Tanja Burgard

Dem Fachbereich IV der Universität Trier zur Erlangung des akademischen Grades Doktorin der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) eingereichte Dissertation

Cumulative Meta-Analysis. Robustness of Evidence in Survey Methodology

<u>Betreuer:</u>

Prof. Dr. Johannes Kopp

Universität Trier – Fachbereich IV, Soziologie

Prof. Dr. Michael Bošnjak

ZPID – Leibniz-Institut für Psychologie

Universität Trier – Fachbereich I, Psychologie

Datum des Antrages auf Eröffnung des Promotionsverfahrens: März 2021

Danksagung

Die vorliegende Dissertation wurde im Juni 2018 vom Fachbereich IV der Universität Trier angenommen und ist maßgeblich am ZPID – Leibniz Institut für Psychologie entstanden.

Mein besonderer Dank gilt zunächst meinem Doktorvater Prof. Dr. Kopp für die spontane Übernahme der Betreuung meiner Dissertation. Für die enge Zusammenarbeit und die vielfältige Unterstützung und Förderung danke ich außerdem meinem Zweitgutachter, Herrn Prof. Dr. Bosnjak.

Darüber hinaus habe ich am ZPID von zahlreichen Kollegen konstruktive Anregungen und wertvolle Hilfestellungen erhalten. Dr. Nadine Wedderhoff hat mich vor Allem zu Beginn meiner Promotionsphase methodisch beraten und mir fachlich mit Feedback zu meinen ersten Manuskripten und Konferenzbeiträgen zur Seite gestanden. Lisa Trierweiler hat sowohl meine englischsprachigen Manuskripte, als auch den Manteltext, als Lektorin mit großer Präzision sprachlich und formal geprüft.

Auch allen anderen Kollegen am ZPID möchte ich an dieser Stelle für die stets konstruktive Zusammenarbeit und das angenehme Arbeitsklima danken. Als Mutter von zwei Kindern bin ich dem ZPID außerdem für die hervorragenden Arbeitsbedingungen dankbar. Insbesondere in Zeiten von Lockdown und Homeschooling waren die zeitliche und räumliche Flexibilität notwendig für die Fortführung meiner Arbeit. Auch der IT danke ich an dieser Stelle für den schnellen und reibungslosen Übergang ins Homeoffice zu Beginn der Pandemie.

Außerdem danke ich meinem ganz besonderen Kollegen und Lieblingsmenschen Robert, der mit mir nicht nur PsychOpen CAMA entwickelt, sondern mir als Partner und Verbündeter in allen Lebensbereichen zur Seite steht.

Schließlich möchte ich diese Arbeit meinen Kindern Hannah und Xesco widmen, die Allem, was ich tue, einen tieferen Sinn geben.

Contents

Abstract

Surveys play a major role in studying social and behavioral phenomena that are difficult to observe. Survey data provide insights into the determinants and consequences of human behavior and social interactions. Many domains rely on high quality survey data for decision making and policy implementation including politics, health, business, and the social sciences. Given a certain research question in a specific context, finding the most appropriate survey design to ensure data quality and keep fieldwork costs low at the same time is a difficult task. The aim of examining survey research methodology is to provide the best evidence to estimate the costs and errors of different survey design options. The goal of this thesis is to support and optimize the accumulation and sustainable use of evidence in survey methodology in four steps:

(1) Identifying the gaps in meta-analytic evidence in survey methodology by a systematic review of the existing evidence along the dimensions of a central framework in the field
(2) Filling in these gaps with two meta-analyses in the field of survey methodology, one on response rates in psychological online surveys, the other on panel conditioning effects for sensitive items
(3) Assessing the robustness and sufficiency of the results of the two meta-analyses
(4) Proposing a publication format for the accumulation and dissemination of meta-analytic evidence

Keywords:

Survey Methodology, Total Survey Error, Meta-Analyses, Knowledge Accumulation, Publication Format

Zusammenfassung

Umfragen spielen eine wichtige Rolle bei der Untersuchung sozialer und verhaltensbezogener Phänomene, die sonst nur schwer zu beobachten sind. Umfragedaten geben Einblicke in die Determinanten und Folgen sozialer Interaktionen und menschlichen Verhaltens. Daher ist die Qualität von Umfragedaten für viele Bereiche und Entscheidungssituationen von hoher Relevanz, unter anderem in der Politik, dem Gesundheitsweisen, der Wirtschaft, sowie der Wissenschaft. Angesichts einer bestimmten Forschungsfrage in einem spezifischen Kontext ist es eine schwierige Aufgabe, das am besten geeignete Erhebungsdesign zu identifizieren, um hohe Datenqualität bei gleichzeitig niedrigen Erhebungskosten zu gewährleisten. Ziel der Untersuchung von Umfragemethoden ist es, die beste Evidenz zur Abschätzung der Kosten und Fehler verschiedener Umfragedesign-Optionen zu liefern. Das Ziel dieser Arbeit ist es, die Akkumulation und nachhaltige Nutzung von Evidenz in der Umfragemethodik in vier Schritten zu unterstützen und zu optimieren:

(1) Die Identifikation von Lücken in der meta-analytischen Evidenz durch eine systematische Aufarbeitung der vorhandenen Evidenz entlang zentraler Dimensionen in diesem Bereich
(2) Beginn der Schließung dieser Lücken mit zwei Metaanalysen im Bereich der Erhebungsmethodik, eine zur Teilnahmebereitschaft in psychologischen Online-Umfragen, die andere zu Panel-Konditionierungseffekten bei sensiblen Items
(3) Untersuchung der Robustheit und Suffizienz der Ergebnisse beider Meta-Analysen
(4) Vorstellung eines Publikationsformates, um meta-analytische Evidenz zu sammeln und verfügbar zu machen

Schlüsselwörter:

Erhebungsmethodik, Gesamterhebungsfehler, Meta-Analysen, Wissensakkumulation, Publikationsformat

List of Figures

List of Tables

List of Publications Included in this Cumulative Thesis

Study 1:

Burgard, T., Bošnjak, M., & Wedderhoff, N. (2020). Response rates in online surveys with affective disorder participants. A meta-analysis of study design and time effects between 2008 and 2019. *Zeitschrift für Psychologie, 228*, 14-24. https://doi.org/10.1027/2151-2604/a000394

Study 2:

Burgard, T., Wedderhoff, N., & Bošnjak, M. (2020). Konditionierungseffekte in Panel-Untersuchungen: Systematische Übersichtsarbeit und Meta-Analyse am Beispiel sensitiver Fragen [Conditioning effects in panel studies: Systematic review and meta-analysis focusing on the example of sensitive questions]. *Psychologische Rundschau, 71*, 89-95. https://doi.org/10.1026/0033-3042/a000479

Study 3:

Burgard, T., Bosnjak, M., & Studtrucker, R. (2021). Community-augmented meta-analyses (CAMAs) in psychology: Potentials and current systems. *Zeitschrift für Psychologie, 229*(1), 15-23. https://doi.org/10.1027/2151-2604 /a000431

Study 4:

Burgard, T., Bosnjak, M., & Studtrucker, R. (under review). PsychOpen CAMA: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods, xx.*

1. Need of Evidence on Determinants of Survey Quality

Surveys play a major role in the social and behavioral sciences. They are often used to investigate phenomena, such as attitudes, well-being, personal values, and emotions, that are difficult to observe otherwise. Survey data reveals insights into human behavior by directly asking people for information (Ponto, 2015). Therefore, surveys serve to study reasons and consequences of human behavior and social interactions. For example, a survey analyst may try to explain differences or changes in voting behavior (Coffé & van den Berg, 2017) or the influence of relationship characteristics and social mechanisms on the probability of separation or divorce (Kopp & Richter, 2016).

Understanding individual decision-making processes and causal determinants of human behavior may guide public policy (Fritz & Koch, 2019) and marketing decisions (Kumar, Bezawada, Rishika, Janakiraman, & Kannan, 2016), as well as the use of psychological interventions (Trudel-Fitzgerald, Millstein, von Hippel et al., 2019) and communication efforts (Graham, Hasking, Clarke, & Meadows, 2015) to improve health behavior and well-being. Therefore, high quality survey data is crucial to provide reliable information for scientific purposes as well as for political, social, and health decisions.

However, survey fieldwork produces costs. The survey budget has to be used efficiently. Quality enhancement is usually accompanied by increased costs (Roberts, Vandenplas, & Stähli, 2014). This cost–error trade-off has to be solved by finding the best available combination of survey design options such as the survey mode, provision of incentives, or number of contact attempts. For example, paying an incentive to potential participants may reduce the number of contact attempts and increase the response rate (Bricker, 2014). Overall, the incentive could thus even reduce the costs per interview. To estimate the costs and errors different survey design options may cause, research on survey methodology is necessary. A strong evidence base in this field enables clear recommendations and guidelines for conducting different types of surveys. A classic collection of recommendations for conducting surveys is Dillman, Smyth, and Christian’s (2014) tailored design method which integrates different survey modes.

Meta-analytic findings from randomized experimental research provide the strongest empirical evidence possible. They offer greater precision and validity than individual studies and allow causal inferences to be made. Therefore, decisions in many fields, such as the health sciences, education, or management, are ideally based on meta-analytic evidence

(Bošnjak, 2018). Decisions concerning survey operations should also be based on the best available evidence. Thus, evidence-based survey operations (EBSO) should rely on meta-analyses in the field of survey methodology for designing and conducting survey projects. However, in the field there remains a reluctance to systematically synthesize the evidence. Bošnjak (2017) identifies about 50 relevant meta-analyses up to March 2016. There are plenty of potential research questions in the field and decisions to be made concerning survey methodology. A systematic collection and synthesis of evidence in the field would therefore be crucial to ensure an efficient use of research resources and a high quality of survey research.

The role of meta-analyses for evidence-based decision-making is explained in the second chapter. The meta-analytic method and common criticisms of the method are presented. Especially, the focus of this presentation is on the continuous accumulation and synthesis of evidence. The idea of cumulative meta-analysis supports the ongoing validation and increasing robustness of meta-analytic evidence on a research question. Furthermore, changes of the subject matter over time can be examined and research programs to strategically close evidence gaps can be informed by cumulative meta-analysis. In this context, the first goal of the thesis is to identify and present approaches to examine how robustness can be assessed in the process of meta-analytically cumulating evidence, and to determine at which point the existing evidence is sufficient to provide robust results.

The second aim of this thesis is to identify gaps in meta-analytic evidence in survey methodology. The starting point, therefore, is a systematic review of the existing evidence in the field of survey methodology along the dimensions of the Total Survey Error. This framework differentiates several types of nonobservation and observation errors that can potentially bias results from survey research. It is presented in chapter 3.

The results uncovered for the methodological issues in chapter 2 serve as the foundation to meet the third goal of the thesis, that is, providing meta-analytic findings to start filling in the gaps in the field of survey methodology, presented in chapters 4 and 5. The first meta-analysis examines response rates in online psychology surveys. In the second meta-analysis, panel conditioning effects for sensitive items are analyzed. The results of both meta-analyses are discussed in the context of previous research and in terms of robustness of evidence using the methods introduced in chapter 2.

The final goal of the thesis is the proposition of a publication format to make evidence available and easily enable its accumulation in chapter 6. The presented publication format will enable cumulative research by making meta-analytic datasets accessible and facilitate the replication and updating of meta-analyses. Existing systems are presented, as well as a newly developed platform for meta-analyses in psychology and neighboring fields: PsychOpen CAMA. Finally, further research gaps in survey methodology are discussed, and recommendations for workflows and data management are given to improve knowledge production and use of existing evidence for decision making in survey operations.

2. Cumulative Meta-Analysis and Evidence-Based Decisions

2.1. The Meta-Analytic Method

Meta-analysis "is the analysis of results from multiple studies" (Card, 2012: 5). Whereas single studies are often underpowered due to small sample sizes, replicating and synthesizing results from multiple studies leads to more reliable results and higher statistical power. Furthermore, summarizing a research field facilitates its advancement by building on previous work. Important questions for conceptualizing new primary studies, such as information for power analyses, potential relevant moderator variables, or open research gaps, can be addressed through meta-analysis. Thus, research resources can be used more efficiently (Glöckner, Fiedler, & Renkewitz, 2018).

To make the results of different studies comparable for the computation of a meta-analytic summary effect, an effect size index is chosen. The kind of data used in the primary studies determines whether the adequate effect size is based on means, binary data, or correlations. The effect size parameter for a single study is a sample estimate for the true underlying effect size, and thus, differs from this true effect to some extent. The precision of the effect size estimate depends primarily on the sample size. More precise estimates yield more information and, thus, are assigned more weight in the calculation of the meta-analytic summary effect (Borenstein, Hedges, Higgins, & Rothstein, 2009).

There are two different assumptions on the underlying true effect size in a meta-analysis. Under the fixed-effects model, it is assumed that one true effect size is underlying all studies in the meta-analysis. This implies that all variance between the studies is due to the sampling error within each study (Borenstein et al., 2009). In the random-effects model, variance on two levels is assumed. First, the true effect in each study is a random draw from a distribution of true study effects and varies around the overall mean of all studies with the variance τ^2 (tau-squared), the variation between studies. Second, within each study, individuals are sampled from a population (Raudenbush, 2009). The estimated effect size therefore will vary from the true effect size of the study by the study-specific sampling variance v_i. The total variation of effect size estimates in a random-effects model thus consists of sampling variance within studies and true heterogeneity between studies (Pigott, 2012).

Consequently, if an infinite sample size is assumed in each study, under the fixed-effects model, the same effect size for all studies would be expected. In the random-effects model, the effect sizes would be expected to differ only due to true heterogeneity between the studies. In other words, the total variation of effect size estimates would reduce to τ^2. Heterogeneity between studies can be highly informative. Effect sizes can vary, for example, due to differences in the composition of the sample or characteristics of the intervention (Borenstein et al., 2009).

