$.

[10] For more information on model specification, please consult the supplementary material.

[11] For details on model specification see supplementary material. Some of the models of this analysis displayed singular fit for the variable *Experience*. The random effect of this variable was 0, or almost 0. While all models converged properly, this could be a sign of underestimating the effect of this variable. However, robustness tests, where all analyses were replicated without this variable, yielded almost identical effects, indicating robustness of results.

[12] Variables have been standardized for the analysis.

[13] For a dendrogram with a detailed variable assignment, see supplementary material.

REFERENCES

1. Ostrom E, Janssen MA, Anderies JM. Going beyond panaceas. *Proc Natl Acad Sci U S A*. 2007;104(39):15176-15178.
2. Ryan M. Using cases in political science: reflections on Keith Dowding's the philosophy and methods of political science. *Polit Stud Rev*. 2017;15(2):194-200.
3. Jensen JL, Rodgers R. Cumulating the intellectual gold of case study research. *Public Adm Rev*. 2001;61(2):235-246.
4. Newig J, Rose M. Cumulating evidence in environmental governance, policy and planning research: towards a research reform agenda. *J Environ Policy Plan*. 2020;22(5):667-681.
5. van der Jagt APN, Raven R, Dorst H, Runhaar H. Nature-based innovation systems. *Environ Innov Soc Transit*. 2020;35:202-216.
6. Alexander SM, Jones K, Bennett NJ, et al. Qualitative data sharing and synthesis for sustainability science. *Nat Sustain*. 2020;3(2):81-88.
7. Lorenc T, Pearson M, Jamal F, Cooper C, Garside R. The role of systematic reviews of qualitative evidence in evaluating interventions: a case study. *Res Synth Methods*. 2012;3(1):1-10.
8. Herron MC, Quinn KM. A careful look at modern case selection methods. *Sociol Methods Res*. 2016;45(3):458-492.
9. Thompson Coon J, Gwernan-Jones R, Garside R, et al. Developing methods for the overarching synthesis of quantitative and qualitative evidence: the interweave synthesis approach. *Res Synth Methods*. 2020;11(4):507-521.
10. Woodside AG. *Case Study Research: Core Skill Sets in Using 15 Genres*. Emerald; 2017.
11. Elman C, Gerring J, Mahoney J. Case study research: putting the quant into the Qual. *Sociol Methods Res*. 2016;45(3):375-391.
12. Lucas WA. *The Case Survey Method: Aggregating Case Experience*. The Rand Corporation; 1974.
13. Yin RK, Heald KA. Using the case survey method to analyze policy studies. *Adm Sci Q*. 1975;20:371-382.
14. Newig J, Fritsch O. The Case Survey Method and Applications in Political Science. 2009.
15. Cook TD. Generalizing causal knowledge in the policy sciences: external validity as a task of both multi-attribute representation and multi-attribute extrapolation. *J Policy Anal Manag*. 2014;33(2):527-536.
16. Beierle TC, Cayford J. *Democracy in Practice: Public Participation in Environmental Decisions*. Resources for the Future; 2002.
17. Newig J, Fritsch O. More input – better output: does citizen involvement improve environmental governance? Blühdorn I, ed. *J Environ Policy Gov*. 2009;19(3):205-224.
18. Geißel B, Heß P. Determinants of successful participatory governance: The case of Local Agenda 21. In: Heinelt H, ed. *Handbook on Participatory Governance*. . Edward Elgar; 2018:246-266.
19. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: unlocking the file drawer. *Science*. 2014;345(6203):1502-1505.
20. Maestas CD, Buttice MK, Stone WJ. Extracting wisdom from experts and small crowds: strategies for improving informant-based measures of political concepts. *Polit Anal*. 2014;22(3):354-373.
21. Newig J, Challies E, Jager NW, Kochskämper E, Adzersen A. The environmental performance of participatory and collaborative governance: a framework of causal mechanisms. *Policy Stud J*. 2018;46(2):269-297.
22. Newig J, Adzersen A, Challies E, Fritsch O, Jager N. Comparative Analysis of Public Environmental Decision-Making