The diversity in study settings and characteristics can increase the external validity of meta-analyses compared to single studies (Shadish, Cook, & Campbell, 2002). Moreover, the differences in study characteristics provide the opportunity to study interrelations that could not be studied in the primary studies. For example, a meta-analysis of studies conducted in different geographic regions can test whether differences between the study outcomes can be explained with the geographic region, whereas each single study only reports on one region. These analyses are called moderator analyses in the terminology of meta-analyses, although the respective moderator variables are included in a meta-analytic regression model just like independent variables in conventional regression models (Cornesse & Bošnjak, 2018). However, Wood and Eagly (2009) note that conclusions from moderator analyses should be drawn with caution, as classification of studies according to study characteristics are often done post hoc, that means without an explicit theoretical framework. Furthermore, the relevance of moderator variables is questionable if they fail to adequately account for the variance between studies. The remaining unexplained heterogeneity rather calls for follow-up research.

2.2. Common Criticisms of the Meta-Analytic Method

One criticism of meta-analyses is often expressed in the statement that “apples and oranges” could be mixed. In the context of meta-analysis, this means that the results of studies with different outcome variables, different operationalizations and treatments, as well as different populations are combined and the meaning of the result is thus unclear (Sharpe, 1997). The apples and oranges criticism above all underlines the meaning of clearly defined concepts and research questions. The research question can be very narrow and only include studies that are very similar, for example, replications of the same type of experiment. It may also be useful to have a broader research question, such as the effectiveness of homework (Cooper, 2017). Investigating this effect would require integrating findings from different interventions and study designs. To ensure comparability, effect sizes are standardized on a common scale. To take into account differences between studies, relevant study characteristics can be used as moderators. Thus, subgroups of similar studies can also be considered separately (Viechtbauer, 2007).

Another common concern is the treatment of publication bias. It is argued that meta-analyses overestimate overall effect sizes, as studies with nonsignificant results often remain in the file drawer and are not published. If the results of these studies are not integrated, the sample of studies is biased and the validity of the meta-analytic results is threatened (Rothstein, Sutton, & Borenstein, 2005). Summarizing relatively few studies usually already results in high statistical significance and strong evidence of an effect. However, the number of studies that remain unpublished due to nonsignificant results remains unknown (Rosenthal, 1979).

But at this point, meta-analysis is not the problem, but the key to treatment. There are plenty of statistical methods to test and correct for publication bias in a meta-analytic dataset, from the funnel plot as a simple visual indication proposed by Light and Pillemer (1984) and extended to adjust for the missing studies by a data augmentation technique (Duval & Tweedie, 2000) to more current approaches that aim at correcting meta-analytic estimates by taking into account the distribution of the p-values (van Aert, Wicherts, & van Assen, 2016). Meta-analyses also allow the inclusion of unpublished studies. To find these studies, a thorough literature search is essential. This effort then enables researchers to investigate the actual differences between published and unpublished studies and thus provides insights into the extent of publication bias in a certain research area.

Since the study quality of primary studies is already questionable, the “garbage in, garbage out” argument assumes that the integration of these studies into a meta-analysis might not yield valid results (Jüni, Altman, & Egger, 2001). Again, this is not a question of the meta-analytic method, but a problem of the scientific validity and quality of primary studies. Meta-analysis has the potential to deal with this issue. Study quality can be coded as a study characteristic and can be used either as an inclusion criterion, by including only studies of a certain quality, or as a weighting factor, by giving more weight to studies of higher quality. Study quality can also serve as a moderator to investigate the influence of study quality on study outcomes (Cooper, 1998). However, up to now, reporting practices related to primary study quality remain poor in psychological meta-analyses (Hohn, Slaney, & Tafreshi, 2019).

To sum up, criticisms of meta-analyses refer primarily to validity. In particular, the external validity is higher than in primary studies due to the variation of study characteristics. This holds at least, if the research question is correctly defined, effect sizes are correctly calculated, study quality and design are considered, and the model is correctly specified. However, conducting meta-analyses opens up many sources of error and requires subjective decisions. Therefore, transparency is important to enable sensitivity analyses. Indeed, meta-analysis then even has the potential to address common problems in scientific practice, such as poor study quality or publication bias. It can even serve as a scientific quality control and provide more valid statements than individual studies. However, the continuous accumulation and updating of meta-analyses is crucial to keep results up-to-date. Therefore, the challenge of accumulating meta-analytic evidence over time will be the focus in the following chapter.

2.3. Accumulation of Evidence for Meta-Analysis

Replicability and Validity of Evidence

Scientific findings can be validated on different levels (Stanley, Carter, & Doucouliagos, 2018). Reproducibility means to produce exactly the same results with the same data and analyses. To validate findings at this level, accessibility of data and analysis code are sufficient. On the next level, we aim for replicability. When the same results and conclusions are obtained as those found in the original study using a new random sample and following the reported procedures, we can report a successful replication of results. If a finding is further independent of unmeasured factors in the original study, such as sample characteristics or country of study, it can be considered as generalizable.

Replicability is already a problem at the level of individual studies. Reasons for failed replications can be found in every phase of the research cycle. If the findings are already known, researchers may propose hypotheses after the fact (a phenomenon known as HARKing—hypothesizing after the results are known, Kerr, 1998), leading to significant results deriving from their dataset. Another questionable research practice used to obtain significant results from the data is p-hacking (Simonsohn, Nelson, & Simmons, 2014). Of 64 studies investigated, Banks, Rogelberg, Woznyj, Landis, and Rupp (2016) identified evidence of questionable research practices in 91%. Poor study quality and relevant differences in study design can also lead to different study results. Finally, published studies may not be representative of all studies, as significant results are published more often (Ioannidis, 2008).

Apart from the questionable validity, individual studies often suffer from low statistical power (Fraley & Vazire, 2014), so that it is unlikely to find effects—especially small ones—even if they actually exist. Tversky and Kahnemann (1971) even argue that it is a misguided intuition to expect that a finding from one small study is replicated by another small study. They call it the “belief in the law of small numbers”. By chance, studies will provide significant results from time to time, even if the effects do not really exist. However, these results are often not replicable.

Table 1: Selected Potential Sources of Error in a Meta-Analytic Process

Step of the meta-analysis according to Cooper (2017)	**Sources of error**	**Threats to validity**
Formulating the problem	Poorly defined constructs and relationships (e.g., Baer, Gu, Cavanagh, & Strauss, 2019)	Questionable construct validity of measures
Searching the literature	Studies found in the literature search may not correctly represent the relevant population of studies (e.g., Pietschnig, Voracek, & Formann, 2010; Kepes & McDaniel, 2015)	Publication bias
Gathering information	Incorrect information retrieval resulting in misrepresentations in the data (e.g., London, 2017)	Unreliability in coding (inter- and intrapersonal), biased effect size sampling (favoring one direction of findings)
Quality appraisal	Evaluation approach not exhaustive concerning study design characteristics, nonexplicit weighting scheme (e.g., Hohn, Slaney, & Tafreshi, 2019)	Biased exclusion of studies, biased weighting of studies
Synthesis methods and analysis	Nonweighting of effect sizes, unjustified model specifications, for example, in case of heterogeneous or dependent effect sizes (e.g., Voracek, Kossmeier, & Tran, 2019)	Biased estimates and inferences
Interpretation of results	Inaccurate treatment of missing data, no consideration of confounded moderator effects and heterogeneity in samples and study setting (e.g., Tran, Hofer, & Voracek, 2014)	Biased estimates, overgeneralization of findings to contexts not represented in the meta-analysis

In meta-analyses, we expect a higher validity of the results due to the stronger evidence base and the heterogeneity in study designs and samples captured with a meta-analysis (Borenstein et al., 2009, p. 9). However, due to the frequency of questionable research practices in individual studies, the risk of bias of these should be accounted for in the first place (Hohn, Slaney, & Tafreshi, 2019). Above this, when carrying out the meta-analysis itself, there are a number of subjective decisions and sources of error threatening the validity of a meta-analysis (Cooper, 2017, p. 318).

These potential errors and threats to validity at each step of a meta-analysis are presented in Table 1. Starting with problem formulation, as in primary studies, constructs and relationships may be poorly defined resulting in questionable validity of measures. The selection of relevant literature for a meta-analysis may be subject to publication bias if relevant sources of unpublished literature are not sufficiently considered. The coding process to extract information from selected literature may be biased if instructions in the coding manual are not clear or coders do not adhere to them correctly. When assessing the quality of primary studies, the evaluations may not be exhaustive and the weighting schemes used may also be biased. Analytic decisions, such as model specification or treatment of missing data, are at least subjective, and in the worst cases even inappropriate for the data at hand.

To sum up, replicability and validity of evidence can be threatened by diverse influencing factors, at the level of primary studies as well as at the level of meta-analysis. Therefore, continuous and strategic accumulation of evidence, as well as collaborative quality assessment can be a means to achieve more reliable research outputs without waste of research resources. That is the basic idea of continuously updating meta-analysis. It enables to address two requirements to improve the design of and inferences from empirical research. Firstly, by continuously cumulating the evidence on a research question, results become more robust over time and changes of a phenomenon across time can also be observed. Secondly, gaps in research can be detected timely. This enables a purposeful allocation of research resources to strategically close these gaps.

Updating Meta-Analyses

Static snapshot meta-analyses, especially in dynamic research fields on hot topics, may outdate quickly. Therefore, regular updates of meta-analyses are necessary. For example, Cochrane reviews should be updated every two years (Shojania et al., 2007) and Campbell reviews within five years (Lakens et al., 2016). Créquit et al. (2016) examined the proportion of available evidence on lung cancer not covered by systematic reviews between 2009 and 2015 with the finding that, in all cases, at least 40% of treatments were missing.

For systematic reviews, an update is defined as a new edition of a published review. It can include new data, new methods, or new analyses. An update is recommended if the topic is still relevant and new methods or new studies have emerged that could potentially change the findings of the original review (Garner et al., 2016).

Shojania et al. (2007) define signals of relevant evidence changes to warrant the update of reviews. These signals are changes in statistical significance, a relevant relative change in effect magnitude, new information on the clinical relevance of a review, or the emergence of new approaches not considered previously. For 100 reviews, the time between the publication and the occurrence of a signal for updating is measured and the median survival time of a meta-analysis in their analysis is 5.5 years. Within two years, almost one-fourth of the reviews was already outdated (Shojania et al., 2007). As the number of publications is continuously growing (Bastian, Glasziou, & Chalmers, 2010), we can expect survival times of reviews to become even shorter.

Reuse of existing meta-analytic data is crucial to keep pace with the continuous publication of new research results. Haddaway (2018) therefore advocates for open synthesis. Applying the principles of open science (Kraker, Leony, Reinhardt, & Beham, 2011) to meta-analyses allows verification of methods and conclusions for an increased reliability. Furthermore, data extracted for a meta-analysis can be reused for novel purposes and for an easier and faster process of updating meta-analyses.

Sufficiency and Stability in Cumulative Meta-Analyses

In contrast to updating of meta-analyses, where only a single pooling is performed, in cumulative meta-analyses, poolings are performed sequentially according to the publication year or to other covariates (Lau, Schmid, & Chalmers, 1995). This enables to study the evolution of evidence over time and allows conclusions on the sufficiency and stability of evidence (Mullen, Muellerleile, & Bryant, 2001).

A cumulative meta-analysis can be illustrated using a forest plot as in Figure 1. The studies in this cumulative forest plot are sorted chronologically. The meta-analytic estimation is conducted successively with each new study. Thus, the plot shows the evolution of the effect size magnitude and the stabilization of evidence over time. The data used is freely available in the metafor package (Viechtbauer, 2010). The original data is derived from a meta-analysis of Linde, Berner, Egger, & Mulrow (2005) on the effect of St. John’s wort for depression. Early studies examining the effect of the substance Hypericum perforatum (St.

John's wort) showed marked effects, but with large confidence intervals. However, the effectiveness of the substance to treat depression was never in doubt. Over the course of time, the results stabilized and decreased to a rather small, but still significant, effect.

A systematic review of cumulative meta-analyses in medicine (Clarke, Brice, & Chalmers, 2014) reports many illustrative examples of how meta-analyzing the evidence at an earlier point in time would have already provided meaningful results about medical interventions. These examples speak for the high relevance of cumulative research and timely synthesis of evidence to enable more informed decisions and at the same time a more efficient distribution of research funds and efforts.

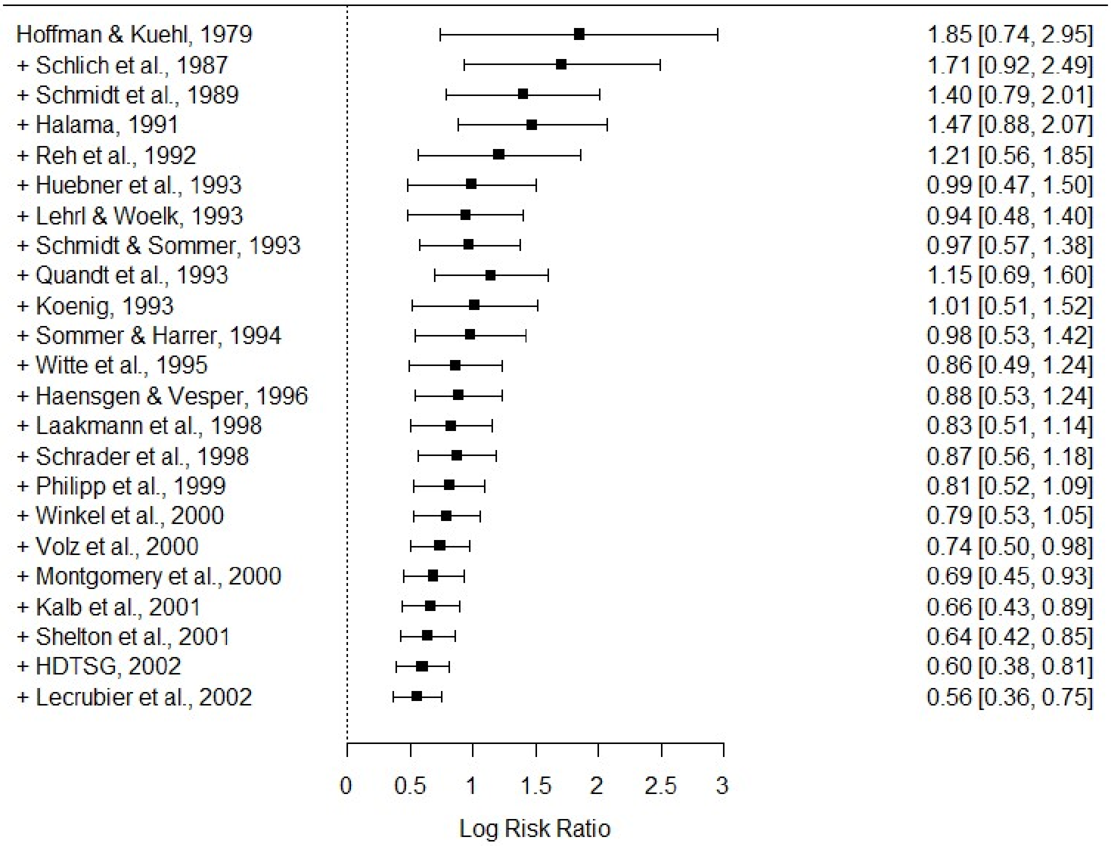


Figure 1: Cumulative Forest Plot on the Effect of St. John's Wort for Depression

The ongoing accumulation of evidence informs researchers about the latest findings in a specific research area. A meaningful concern is to determine when evidence is sufficient to no longer justify further research investment, at least without taking into account existing results and perhaps specific research gaps. At some point in time, evidence should allow to either conclude that there is no relevant effect to detect, or to draw conclusions on the direction of an effect or the efficiency of a treatment (Simmonds et al., 2017). Turner, Bird, & Higgins

(2013) examine the power of almost 15 000 meta-analyses from Cochrane reviews and conclude, that only 22% achieve the common threshold of 80% power to find a relative risk reduction of 30%. The power of the meta-analyses assessed is higher than the power of the corresponding included primary studies. However, heterogeneity between studies can also decrease precision, increasing the sample size required to reach sufficiently powered meta-analyses.

Next to providing sufficient information on a research question due to greater statistical power, meta-analyses can also reveal research gaps by giving an overview of potential moderators or moderator combinations not yet sufficiently studied. In the case of Zhu, Jiang, & Ding (2014), previous meta-analyses on the effect of substances for diabetes treatment on the risk of fractures had mainly focused on postmenopausal women. New evidence provided the opportunity to study gender as a potential moderator of the effect, and results revealed that increased risk of fractures was in fact only detected for women.

Finally, the research object could also be subject to temporal changes. Cumulative meta-analysis can then serve as a visual aid to detect trends over time in research findings. Leimu & Koricheva (2004) illustrate the benefit of cumulative meta-analysis over a correlative approach to examine the relationship between publication year and effect size magnitude. Cumulative meta-analysis is more informative. In addition to the temporal changes of the mean effect, the evolution of the variation around the mean is also considered, thus also indicating stabilization of an effect.

2.4. Approaches to Assess Sufficiency and Robustness of Meta-Analytic Evidence

Cumulative research aims for increasing the reliability and robustness of existing research on a specific topic over time. Strong empirical evidence is particularly needed when recommendations for decision making are supposed to be derived. In the following, the considerations for and outline of instruments to evaluate the reliability of meta-analytic findings will be divided into two parts: First, the question of when do meta-analytic estimations have enough power to detect relevant effects due to a sufficient number of effect sizes will be addressed. This question can be answered from a classical statistical point of view, with the help of power analyses and signal detection theory, or using Bayesian statistics. Second, the robustness of the meta-analytic results against variations in analytical decisions (the “garden of forking paths,” Gelman & Loken, 2014) will be discussed. In the case of insufficient evidence, research gaps and needs of future research are identified and defined using evidence gap maps (EGMs).

2.4.1. Power in Meta-Analyses

Statistical power is a concept related to hypothesis testing. It is the likelihood of yielding a statistically significant effect by correctly rejecting the null hypothesis if it is false. Power is determined by three factors: First, a larger effect size increases the probability of finding a significant effect. Second, if the effect size estimate is more precise (i.e., if the corresponding standard error is smaller), the likelihood of a significant test statistic increases. Third, the probability of statistical significance increases if the significance level α moves away from zero. Whereas the expected effect size and α typically do not differ from those found in primary studies, the effect size precision is usually higher in meta-analyses than in single studies, as the combined sample size in a meta-analysis is always higher than the sample size in each single study. Accordingly, statistical power is often higher in meta-analyses than in single studies (Borenstein et al., 2009).

However, there is a difference between the fixed-effect and the random-effects model. Whereas in a fixed-effect analysis, precision is mainly determined by the total sample size, under the random-effects model, there are two sources of error. The sampling variance decreases with the total sample size by including more studies in a meta-analysis, as in the case of the fixed-effect model. However, the variance between the studies decreases with the number of studies included and the consistency of effect sizes. Thus, in case of substantial variation between studies and a small number of included studies, the power of a meta-analysis may also be smaller than in each single study included, due to the high variance between the studies limiting the potential power of the meta-analysis. Accordingly, in their empirical comparison of meta-analytic and study-specific power, Jackson and Turner (2017) conclude that, under the random-effects model, meta-analyses including five studies or fewer typically provide less power than the included studies individually.

Another limitation to the power of meta-analyses is the test of moderator variables. Testing a moderator effect basically means that the effect sizes for different levels of a moderator variable (in case of categorical moderators) are compared. As in the case of interaction effects in single studies, the sample size within groups of the same moderator level is smaller than the total sample size. Moreover, the differences in effect sizes between these groups, which are the effect sizes of interest in moderator analyses, are typically smaller than overall effect sizes (Borenstein et al., 2009). Thus, effect sizes are smaller and available sample sizes per group are lower. As a consequence, power for moderator tests is often low (Hedges & Pigott, 2004).

Cafri, Kromrey, and Brannick (2010) investigated the statistical power of meta-analyses published in the journal *Psychological Bulletin* between 1995 and 2005 and conclude that power for tests of moderator effects are the lowest, with 60 to 75% of the studies not meeting at least 80% power, which is regarded as adequate. As power depends on the number of studies included and the average sample size per study, especially for the test of moderator variables, a substantial number of primary studies is needed in a meta-analysis to be able to detect differences in effect sizes due to moderator effects. However, it is difficult to determine at which point the evidence concerning a research question is really sufficient. Therefore, in the following, different approaches focusing on this question are presented.

2.4.2. Evidence and Uncertainty: Signal Detection Theory

When using hypothesis testing, the costs and benefits of correct decisions, false positives, and false negatives should be considered equally to decide on the sufficiency of evidence. A metric for such a decision criterion is provided by **signal detection theory** (SDT; Fiedler, 2018). SDT, in general, analyzes the case of binary decision making on the basis of given evidence and in the presence of uncertainty. There are two alternative states. For example, a patient suffers from depression or he does not. The available evidence, in this case perhaps diagnostic tests with the patient, does not allow perfect discrimination between these two states. Thus, there are four possible events, as illustrated in Table 2: If the patient is depressive, it is a hit, if he is diagnosed and it is a false negative or a miss, if he is not. If the patient is not depressive, a diagnosis of depression is a false alarm and otherwise, it would be a correct rejection (McCarley & Benjamin, 2013).

The aim of a decision maker is, of course, to have a high rate of hits, which is the test sensitivity, and a low rate of false alarms, which is equivalent to the specificity of the test. But due to noise in the distributions of the two alternatives, it is impossible to correctly specify every event. Thus, the decision maker is confronted with a trade-off (McFall & Treat, 1999).

This trade-off is depicted in Figure 2. There are two hypothetical density distributions of test scores, one for a depressed population and one for a population not suffering from depression. The test scores in the depressed population are higher on average, but both distributions do overlap. Thus, a diagnosis of depression based on the test scores within the area of overlap always entails the risk of false decisions. The more conservative decision criterion with a small type I error rate α causes a high type II error rate β and, thus, low power to detect an effect. A more liberal cutoff value is accompanied by higher test sensitivity, but also by a high rate of false alarms.

Table 2: Events and Error in Evidence-Based Binary Decision Making

	Patient suffers from depression	Patient does not suffer from depression
Patient is diagnosed	Correct diagnosis/hit (Probability of a hit = Test power 1 - β)	False alarm (Type I error α)
Patient is not diagnosed	False negative/miss (Type II error β)	Correct rejection (1 - α)

The appropriateness of the applied decision criterion is dependent on the unconditional probabilities of the two events. If one possible outcome is much more probable than the other, the response criterion should take this information into account and favor the first outcome. This results directly from Bayes theorem and improves the decision making by using further information. Another factor influencing the choice of the decision criterion is the expected utility of a decision. This depends on the costs and benefits of possible decisions (Swets, Dawes, & Monahan, 2000). Fiedler (2018) advocates a more liberal strategy, especially in scientific areas with a low rate of real findings. The underlying rationale is that false-positive findings are examined further, whereas false-negatives may result in missing and, later on, ignoring potentially important findings. A liberal strategy, therefore, could support a beneficial culture of errors in science.

A measure that is based on statistical decision theory (Wald, 1949) and gives an index of accuracy independent of certain decision criteria and subjective influences on the response is the **relative operating characteristic** (ROC), which is displayed in Figure 3. The ROC plots hit rates against false alarm rates. Thus, the ROC curve enables illustrating the estimates of the four possible events of decisions and outcomes (Swets & Pickett, 1982) as described in Table 2. Hence, one graphic account indicates the power of a test and the probabilities for type I and type II errors (Swets, 1996).

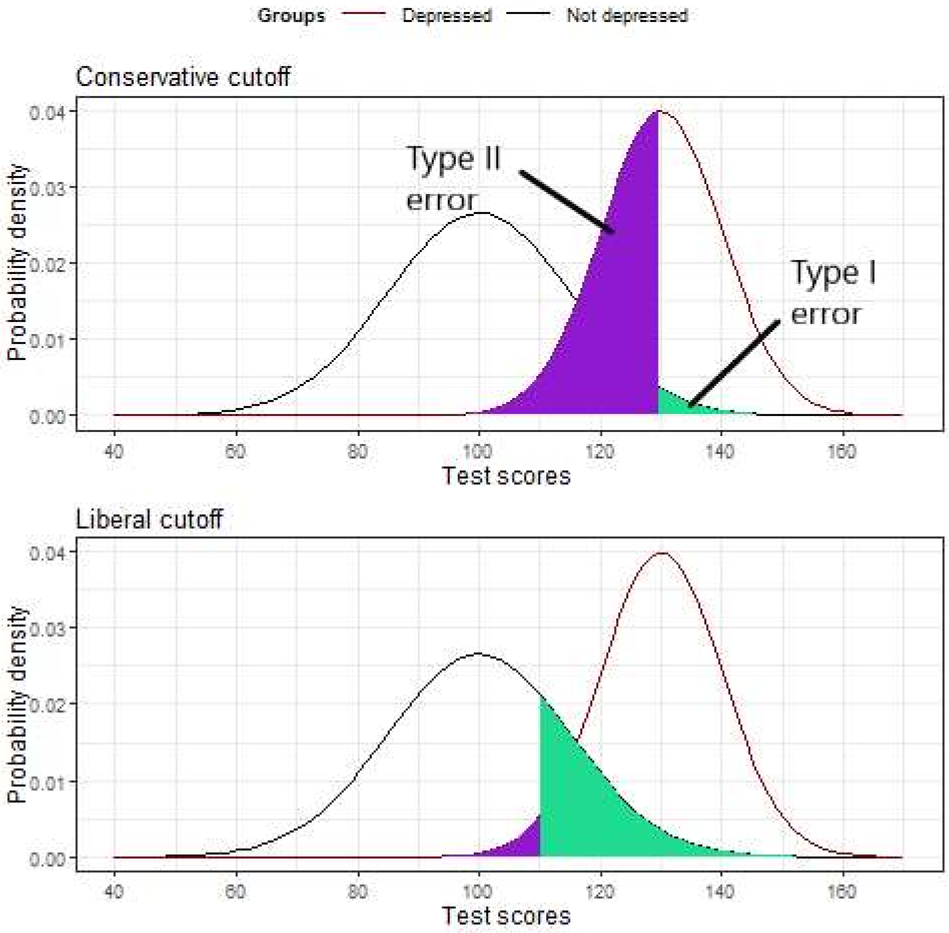


Figure 2: Types of Error in Decision-Making Under Uncertainty

The blue ROC curve in Figure 3 represents a defined level of discriminatory power of an applied test. The ROC curve is produced with the R package pROC (Robin et al., 2011) and based on data from the study of Turck et al. (2010). It shows all possible pairs of hit rates and corresponding false alarm rates from conservative to liberal decision criteria, whereby more conservative testing with a smaller rate of false alarms increases the rate of misses and vice versa in the case of more liberal testing. In Figure 3, the red dashed lines display two possible pairs of error rates: Assuming a typical false alarm rate alpha of 10% results in only about 39% power for the test. To reach 80% power, a false alarm rate of 55% would have to be accepted. The area between the curve and the diagonal black line, which represents a situation of no evidence, indicates the information value of the data (McFall & Treat, 1999).