Processes – a Variable-Based Analytical Scheme. *Leuphana University Lüneburg*; 2013.
23. Lipsey MW, Wilson DB. *Practical Meta -Analysis*. Sage; 2001.
24. Cooper H. *Research Synthesis and Meta-Analysis. A Step-by-Step Approach*. Vol 4. Sage; 2010.
25. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res*. 1976;5(10):3-8.
26. Larsson R. Case survey methodology: quantitative analysis of patterns across case studies. *Acad Manag J*. 1993;36(6):1515-1546.
27. Haddaway NR, Bethel A, Dicks LV, et al. Eight problems with literature reviews and how to fix them. *Nat Ecol Evol*. 2020;4 (12):1582-1589.
28. Bullock RJ, Tubbs ME. The case meta-analysis method for OD. *Res Organ Chang Dev*. 1987;1:171-228.
29. Rothstein HR, Sutton A, Borenstein M, eds. *Publication Bias in Meta-Analysis. Prevention, Assessment and Adjustments*. Wiley; 2005.
30. Gerber AS, Green DP, Nickerson D. Testing for publication bias in political science. *Polit Anal*. 2001;9:385-392.
31. Marks-Anglin A, Chen Y. A historical review of publication bias. *Res Synth Methods*. 2020;11(6):725-742.
32. Banks GC, Kepes S, McDaniel M. Publication bias. Understanding the myths concerning threats to the advancement of science. In: Lance CE, Vandenberg RJ, eds. *More Statistical and Methodological Myths and Urban Legends*. Routledge; 2015:36-64.
33. Mahood Q, Van Eerd D, Irvin E. Searching for grey literature for systematic reviews: challenges and benefits. *Res Synth Methods*. 2014;5(3):221-234.
34. Hopewell S, Clarke M, Mallett S. Grey literature and systematic reviews. In: Rothstein HR, Sutton AJ, Borenstein M, eds. *Publication Bias in Meta-Analysis, Prevention, Assessment and Adjustments*. Wiley; 2005:49-72.
35. Haddaway NR, Bayliss HR. Shades of grey: two forms of grey literature important for reviews in conservation. *Biol Conserv*. 2015;191:827-829.
36. Krippendorff K. *Content Analysis. An Introduction to Its Methodology*. Sage; 2013.
37. Armstrong D, Gosling A, Weinman J, Marteau T. The place of inter-rater reliability in qualitative research: an empirical study. *Sociology*. 1997;31(3):597-606.
38. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. *Sociol Methods Res*. 2021;50(2):837-865.
39. Van Bruggen GH, Lilien GL, Kacker M. Informants in organizational marketing research: why use multiple informants and how to aggregate responses. *J Market Res*. 2002;39(4):469-478.
40. Wilson DB. Systematic coding. In: Cooper H, Hedges LV, Valentine JC, eds. *The Handbook of Research Synthesis and Meta-Analysis*. Sage; 2009:159-176.
41. Miller N, Lee J-Y, Carlson M. The validity of inferential judgements when used in theory-testing meta-analysis. *Pers Soc Psychol Bull*. 1991;17(3):335-343.
42. Jager NW, Newig J, Challies E, Kochskämper E. Pathways to Implementation: Evidence on How Participation in Environmental Governance Impacts on Environmental Outcomes. *J Public Adm Res Theory*. 2020;30(3):383-399.
43. Newig J, Jager NW, Kochskämper E, Challies E. Learning in participatory environmental governance – its antecedents and effects. Findings from a case survey meta-analysis. *J Environ Policy Plan*. 2019;21(3):213-227.
44. Libby R, Blashfield RK. Performance of a composite as a function of the number of judges. *Organ Behav Hum Perform*. 1978; 21(2):121-129.
45. Kumar N, Stern LW, Anderson JC. Conducting Interorganizational research using key informants. *Acad Manag J*. 1993;36(6):1633-1651.
46. Renn O. Partizipation – ein schillernder Begriff. *Gaia*. 2005;14 (3):227-228.
47. Cox M. Understanding large social-ecological systems: introducing the SESMAD project. *Int J Commons*. 2014;8(2):265-276.
48. Scott TA. Does collaboration make any difference? Linking collaborative governance to environmental outcomes. *J Policy Anal Manag*. 2015;34(3):537-566.
49. Ulibarri N. Tracing process to performance of collaborative governance: a comparative case study of Federal Hydropower Licensing. *Policy Stud J*. 2015;43(2):283-308.
50. Collaboration for Environmental Evidence. *Guidelines for Systematic Review and Evidence Synthesis in Environmental Management*. Environmental Evidence; 2013. http://www.environmentalevidence.org/Documents/Guidelines.
51. McKay PF, McDaniel MA. A reexamination of black-white mean differences in work performance: more data, more moderators. *J Appl Psychol*. 2006;91(3):538-554.
52. Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *Am Psychol*. 1993;48(12):1181-1209.
53. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess*. 2003;7(1):1-76.
54. Banks GC, Kepes S, Banks KP. Publication bias: the antagonist of meta-analytic reviews and effective policymaking. *Educ Eval Policy Anal*. 2012;34(3):259-277.
55. Aguinis H, Pierce CA, Bosco FA, Dalton DR, Dalton CM. Debunking myths and urban legends about meta-analysis. *Organ Res Methods*. 2011;14(2):306-331.
56. Kreft I, de Leeuw J. *Introducing Multilevel Modelling*. Sage; 1998.
57. Gerber AS, Malhotra N, Dowling CM, Doherty D. Publication bias in two political behavior literatures. *Am Polit Res*. 2010;38 (4):591-613.
58. Steiger JH. Tests for comparing elements of a correlation matrix. *Psychol Bull*. 1980;87(2):245-251.
59. Seawright J, Gerring J. Case selection techniques in a menu of qualitative and quantitative options. *Polit Res Q*. 1975;2008:294-308.
60. Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. *Res Synth Methods*. 2019;10(4):539-545.
61. James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. *J Appl Psychol*. 1984;69(1):85-98.
62. Wagner SM, Rau C, Lindemann E. Multiple informant methodology: a critical review and recommendations. *Sociol Methods Res*. 2010;38(4):582-618.
63. Putka DJ, Le H, McCloy RA, Diaz T. Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *J Appl Psychol*. 2008;93(5): 959-981.