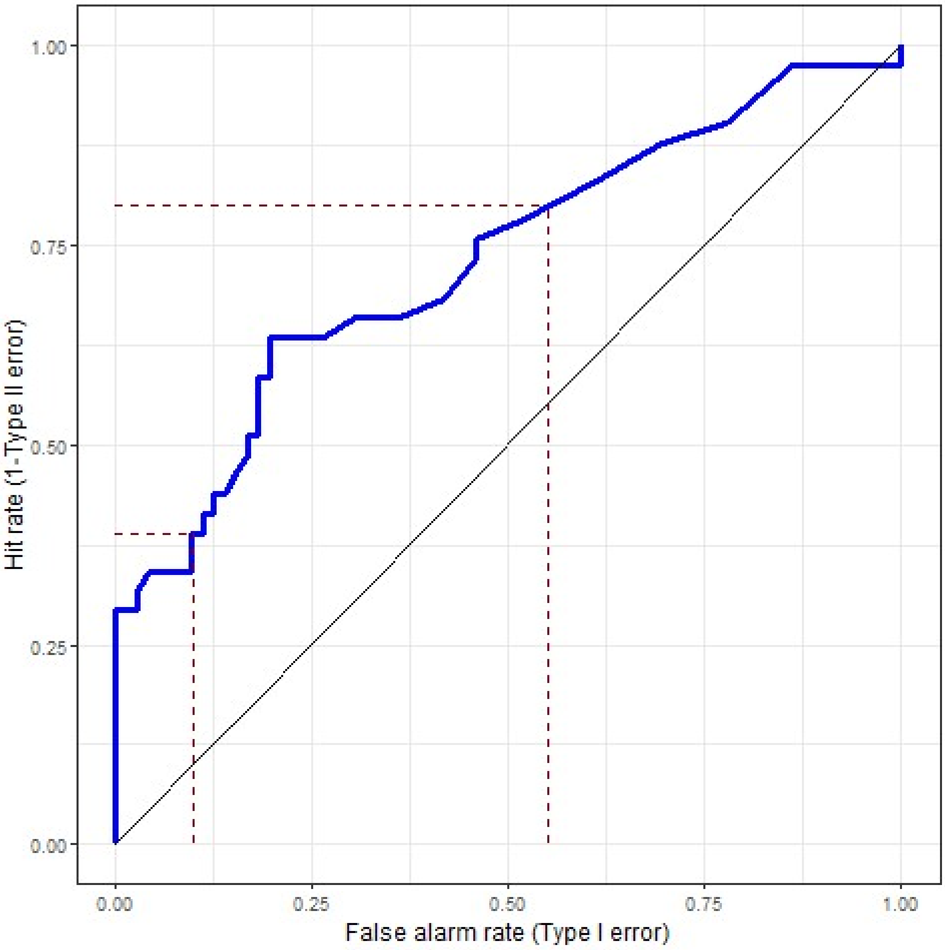


Figure 3: ROC Curve and Error Rates in Statistical Hypothesis Testing

There are two possibilities to increase the area under the ROC curve and, hence, the discriminatory power. The first possibility is a stronger effect and consequently a greater difference between the distributions of the two alternative states (e.g., the patient suffers from depression or does not suffer from depression). The second possibility is less noise or a more accurate measurement that leads to less variance in the distributions. In both cases, the overlap of the two distributions as depicted in Figure 2 is smaller. It is then possible to increase the hit rate and simultaneously keep the false alarm rate low.

Applying SDT in statistical testing allows researchers to assess the sufficiency of meta-analytic data with the help of the ROC curve. As in the case of a psychological diagnosis, there are four possible events in hypothesis testing: If the null hypothesis is rejected, it is a hit when the alternative is true. Otherwise, it would be a false alarm, which is equivalent to type I error. If the null hypothesis is not rejected, this is correct when the alternative hypothesis is wrong. But it is a miss when the alternative hypothesis is true. This case is equivalent to type II error (Swets, 1996).

2.4.3. Cumulative Evidence and Bayesian Statistics

Instead of null hypothesis significance testing, a shift to cumulative science aiming at continually improving the estimation of effect sizes and the related uncertainty is expected (Cumming, 2014). Kruschke and Liddell (2018) point out that Bayesian statistics are more appropriate for this purpose than the frequentist statistics.

The basic idea of **Bayesian estimation** is to update preexisting beliefs (prior distribution) on a probability distribution, as new evidence becomes available.
The Bayes theorem, $P(\theta|x) = (P(x|\theta) * P(\theta))/P(x)$, implies that the belief after taking into account new evidence (the posterior distribution $P(\theta|E)$), is a function of the prior $P(\theta)$ and the strength of the new evidence $P(x)$. If there is no evidence on a research question, Bayesian estimation typically starts with an uninformative prior, which is a uniform distribution across a range of plausible values. Thus, the posterior only depends on the new evidence, and the prior has no influence on the estimation. After each round of data observation, the assumptions in the prior are more informed (McCarley & Benjamin, 2013). Consequently, the results become more stable, as the influence of the already existing evidence in the form of the prior increases.

The posterior distribution consists of the modal value, which is the most credible value of the parameter θ, and the spread of the posterior distribution, which can be summarized with the 95% highest density interval (HDI). The HDI contains the 95% most credible values and reflects the precision of the estimation. Every time prior beliefs are updated with new evidence, the HDI typically becomes more narrow, as the precision of the estimation increases (Kruschke & Liddell, 2018). This reflects the idea of cumulative science by taking into account already existent knowledge and, thus, iteratively increasing the precision of the estimation of credible parameter values.

In Bayesian hypothesis testing, common decision rules are based on the **Bayes factor**, illustrated in Figure 4. It provides a continuous scale to compare the likelihoods of two competing statistical models given the available data. In case of Figure 4, the null hypothesis, represented as black dashed line, assumes equal probabilities at each point within the range of possible x-values. The most probable value under the alternative hypothesis is 0.7, whereas values below 0.4 are not likely to be observed. After having collected data, the Bayes factor is computed by comparing the probability of the given data under the alternative hypothesis versus the probability under the null hypothesis. If the data is more probable under the null hypothesis, a Bayes factor smaller than 1 results. In case of data favoring the alternative hypothesis, the Bayes factor is greater than 1.

The Bayesian framework thus allows evidence to be shown in favor of the null hypothesis relative to other alternatives. Selectivity towards hypothesis rejection can thus be reduced (Wagenmakers et al., 2018). The Bayes factor can be compared against a decision threshold to evaluate the strength of evidence and decide on the sufficiency of evidence. A common threshold is a Bayes factor of 10, indicating that the data are 10 times more likely under the alternative hypothesis compared to the corresponding null hypothesis (Kruschke & Liddell, 2018).

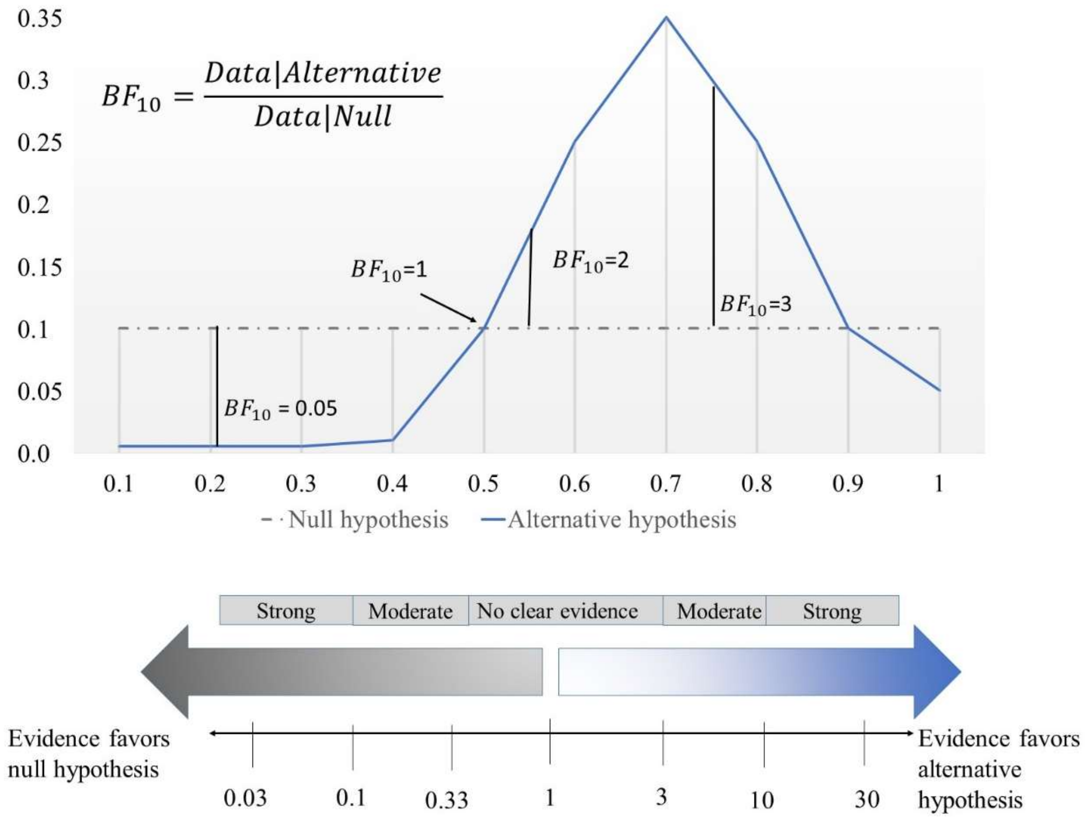


Figure 4: The Bayes Factor and Strength of Evidence

This approach is highly appropriate for cumulative science and can be used in the form of **sequential Bayes factors** (SBFs). Bayes factors are then employed in sequential designs with optional stopping rules, allowing unlimited multiple testing. Sample sizes are increased until a predefined threshold for the Bayes factor is reached (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). Data is collected iteratively. After each round of data collection, the Bayes factor is computed. As soon as the desired level of evidence is reached, sampling is stopped. In addition to the final Bayes factor, the mean and corresponding highest posterior density (HPD) interval are reported. This design ensures that data collection stops when the evidence is sufficient and, therefore, is more efficient than fixing sample sizes based on expected effect sizes from the beginning (Schönbrodt & Wagenmakers, 2018).

In **meta-analyses, Bayesian methods** provide an opportunity for incorporating past evidence that cannot be included in the meta-analysis itself, for example, due to methodological restrictions. However, meta-analysts have to carefully consider the specification of informative priors, as these can have a huge effect on the posterior (Lewis & Nair, 2015). To facilitate the use of Bayesian meta-analytic techniques for cumulative science, Lakens, Hilgard, and Staaks (2016) point out that Bayesian inferences might require the reporting of test statistics instead of effect sizes. Such considerations are crucial to foster future-proof, cumulative meta-analyses.

2.4.4. Sensitivity and Robustness of Results in Face of Analytic Degrees of Freedom

A prominent objection against classical hypothesis testing is the possibility to conduct multiple potential comparisons with a dataset. Supposing the worst of a researcher, one could suspect p-hacking (Simmons, Nelson, & Simonsohn, 2011), the act of consciously searching for significant effects or the best results from multiple tests performed on the data. Gelman and Loken (2014) argue that even without any unethical intentions, analyses can be refined in terms of data exclusion or coding based on the data given. These degrees of freedom, that can be completely appropriate and within well-accepted research practices, call into question the informative value of statistical significance, as this lies in the generalizability across multiple potential datasets. Therefore, it is relevant to consider the dependency of analysis choices on the data.

One way of assuring independence of the analysis strategy from the data is by preregistration, as this means to define all steps of data collection and analysis before data collection actually begins. It can be advantageous for researchers that reviewers evaluate the importance of the planned research and the methodological quality before starting their data collection. Authors are provided feedback during the time it is still possible to improve the methodological design. Furthermore, there are preregistration formats, such as Registered Reports (Nosek & Lakens, 2014), where researchers can get an in-principle acceptance of their work before data collection starts. Thus, they are assured that their outcomes will be published as long as they adhere to the predefined research plans, irrespective of the results.

To rule out detrimental effects of arbitrary choices in the construction of data for analysis purposes, the idea underlying multiverse analyses is to perform analyses across the multiverse of datasets that could be obtained by different reasonable data processing choices. Thus, it can be seen as a systematic extension of outlier analysis, displaying the robustness of outcomes not only across different exclusion criteria, but across all relevant steps in data processing (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

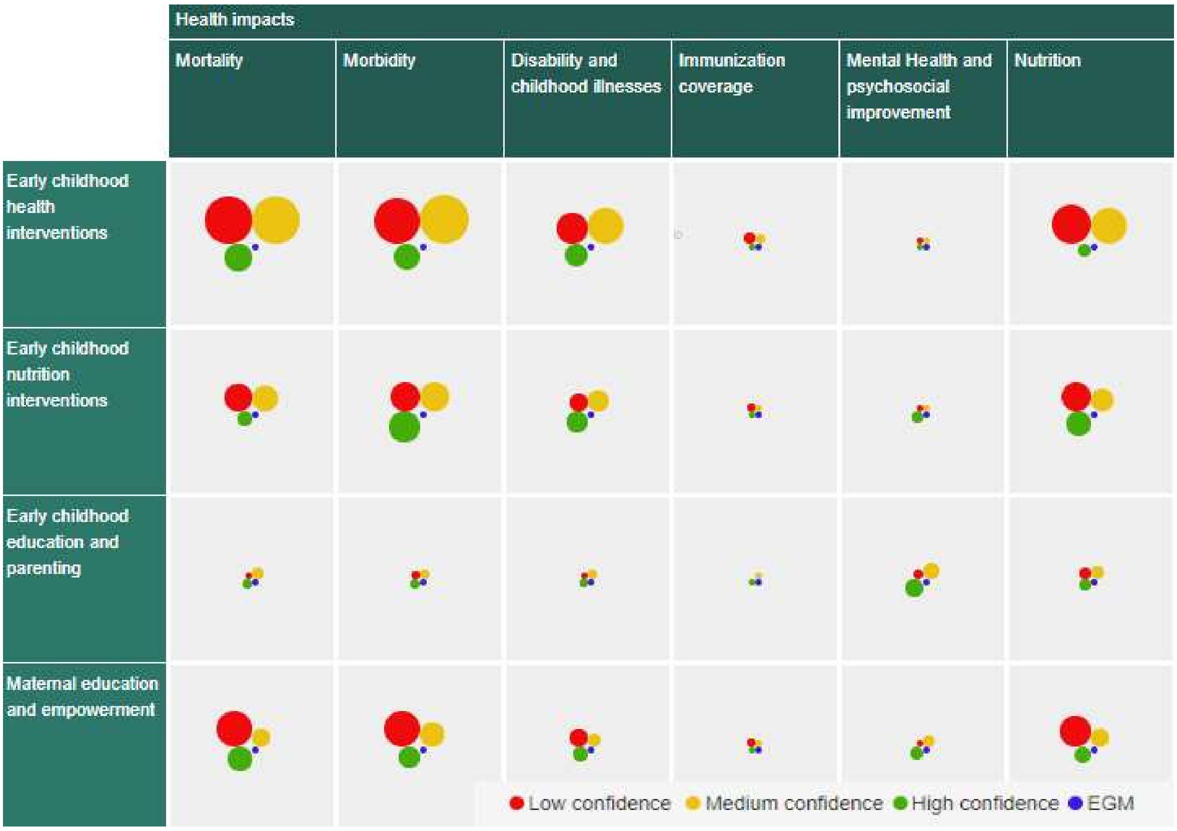
Specification curve analysis is a similar approach that not only focuses on which data to analyze, but also on how to specify the analyses. The aim is to simultaneously report results for all valid and nonredundant analysis specifications within a certain theoretical view. Thereby, disentangling whether different conclusions derive from theoretical disagreements on valid specifications or from arbitrary selective reporting of results is facilitated. There is already a long tradition of analyzing the sensitivity of results to analytical specifications. An example is the reporting of regression results from different specifications in a table with multiple columns or the use of model selection. Specification curve analysis extends and formalizes these approaches following three steps: (1) definition of the set of reasonable specifications, (2) estimation of all the specifications reported in the descriptive specification curve, and (3) joint statistical tests based on an inferential specification curve (Simonsohn, Simmons, & Nelson, 2015).

At each step in a meta-analysis, as in single studies, decisions have to be made that are often arbitrary, but affecting the overall results. For instance, Voracek, Kossmeier, and Tran (2019) are the first to apply specification curve analysis and multiverse analysis in combination with combinatorial meta-analysis (Olkin, Dahabreh, & Trikalinos, 2012). They divide the specification factors into external factors (such as effect size or model choice) and internal factors (potentially relevant study features making up for different potential study designs). Voracek et al.'s (2019) consideration of all potential study designs and all combinations of analysis decisions of interest yielded a total of 1,529 different meta-analytic specifications that were calculated. This approach allows a thorough and systematic sensitivity analysis to assess the robustness of meta-analytic evidence. The specification factors have to be determined for each meta-analysis individually and controversies in a research field can be treated purposefully.

2.4.5. Identification of Research Gaps Using Evidence and Gap Maps

Beyond assessing the sufficiency and robustness of meta-analytic evidence, it is important to specify existing research gaps to guide the conception of future research. A relatively new approach to systematically map the research activity in a field to support evidence-based decisions and provide a foundation for future research priorities is the use of **evidence maps** (Saran & White, 2018). The earliest evidence maps were published in 2003 (Katz, Williams, Girard, & Goodman, 2003), and there are various different approaches to evidence mapping.

A popular example of mapping evidence to make research gaps visible is found in the evidence gap maps (EGMs) of the International Initiative for Impact Evaluation (3ie) (Snilstveit, Bhatia, Rankin, & Leach, 2017). An EGM is a matrix of relevant interventions and empirical outcomes in a particular research area. The cells of the matrix contain colored bubbles indicating the type of evidence. EGMs have been developed to make evidence on social intervention programs accessible to decision makers and to guide research funding efficiently.



	Health impacts					
	Mortality	Morbidity	Disability and childhood illnesses	Immunization coverage	Mental Health and psychosocial improvement	Nutrition
Early childhood health interventions						
Early childhood nutrition interventions						
Early childhood education and parenting						
Maternal education and empowerment						

● Low confidence ● Medium confidence ● High confidence ● EGM

Figure 5: Excerpt of the Campbell-UNICEF Child-Well-being Mega Map

An example of an EGM is the Campbell-UNICEF Child-Well-being Mega Map (Saran, Albright, Adona, & White, 2020) depicted in Figure 5. Each cell of the matrix shows the evidence in terms of systematic reviews or EGMs according to the impact of the intervention in the corresponding row on the outcome in the column. The bubble sizes represent the amount of evidence included in systematic reviews, the colors indicate the confidence in the evidence. Thus, an EGM gives a quick overview on which combinations of interventions and outcomes might be understudied and how much confidence in the impact of relevant moderators is provided by the existing evidence. The design of both future primary studies and meta-analyses can be informed accordingly.

In the following, evidence gaps in the field of survey methodology will be identified and two meta-analyses aiming at closing these gaps are presented and discussed according to the sufficiency and robustness of their results. To enable the continuous accumulation of evidence, a scientific infrastructure for replicable and extendable meta-analyses is needed. The concept of Community-Augmented Meta-Analyses is suggested as a solution and a corresponding system in the field of psychology called PsychOpen CAMA is presented in detail in chapter 6.

3. Evidence and Research Gaps in Survey Methodology

3.1. The Total Survey Error Framework

A central framework for research in the field of survey design methods is the total survey error (TSE) paradigm (Groves et al., 2009; Weisberg, 2005). Therefore, it is an adequate approach to identify research gaps and derive research questions in the field of survey methodology. The TSE covers all types of errors that may cause deviation of survey results from the underlying true values. It is used to identify the possible sources of error to guide design and method choices minimizing the TSE given the resources available (Biemer, 2010). The components of the TSE can be partitioned to nonobservation and observation errors (Groves & Lyberg, 2010).

Nonobservation errors occur due to not obtaining data from all elements of the target population. As a consequence of nonobservation errors, the target population is not represented properly by the survey sample. There are different potential reasons and sources of error that may cause this misrepresentation:

A coverage error is the misfit between the target population and the sampling frame that is used to select the survey sample. Ideally, the sample frame and the target population should be congruent. In practice, information on the target population is often not accurate and undercoverage or overcoverage may occur (Lessler & Kalsbeek, 1992).

Even in case of sampling from an ideal sampling frame, a sample is not a perfect representation of the target population. The uncertainty due to not having census information is called *sampling error*, and it depends on the sample size. The more units of the population that researchers can survey, the less uncertainty remains and the more precise are the inferences drawn from the sample to the target population.

Nonresponse errors occur if the survey respondent refuses to answer either the survey as a whole in the case of unit nonresponse or only some of the items in the questionnaire. Nonresponse does not only reduce the amount of available information, it often also leads to serious biases, as the refusal to provide answers to survey questions is often not at random, but associated with relevant characteristics of the respondent, or even related to the true value in the variable that is subject to item nonresponse (Groves & Couper, 1998).

An adjustment error is a missing or an erroneous calculation of survey weights to compensate for unequal selection probabilities, nonresponse, and coverage errors. This error type is sometimes subsumed under data processing errors (Biemer, 2010). As survey weighting is related to representation, associated errors are defined as a separate source of error here.

Observation errors involve differences between reported and true values. These differences may occur at different stages of data collection and processing and, therefore, can be further decomposed into more specific error sources:

A *specification error* is a problem of construct validity. It occurs if a survey item does not measure the construct that it is intended to measure.

Measurement errors describe the collection of incorrect responses to survey questions. The reasons why respondents provide incorrect information can be diverse. Interviewer effects are, for example, unintentional influences on responses due to their speech, appearance, or gender, and noncompliance with the survey procedures (West & Blom, 2017). A poorly designed questionnaire can lead to measurement error by confusing or overtaxing respondents (Peytchev & Peytcheva, 2017; Sanchez, 1992). In addition, the information retrieval and processing within the respondent can be a source of error (Tourangeau & Hanover, 2018) and is also affected by the survey context, such as the mode of survey administration and the setting (Gnambs & Kaspar, 2017). For example, in the case of sensitive questions, the assurance of privacy and confidentiality can be crucial to prevent intentional misreporting (Tourangeau & Yan, 2007). As an extension of the TSE, Smith (2011) also suggests the ‘conditioning error’ that is a type of measurement error unique to multi-wave panel studies. In these studies, the experience of being interviewed previously may influence question understanding and response behavior in subsequent waves.

Data processing errors cover mistakes made in data entry, editing, coding, and the computation and tabulation of estimates and final analysis results (Groves et al., 2009).

In the following, an overview of the evidence on each of these sources of error will be given to detect the existing gaps in research on survey methodology. A research program to close these gaps and improve evidence for decisions in planning and conducting surveys will be derived.

3.2. Evidence of Nonobservation Errors

Meta-analytic evidence for nonobservation errors is only available for nonresponse; no evidence currently exists on coverage, sampling, or adjustment errors (Cehovin, Bošnjak, & Lozar Manfreda, 2018).

Coverage errors are mainly discussed in relation to data collection modes (Roberts & Vandenplas, 2017), especially concerning Internet coverage and related bias, and in sampling hard-to-reach populations (Schnell, Trappmann, & Gramlich, 2014). They are closely linked to nonresponse errors (e.g., Dewaele, Caen, & Buysse, 2014). Both forms of nonobservation errors depend on the rate of missing data and the difference between elements that are missing and those not missing. Actually, there is often no exact information on the magnitude of coverage and nonresponse errors (Biemer & Lyberg, 2003). Differences in characteristics between a sample and the target population may result from differences between the target and the frame population (coverage) as well as from differences between the frame population and the final respondents (nonresponse). To differentiate coverage from nonresponse error, it is necessary to discern the extent to which the differences are due to errors in the sampling frame and the extent to which they are the result of not obtaining data from sample units. Even if noncoverage rates and nonresponse rates are known (and this is sometimes not even the case), the bias can arise due to differences between the covered and noncovered population or due to differences between respondents and nonrespondents or both.

To estimate the coverage error separately from nonresponse bias, information on the noncovered population or the complete target population is necessary to compare it to the population actually covered. This is usually not available and, thus, coverage bias cannot be computed properly (Halmdienst & Radhuber, 2018). This might be a reason explaining why we find a lack of targeted studies on coverage errors and a stronger focus of meta-analyses on nonresponse error. There are studies approximating the coverage error by using samples of large household surveys, assumed to be representative for the target population, as reference population. However, these also suffer from nonresponse and do not provide an exact calculation for a coverage error. Yet this procedure can be used to estimate the potential effect of data collection modes or a sampling method on the coverage of the target population.

In the study of Mohorko, de Leeuw, and Hox (2013), the hypothetical coverage error of an online sample was examined by specifically comparing participants with online access to the whole sample of the public opinion survey, Eurobarometer. Their findings reveal that the increase in Internet penetration from 2005 to 2009 is accompanied by decreasing coverage errors for the online subsample. The Eurobarometer sample serves as the reference population here, even though in some countries, less than 50% of sampled units responded to the survey. The question assessing Internet access is used as an indicator of which elements would be covered in a hypothetical online-only data collection mode.

The sampling of hard-to-reach populations is another issue of coverage errors. Schnell et al. (2014) evaluated the approach of name-based sampling (NBS) to identify migrants by name. The reference sample is one of the largest German panel studies (PASS; Trappmann, Gundert, Wenzig, & Gebhardt, 2010). Differences in sociodemographic characteristics of migrants in PASS data and migrants that would have been selected by NBS were measured, and Cohen's d was calculated as the effect size for the resulting coverage error. The migrants who were identified by NBS tended to live in larger households with more children, less likely to be married to a German partner, and poorly educated in comparison to the migrants not identified by NBS.

Sampling errors primarily depend on the sample size of a study and on the sampling method. In the case of simple random sampling, all respondents have the same probability to be included in the sample. Other random sampling methods, such as clustered or stratified sampling, are sometimes more efficient for practical or statistical reasons, but increase sampling errors. Nonrandom sampling does not allow the computation of sampling error and estimates for population parameters reliably (Halmdienst & Radhuber, 2018). To conclude, sampling error is rather a statistical issue, and the intervention options to deal with sampling errors are clearly defined and limited to mathematical estimation methods.

Many survey design characteristics and intervention variables of interest in the framework of survey methodology (data quality, contacting protocols, questionnaire design, interviewer effects) cannot be studied in relation to coverage or sampling errors (Cehovin et al., 2018). Hence, the number of primary studies in survey methodology on coverage or sampling errors is small. Quick searches in Web of Science and JSTOR for coverage error and survey methods resulted in a total of 11 records (May 2020). Not all of these records necessarily include an analysis or computation of coverage errors nor does a single record mention coverage error in the title. If there are relevant studies at all, they still might not provide comparable measures to enable a meta-analysis of the results. For survey methods and sampling error, a total of 51 hits were found in the two abovementioned databases. For these records, the same limitations hold: Sampling errors are often simply mentioned, but are not the primary topic of the study. In sum, evidence for coverage and sampling errors in survey methodology remains scarce, partly for good reasons, and it might not yet be fruitful to conduct meta-analyses in this context.

Meta-analytic evidence on nonobservation errors in survey methodology focus exclusively on **nonresponse error**. Cehovin et al. (2018) have identified experimental meta-analyses on nonresponse dealing with contacting protocols (Church, 1993; David & Ware, 2014; de Leeuw, Callegaro, Hox, Korendijk, & Lensvelt-Mulders, 2007), administration mode (Lozar Manfreda et al., 2008; Medway & Fulton, 2012; Shih & Fan, 2007/2008), and questionnaire design (Rolstad, Adler, & Ryden, 2011; Villar, Callegaro, & Yang, 2013).

Moreover, nonexperimental meta-analyses on nonresponse deal with data quality issues (Mercer, Caporaso, Cantor, & Townsend, 2015), sample characteristics (Mavletova & Couper, 2015, van Horn, Green, & Martinussen, 2009), the type of target population (Cho, Johnson, & VanGeest, 2013; Cook, Heath, & Thompson, 2000; Shih & Fan, 2007), and survey setting (Mavletova & Couper, 2015; Yarger et al., 2013).