64. Lebreton JM, Burgess JRD, Kaiser RB, Atchley EK, James LR. The restriction of variance hypothesis and interrater reliability and agreement: are ratings from multiple sources really dissimilar? *Organ Res Methods*. 2003;6(1):80-128.
65. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Stat Med*. 2002;21(22):3431-3446.
66. Dufrêne M, Legendre P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol Monogr*. 1997;67(3):345-366.
67. Pasek J, Krosnick JA. Optimizing survey questionnaire design in political science: insights from psychology. In: Leighley JE, ed. *The Oxford Handbook of American Elections and Political Behavior*. Oxford University Press; 2010:27-50.
68. Ratajczyk E, Brady U, Baggio JA, et al. Challenges and opportunities in coding the commons: problems, procedures, and potential solutions in large-N comparative case studies. *Int J Commons*. 2016;10(2):440-466.
69. Velten S, Jager NW, Newig J. Success of collaboration for sustainable agriculture: a case study meta-analysis. *Environ Dev Sustain*. 2021.
70. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-152; 2020.
71. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biom J*. 2008;50(3):346-363.
72. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2016.
73. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. *J Stat Softw*. 2006;17(5):1-25.
74. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48.
75. Bliese P. Multilevel: multilevel functions. R Package Version 2.6; 2016.
76. Wilke C. cowplot: streamlined plot theme and plot annotations for "ggplot2." R package version 1.1.1; 2020.
77. Mächler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster: cluster analysis basics and extensions. R package version 2.1.2; 2021.
78. Roberts DW. Labdsv: Ordination and multivariate analysis for ecology. R package version 2.0-1; 2019.
79. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Softw*. 2014;61(6):1-36.

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: Jager NW, Newig J, Challies E, Kochskämper E, von Wehrden H. Case study meta-analysis in the social sciences. Insights on data quality and reliability from a large-N case survey. *Res Syn Meth*. 2022;13(1):12-27. https://doi.org/10.1002/jrsm.1514