Nonresponse bias, according to Groves and Peytcheva (2008), is examined in terms of differences between respondents and nonrespondents. In their meta-analysis, they present almost 1,000 nonresponse bias estimates observed in 59 studies. Estimated population means were produced from sample frame data, screening interviews, follow-up studies, and reports of intentions to respond to later surveys. These techniques all guarantee that the reference is the population sampled and, thus, they clearly distinguish nonresponse bias from coverage errors.

Weighting or imputation procedures can be used to treat missing data bias due to coverage as well as nonresponse. These typically rely on assumptions that are hardly ever completely fulfilled. Therefore, missing data and corresponding bias usually cannot be fully compensated (Brick, 2013). Especially in the case of large noncoverage and nonresponse, or if necessary covariates are missing, weighting does not even remove half of the missing data bias (Vehovar, Lozar Manfreda, & Batagelj, 1999). The use of adjustment procedures is strongly recommended. Yet their application requires advanced statistical skills and should be done with caution. If survey weights are calculated erroneously or imputation models are wrong, additional **adjustment error** may be introduced (Biemer & Lyberg, 2003).

Typically, adjustment errors do not play a crucial role in the design of a survey. Thus, it is plausible that the quick searches undertaken in Web of Science and JSTOR for adjustment error and survey in the fields of psychology, sociology, and statistics resulted in only 25 hits (May 2020). In these records, if adjustment errors are the main focus of the papers at all, they are mainly examined in terms of statistical comparisons of methods for dealing with nonobservation bias (e.g., Slud & Bailey, 2010; Zhang, Thomsen, & Kleven, 2013). This is rather relevant for the data analysis and can be calculated and compared specifically for single datasets by using, for example, Monte Carlo simulation methods (Schanze & Zins, 2019). This could be an explanation for the finding that there are only few primary studies and no meta-analyses on adjustment errors in the area of survey methodology.

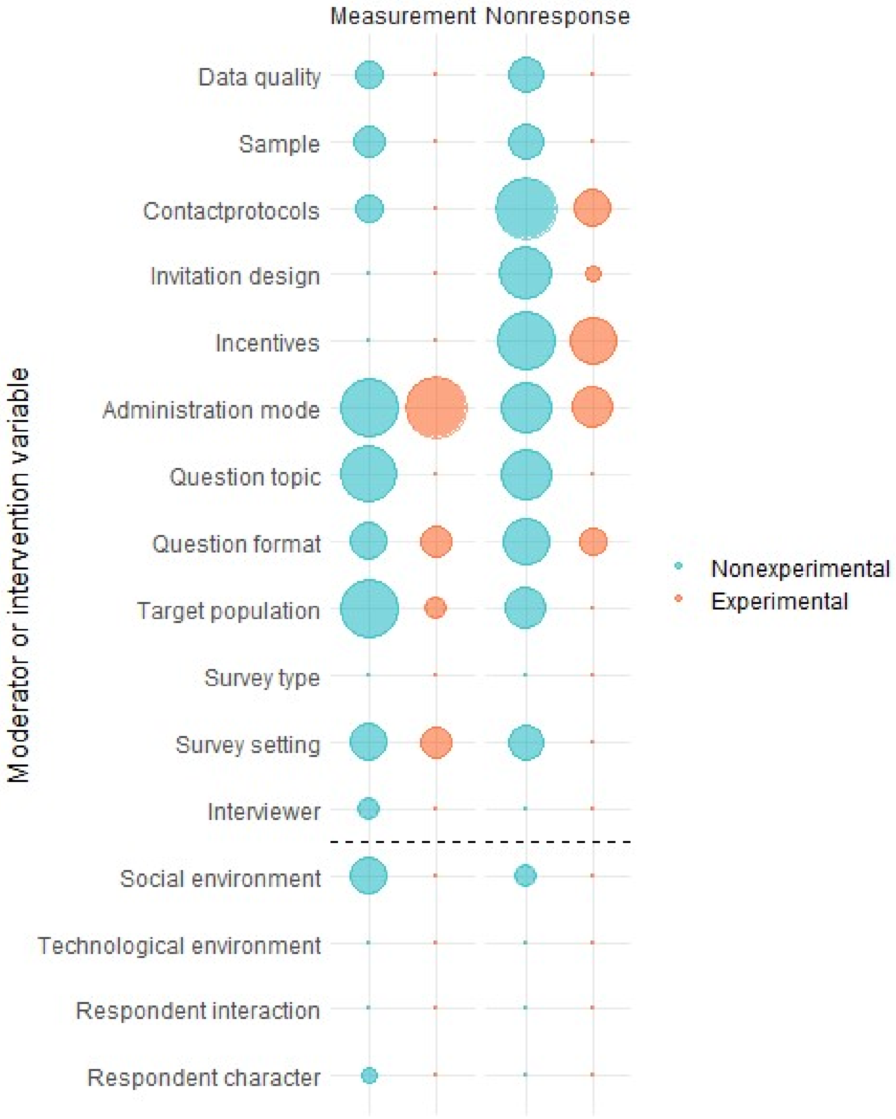


Figure 6: Meta-Analytic Evidence in Survey Methodology, 1984-2019

The evidence map in Figure 6 illustrates the extent of meta-analytic evidence in survey methodology. The data for Figure 6 is an extension of the review of Cehovin et al. (2018). Utilizing the same search terms as in this review, a database search in Web of Science, JSTOR, PubPsych, and Sociological Abstracts resulted in five additional relevant meta-analyses published between 2017 and 2019 (Ang & Eisend, 2018; Cornesse & Bošnjak, 2018; Daikeler, Bošnjak, & Lozar Manfreda, 2020; Li & van den Noortgate, 2019; Saywitz, Wells, Larson, & Hobbs, 2019). Measurement and nonresponse are the only dimensions of the TSE framework serving as outcome variables in these reviewed meta-analyses. However, there are 16 relevant intervention variables in experimental research or moderator variables in nonexperimental meta-analyses that could potentially have an effect on these TSE dimensions.

Most of the gaps in meta-analyses on nonresponse are due to moderators that are not under the researcher's control. These are located below the dashed black line in Figure 6. These effects cannot be studied experimentally and, moreover, evidence has only limited benefit for concrete recommendations on survey design decisions. Moderators not explored in experimental meta-analyses on nonresponse are data quality, which includes survey burden and response quality here, sample characteristics, questionnaire topic, type of target population, and survey setting. Of these, survey burden is a moderator (1) that can be varied experimentally and (2) that is promising in terms of providing results that enable clear recommendations and (3) for which enough evidence from primary studies is available for use in meta-analyses. Bogen's (1996) literature review summarizes nonexperimental and experimental studies on the effect of questionnaire length and response rates. Further evidence published more than a decade after this review is also available (e.g., Galesic & Bošnjak, 2009).

Moderators yet to be examined in experimental or in nonexperimental meta-analyses are survey type and interviewer characteristics. In their research synthesis on interviewer effects, West and Blom (2017) discuss the potential effects of interviewer characteristics and behavior on nonresponse error and other types of survey error. This synthesis could serve as a starting point for a meta-analytic investigation of interviewer effects on nonresponse.

3.3. Evidence of Observation Errors

Another blind spot in terms of meta-analyses in survey methodology is **specification error** (Cehovin et al., 2018). This is an issue of operationalization and construct validity. It needs qualitative assessment and is highly topic specific. An exemplary study on specification error is Regan and Oaxaca's (2009) investigation of different operationalizations of work experience as a source of bias. There are no standard procedures to measure specification error (Fuchs, 2010). Consequently, similar studies using comparable study designs and procedures to estimate specification errors that could be used for a meta-analysis are not available.

Measurement errors can occur due to various reasons. The information retrieval, processing, and provision of the respondent can be affected by the administration mode, the questionnaire design and wording, and the context of the survey, as well as by characteristics of the interviewer and the respondents.

Experimental meta-analyses on the effects of the survey mode on social desirability distortion consistently conclude that there is no relevant difference in social desirability between the various self-administered survey modes (Dodou & de Winter, 2014; Gnambs & Kaspar, 2017). However, being alone and self-administering the survey reduces socially desirable responding compared to being surveyed by an interviewer (Richman, Kiesler, Weisband, & Drasgow, 1999; Tourangeau & Yan, 2007). Moreover, socially undesirable behaviors are reported more often in computerized surveys than in cases of paper administration (Gnambs & Kaspar, 2014).

Experimental meta-analytic evidence on effects of the questionnaire design, format, and wording on the response behavior is still scarce (Callegaro, Murakami, Tepman, & Henderson, 2015). There are also contradictory results of meta-analyses on the same research question, for example, whether patients report higher health state valuations than the general population (Peeters & Stiggelbout, 2010) or no differences are found between these target groups (Dolders, Zeegers, Groot, & Ament, 2006).

Nonexperimental meta-analyses also reveal that question wording (Pupovac & Fanelli, 2015), survey administration, and contact method (Fanelli, 2009) affect the reporting of sensitive behaviors and satisfaction levels (Voutilainen, Pitkäaho, Vehviläinen-Julkunen, & Sherwood, 2015). The validity of survey results, especially regarding sensitive topics, can be improved by special survey designs (Lensvelt-Muders, Hox, van der Heijden, & Maas, 2005). Sometimes, characteristics of the target population may also influence the validity of self-reports. For proenvironmental behavior, Kormos and Gifford (2014) found that men report their behavior more validly than women.

Data processing errors occur while editing and coding the survey responses. Examples of data processing errors are problems with deciphering handwriting in paper surveys, the treatment of inconsistent responses, and coder errors. A frequent coder error is the incorrect classification of open-ended reports, such as occupation codes (Conrad, Couper, & Sakshaug, 2016), activities in time use surveys (Sturgis, 2004), or political knowledge (DeBell, 2013). Processing errors are discussed in the context of interviewer effects, and there is evidence for variability in the competencies of data entry and coding of interviewers (Sayles, Belli, & Serrano, 2010; Smyth & Olson, 2020). A comparison of the work of professional coders from different coding agencies revealed relatively low reliability between agencies (Massing, Wasmer, Wolf, & Zuell, 2019).

There are attempts to reduce data processing errors, such as presenting suggestions of codings to the respondents during the interview (Schierholz, Gensicke, Tschersich, & Kreuter, 2018) or constructing more precise and comprehensive coding rules (DeBell, 2013). Data processing errors are mainly associated with the interviewer, as illustrated by the findings of a quick search in JSTOR (May 2020) of 136 hits for the search terms "processing error" OR "coding error" OR "data editing" AND "survey" AND “interviewer.” Other moderators of interest, such as administration mode or question type, resulted in fewer than 10 hits. Thus, interviewer characteristics could be of interest in the context of data processing to further improve data quality. Whether there are enough comparable primary studies among the 136 hits to provide clear recommendations based on sufficient evidence should be investigated.

In general, Figure 6 shows that there are no experimental meta-analyses on the effects of interviewers, question topic, survey burden, sample characteristics, and contact protocols on measurement error. Meta-analytic evidence on measurement errors is completely lacking for invitation design, incentives, and different types of surveys (panel, longitudinal, cross-sectional).

3.4. Design of Two Meta-Analyses in Survey Methodology

It can be concluded that meta-analyses in survey methodology have dealt exclusively with nonresponse errors and measurement errors. For other types of survey errors, evidence of comparable primary studies is often scarce. This may be due to a lack of relevant information for proper comparisons, as in the case of coverage errors, or because a specific error type is commonly examined statistically (e.g., sampling or adjustment errors) or needs qualitative assessment (e.g., specification errors).

Findings in survey methodology are of practical importance to guide survey operations aiming at reducing survey errors and biases by optimizing data collection procedures (Bošnjak & Danner, 2015). Thus, a strong evidence base for making decisions on survey operations, such as the choice of survey modes and incentives or appropriate measurements for sensitive topics is desirable. To increase knowledge in survey methodology, it is important to close the identified gaps, to provide instruments to collect available information, and to promote cumulative research.

As a starting point, two meta-analyses are conducted to address existing gaps in the research on nonobservation errors and observation errors. The design of both meta-analyses is guided by the research gaps detected and by current developments in the field of survey methodology. Namely, the decrease in response rates, the increase in the use of online surveys, and the growing relevance of panel infrastructures leading to the question on the impact of potential "professional respondents" (Zhang, Antoun, Yan, & Conrad, 2020) on data quality.

3.4.1. Nonobservation Errors: Response Rates in Online Surveys

Nonresponse is one of the most severe problems in social and behavioral research challenging both the internal and external validity of surveys (Hox & De Leeuw, 1994). If the causes for missingness are independent to any other parameter (Little & Rubin, 2019), nonresponse reduces the amount of data collected and thus results in less precise estimates and lower statistical power. However, if the reason for nonresponse is nonrandom, missing data can even cause severe bias and invalid conclusions, as the final respondents are no longer representative for the population of interest (Groves & Peytcheva, 2008).

The willingness to participate in surveys has decreased in the past decades in the social and political sciences (de Leeuw & de Heer, 2002) as well as in counseling and clinical psychology (Van Horn, Green, & Martinussen, 2009). This trend can aggravate the possible bias due to nonresponse. As survey participation is interrelated with communication (Schwarz, 2003), a factor that might have affected response rates is the increase of Internet usage in recent years (World Bank, 2018). The Internet has become a popular platform for conducting surveys, due to the fast and easy implementation and low costs of online surveys. Yet they are thought to suffer even more from issues of nonresponse and a lack of representativeness (Cook, Heath, & Thompson, 2000).

The extensive use of online surveys for data collection may have caused oversurveying and thus, a decrease in the participation in online psychology surveys (Groves et al., 2009; Weiner & Dalessio, 2006), reflecting the trend in other scientific branches and for other modes of data collection (Brick & Williams, 2013). This is due to less attention to single communication requests, because of the amount of information to be processed (Groves, Cialdini, & Couper, 1992).

On the other hand, the ever-increasing growth of Internet use could have changed the willingness to participate in these types of surveys relative to other survey modes. However, the overall conclusions on the response rate differences between online surveys and other survey modes did not change significantly between the meta-analysis conducted by Lozar Manfreda, Bošnjak, Berzelak, Haas, and Vehovar (2008) and a recent update by Daikeler, Bošnjak, and Lozar Manfreda (2020). However, coverage bias in Internet surveys has decreased over time with the increasing use and diffusion of the Internet (for the Eurobarometer: Mohorko, de Leeuw, & Hox, 2013).

As considerable changes over time can be expected for the participation in web surveys, the continuous updating of the evidence and the examination of evidence over time is highly relevant. The moderating effects of time and survey design will be tested using study characteristics (contact mode, number of items, and use of incentives). The first hypothesis focuses on the time effect:

H1: The response rates in online psychology surveys have decreased over time.

The evidence map in Figure 6 has shown that special target populations and survey settings are understudied in the field of nonresponse errors. Thus, in this meta-analysis, the trend of declining response rates is examined for online surveys in psychology, specifically focusing on participants with depression or anxiety disorders. From an epidemiological perspective, this is an important population that may be hard to reach and difficult to motivate to participate in studies.

It is crucial to study possible effects of a study’s design on people’s willingness to participate in the study to guide survey design decisions. In times of oversurveying, one method to achieve higher response rates is personal contact. Participants can be invited to access online surveys in various ways that differ in the extent of personal contact. For example, contacting potential respondents by phone is a more personal invitation than sending an e-mail invitation to participate via a mailing list. Schaefer and Dillman (1998) stress the importance of a personal contact to potential respondents, an act which conveys their importance for the survey institution. The meta-analysis of Cook et al. (2000) also shows that more personalized contacts yield higher response rates in online surveys. Examining the type of contact to deliver the invitation to participate in a survey, it is assumed:

H2: Personal or phone contact as an invitation mode yields higher response rates in online psychology surveys than e-mail invitations.

The influence of survey length on response rates was examined meta-analytically by Rolstad, Adler, and Rydén (2011): They found a clear association between questionnaire length and response rates. Yet it is not clear whether the difference in response rates is directly attributable to the length of the questionnaires. For the association between questionnaire length and experienced response burden, only weak support is found. Galesic and Bošnjak (2009) conducted an experiment in which the announced length of the survey, incentives, and the order of thematic blocks were randomly assigned to participants. Findings revealed that the respondents were more likely to start the survey when the stated length was shorter. In the context of the higher importance of the cost-benefit ratio due to cultural individualization (Santos, Varnum, & Grossmann, 2017), over time it can be expected that longer studies suffer more from the decrease in participation than shorter ones. Thus:

H3: The higher the number of items in an online survey questionnaire, the lower is the response rate.

An intensively researched topic in the area of survey participation is the effect of incentives. An early meta-analysis showed that only initial incentives had an effect on response rates (Church, 1993). In general, cash incentives have a stronger effect on response rates than nonmonetary incentives (Pforr et al., 2015). The difference between prepaid and promised incentives was also corroborated by Mercer et al. (2015), but only for telephone and mail surveys. These findings from cross-sectional research indicate that incentives, under certain conditions, may have an effect on response rates. However, in the present research, a special population is considered, namely samples with a relevant share of respondents suffering from depressive or anxiety disorders. Following the reinforcement sensitivity theory (Corr, 2002), it is expected that this population, scoring high on neuroticism, will be less sensitive to rewards (Beevers & Meyer, 2002; Bijttebier, Beck, Claes, & Vandereycken, 2009;

Pinto-Meza et al., 2006). This would also imply that the effect of incentives for survey participation will be lower than expected for the general population. Thus, it is hypothesized:

H4: Response rates in online psychology surveys in a group scoring high on neuroticism are not affected by incentives awarded for participation.

3.4.2. Observation Errors: Panel Conditioning in Sensitive Items

Starting from the finding of a lack of meta-analytic evidence on the effects of special survey types, interviewers, survey topic, and survey burden on measurement error, a highly relevant topic of measurement quality in panel studies is panel conditioning. The concern that repeated interviewing of respondents may affect their opinions (Lazarsfeld, 1940), attitudes, and behaviors (Kalton, Kasprzyk, & McMillen, 1989) has not been sufficiently studied to draw clear conclusions yet (Struminskaya, 2016). The mechanisms behind panel conditioning as well as the possible outcomes are diverse, complicating the development of a unified framework (Bergmann & Barth, 2018). Moreover, the research designs used in studying panel conditioning are often insufficient and cannot clearly distinguish panel conditioning effects from biases due to panel attrition (Warren & Halpern-Manners, 2012).

Panel data are indispensable for answering longitudinal questions and drawing causal conclusions. Because both the collection and the long-term maintenance of participant pools are complex and expensive, open panel infrastructures exist in many disciplines that provide the research community with data collected on a regular basis. Examples are the GESIS Panel (Bošnjak et al., 2018), the Understanding America Study (Alattar, Rogofsky, & Messel, 2018), KAMOS (Cho, LoCascio, Lee, Jang, & Lee, 2017), and the LISS Panel (for a description of these infrastructures, see Das, Kapteyn, & Bošnjak, 2018; Weiß et al., 2020). These infrastructures pool resources and increase the objectivity of a survey. A strengthening of such infrastructures is also called for in psychology (Bruder, Göritz, Reips, & Gebhard, 2014). To accompany the establishment and use of such a service at the Leibniz Institute for Psychology (ZPID), the question of possible conditioning effects on the quality of panel data is crucial.

Depending on the underlying mechanism, panel conditioning can have both positive and negative effects on the validity of the data. According to Tourangeau, Rips, and Rasinski's (2000) model of the survey process, at least four steps are necessary to answer a question: understanding the question, retrieving relevant information, processing and evaluating the information, and selecting the appropriate response option. At each of these steps, the survey participants may be affected by previous survey experience.

An experienced participant conceivably understands both the question and the response options better, and is more familiar with the rules of the interview. This reduced cognitive load could lead respondents to stating their opinions more often rather than simply selecting categories such as "don't know," especially in the case of more complex attitudinal questions (Binswanger, Schunk, & Toepoel, 2013). Moreover, a survey can trigger reflection processes beyond the survey and lead to greater attention to and discussion of survey issues (Sturgis, Allum, & Brunton-Smith, 2009). This cognitive stimulation can change attitudes as well as knowledge (e.g., on demographic data such as income; Fisher, 2019) and have an effect on data in subsequent waves.

The way in which the retrieved information is processed and assessed in subsequent waves can also be influenced by previous surveys. In the case of “survey fatigue”, test persons may be prone to satisficing behavior (Krosnick, 1991) or speeding (Schonlau & Toepoel, 2015). The effort of the survey can furthermore be reduced by avoiding follow-up questions, if respondents know the rules of the interview from previous surveys. An example is to answer filter questions negatively (Kreuter, McCulloch, Presser, & Tourangeau, 2011), or to report a smaller social network (Silber et al., 2019). To provide clear evidence for these diverse mechanisms, targeted meta-analyses on each mechanism will be needed in the long run.

Concerning cognitive stimulation, conditioning effects are assumed to a greater extent for sensitive items. Therefore, panel conditioning in sensitive attitudinal and behavioral questions are the focus of the second meta-analysis. An item can be classified as sensitive if it possesses at least one of the following three characteristics (Tourangeau et al., 2000): (1) the question demands a socially undesirable answer (e.g., do you regularly consume illegal drugs?), (2) the question is perceived as intrusive and personal (e.g., how many sexual partners have you had in the last three years?), and (3) the question is particularly relevant in terms of data privacy (e.g., did you earn income in the last year which you did not declare on your taxes?).

The evidence on the effects of panel conditioning on social desirability bias is controversial and might depend on the type of question. Due to the greater familiarity with the interview situation, respondents are less fearful of the consequences and respond more honestly to attitude questions. Thus, experienced participants are less likely to provide socially desirable responses compared to new participants (e.g., Binswanger et al., 2013; Fowler & Floyd, 1995; Nancarrow & Cartwright, 2007; Phillips & Clancy, 1972). The first hypothesis thus assumes a reduction of social desirability bias in attitudinal questions:

H1: The responses of experienced panel participants to sensitive attitudinal questions are less likely to be socially desirable than the responses of new panel participants.

In the case of socially undesirable behavior, it is argued that reporting such actions triggers negative emotions and thereby initiates a reflexive process that leads to the adaptation of responses toward social conformity in subsequent study waves (Baumeister, Vohs, DeWall, & Zhang, 2007). Especially in studies on drug abuse among adolescents and young adults, evidence for the so-called recanting effect is found (Percy, McAlister, Higgins, McCrystal, & Thornton, 2005). Previously reported drug abuse is denied in subsequent waves (Torche, Warren, Halpern-Manners, & Valenzuela, 2012). Similar effects have also been found for other types of sensitive questions assessing behaviors (Fitzsimons & Moore, 2008; Halpern-Manners, Warren, & Torche, 2014; Williams, Block, & Fitzsimons, 2006). This leads to the assumption of an increase in social desirability bias in sensitive behavioral questions with increasing survey frequency:

H2: The responses of experienced panel participants to sensitive behavioral questions are more likely to be socially desirable than the responses of new panel participants.

A general assumption for conditioning effects, which applies to both attitudinal and behavioral items, is the existence of dose effects. That means, that cognitive stimulation is more pronounced, if respondents are conditioned more often, and if the previous survey experience took place more recently. That is:

H3: A stronger conditioning effect, i.e., more pronounced differences between the standardized mean values of experienced and new panel participants, is found with increasing frequency of interviews.

H4: The longer the time interval from the previous survey wave, the weaker the conditioning effect.

The meta-analyses in the following two chapters have been conducted and reported in accordance with the PRISMA statement[1] (Moher et al., 2009).

[1] The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement consists of a checklist with 27 items that have to be reported in a research synthesis and a diagram for reporting the flow of information through the four phases of the literature selection. It helps authors to improve their reporting.

4. Meta-Analysis 1: Response Rates in Psychological Online Surveys

Burgard, T., Bošnjak, M., & Wedderhoff, N. (2020). Response rates in online surveys with affective disorder participants. A meta-analysis of study design and time effects between 2008 and 2019. *Zeitschrift für Psychologie, 228*, 14-24. https://doi.org/10.1027/2151-2604/a000394

4.1. Methods

In the first meta-analysis, time and study design effects on response rates in online surveys are examined. Studies that do not report online survey-only rates or studies where the type of survey is not explicitly reported were excluded. The meta-analysis is restricted to samples of adults with general anxiety disorder or depression, as this population is of growing epidemiological importance (WHO, 2017) and, therefore, of special interest in the domain of psychology. An overview of the inclusion and exclusion criteria for study selection is presented in Table B1.

In the meta-analysis, moderating effects on survey response rates are tested. As potential moderators, basic information from the report, such as publication year and funding, are coded, as well as information on the sample and potentially relevant study design characteristics, as the use of incentives. Finally, a response rate is either given in the report or calculated using the formula defined by the American Association for Public Opinion Research (2016): the number of complete interviews divided by the number of interview attempts.

To search for relevant records, 10 databases were used. The results of the database searches, as well as the search strategies, that differed slightly between the respective databases, are reported as electronic supplements. In addition, the following conference proceedings were searched manually for potentially relevant records: European Survey Research Association (conference year 2017) and American Association for Public Opinion Research (conference years 2016, 2017, 2018).

The outcome is the response rate for each treatment, a relative measure restricted to values between 0 and 1. The treatment is an invitation to participate in an online survey. Data were collected on several levels (report, study, sample, outcomes), but as there is only one response rate per study and sample, and in each report there is only one usable sample reported, there is actually no multilevel data structure. Using the metafor package in R (Viechtbauer, 2010), mixed-effects meta-regressions are computed to test the influence of the year of data collection and the characteristics of the survey design on the response rate. In the mixed-effects model, it is assumed that the true effect sizes may vary between studies. The observed variance in effect sizes is then comprised of the variance of the true effect sizes (the heterogeneity) and the random error. The proportion of true variance in response rates explained by the model used, is assessed using an index analogous to the R^2 index for primary studies (Borenstein et al., 2009).

4.2. Results

The literature search yielded 2,874 potentially relevant records for screening. Of these, 2,769 records could be excluded due to obviously not meeting inclusion criteria. 105 articles were screened as full text. Of these records, 20 were found to be relevant for coding. The main reasons for exclusion of full text articles was missing information on the flow of participants, thus that no response rate could be computed. Figure B1 shows the selection process of literature in detail. Table B2 gives an overview of the characteristics of the studies ultimately included. The major drawback resulting from the small sample of included studies is the lack of studies published before 2008. This was not intended, but is a result of the restriction on samples with anxiety disorder or depression and the requirement of information for the calculation of the response rate.

As a first indication of the time effect on response rate, Figure 7 depicts the bivariate distribution of publication year and response rate. The linear regression line shows the negative relationship between both variables. This relationship is also significant, and the corresponding R^2 is 15.75%. That means that almost 16% of the variance in the response rates of the studies in the meta-analysis can be explained by the publication year. As can be seen in Figure 7, we obviously have one study without nonresponse. Omitting this study from the analysis, as its effect size deviates significantly from the other studies, does not change the conclusions presented in Table 3. Neither the size nor the significance level of effects are affected. This may be due to the small sample size ($n = 30$) of this outlier study and speaks for the robustness of the results.

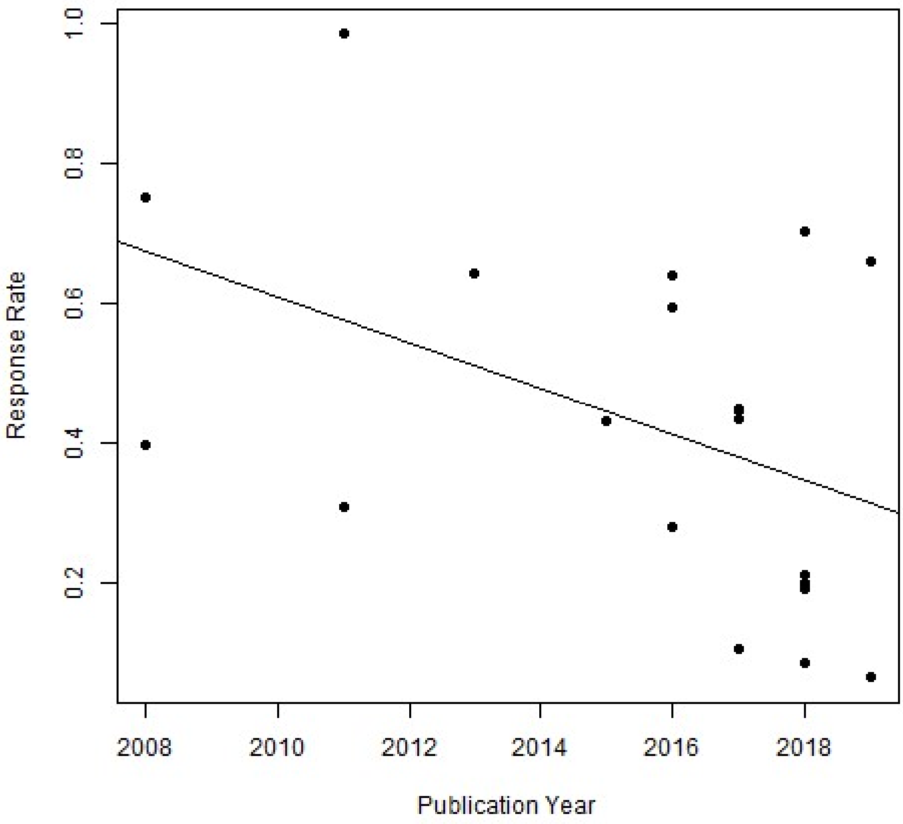


Figure 7: Scatterplot of Publication Year and Response Rate

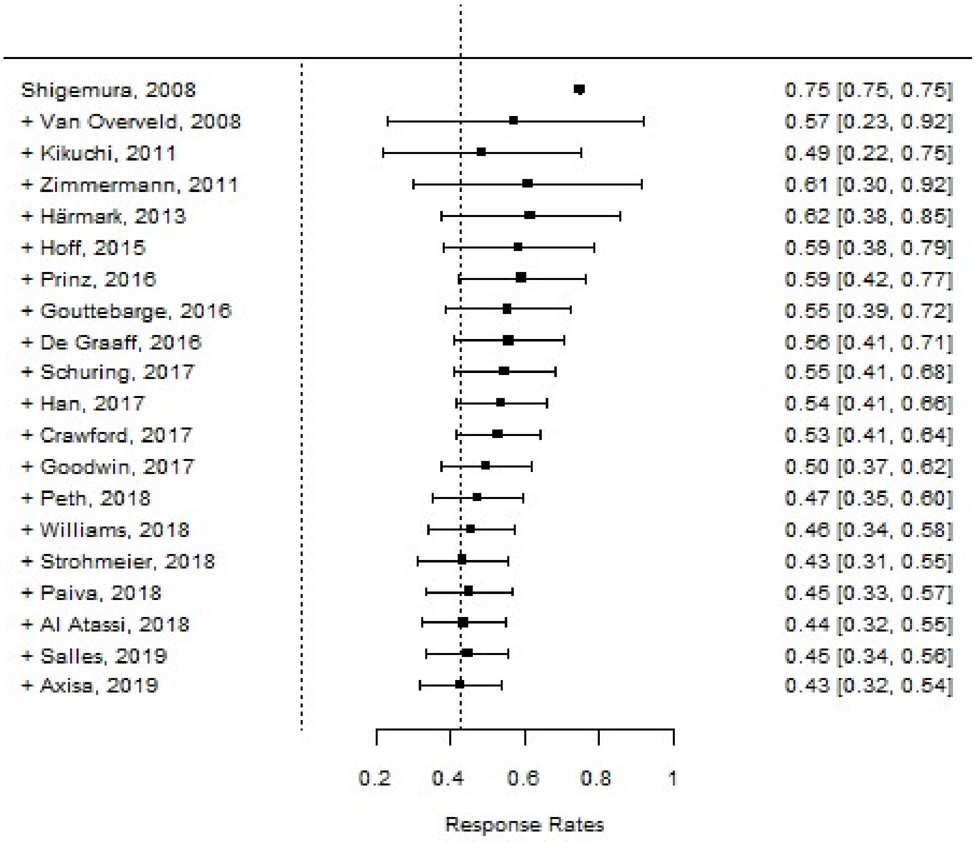


Figure 8: Cumulative Forest Plot of Response Rates over Time

Figure 8 displays the cumulative forest plot of the 20 studies included in the meta-analysis. The studies are sorted chronologically, and the evidence is summed up from study to study. At the beginning (studies published in 2008), the confidence intervals are broad and the overall mean response rates are volatile. The estimates become more precise after few studies. The mean response rate decreased in recent years, confirming the conclusion from the bivariate approach in Figure 7. In 2018, the overall effect finally remains stable and is hardly affected by new evidence. This indicates that evidence to estimate the mean response rate for the studies in this meta-analysis is satisfactory at this point, at least for the time interval examined. The mixed-effects meta-analysis conducted in R reveals an overall response rate of 42.8% for the 20 studies, with a 95% confidence interval between 31.7% and 53.9%.

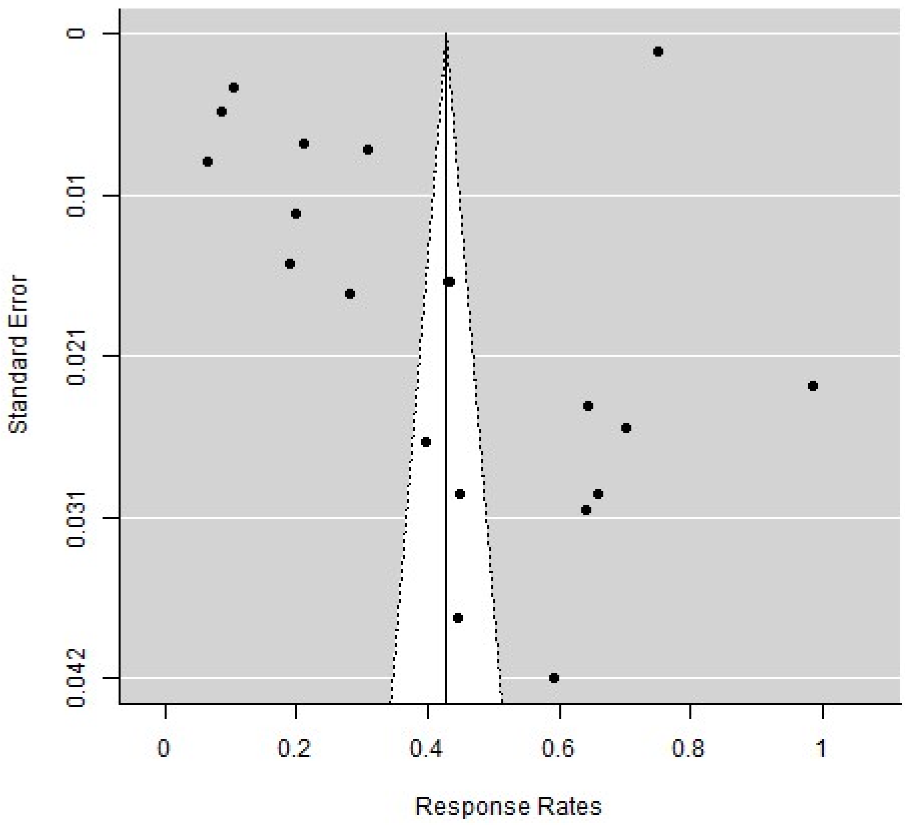


Figure 9: Funnel Plot for Response Rates

The funnel plot in Figure 9 depicts a relationship between the response rates and the standard errors. It seems that the response rates in smaller studies are higher than in larger studies. The result of Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997) confirms that the relationship is significant ($z = 2.29$, $p = 0.0219$). As the response rate is not the outcome of interest in the studies, a publication bias is not the most plausible explanation for this finding. Taking into account the assumptions of positive effects of personal contact to potential participants, an alternative rationale for this relationship might be that the participants in smaller studies were more likely to be contacted personally and that this contact might have resulted in higher response rates.

Table 3: Results of the Meta-Regressions

Moderator	**Full model of study design characteristics**	**Full model study design + additional controls**
Intercept	0.729*** [0.426; 1.033], $p < 0.001$	0.698** [0.271; 1.124], $P = 0.001$
Publication year (H1)	-0.177**[-0.287; -0.068], $p = 0.002$	-0.172** [-0.301; -0.043], $p = 0.009$
Contact mode of invitation (H2) **(e-mail vs. other)**	-0.342* [-0.683; -0.000], $p = 0.050$	-0.317 [-0.742; 0.108], $p = 0.144$
Number of items (H3)	-0.144*[-0.264; -0.024], $p = 0.018$	-0.137. [-0.296; 0.022], $p = 0.092$
Incentives (H4)	-0.055 [-0.295; 0.185], $p = 0.654$	-0.052 [-0.313; 0.209], $p = 0.695$
Funds	-	0.030 [-0.226; 0.286], $p = 0.819$
Mean age sample	-	0.001 [-0.120; 0.121], $p = 0.991$
R^2	28.79%	18.08%

In Table 3, the results of the meta-regressions conducted are reported. There is evidence for an overall decrease in response rates over time. The mode of contacting participants is also relevant. The least personal contact mode was via e-mail. Samples contacted this way showed less willingness to participate in an online survey than samples approached personally, by phone, or mail. A higher number of items in the questionnaire is also related to lower response rates. In total, the moderator variables hypothesized to effect response rates in online surveys explain almost 30% of the variation in response rates. Corroborating the expectations, an effect of incentives is not supported for the population considered in this meta-analysis. Yet this finding does not necessarily mean that incentives have no effect at all. On the contrary, it might also be possible, as previous research indicates, that incentives only have an effect on response rates under certain conditions.

4.3. Sufficiency and Robustness of Results

To test the sufficiency and robustness of the results, we first conducted a power analysis and examined the effects of potentially influential studies on the conclusions of the meta-analysis.

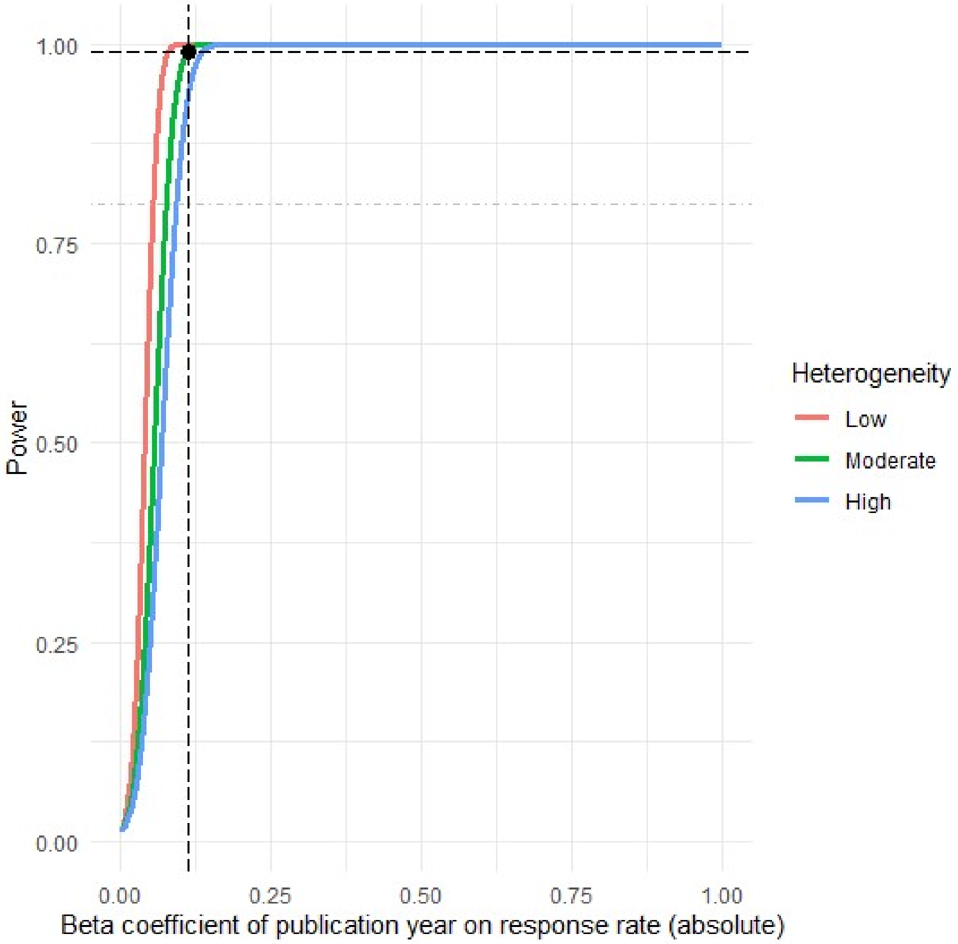


Figure 10: Power of the Meta-Analytic Estimation of the Time Effect on Response Rates

In Figure 10, the statistical power for estimating the time effect on response rates is depicted. The three colored lines represent the statistical power dependent on the size of the true beta coefficient of publication year. The difference between the three lines is the assumption of the underlying heterogeneity between the studies. In case of higher heterogeneity, the advantage of the increase of sample size due to the inclusion of additional studies is mitigated by the increase of variance due to the heterogeneity between the studies.

Due to the large overall sample size in this meta-analysis, power reaches the desired level of 80% at a true beta coefficient of about 0.1 with a significance level of p=0.01. The black dashed line denotes the power level assuming the estimated beta coefficient of publication year in the meta-regression on response rates as the true effect size. Even for this small effect size of -0.11, power reaches almost 100%, even in case of high heterogeneity. The relationship of response rates and publication year cannot be examined in a single study, as there is only one publication year per study and, therefore, no single effect sizes on study level. Thus, an estimation of the hypothetical power in an additional single study does not make sense in this case.

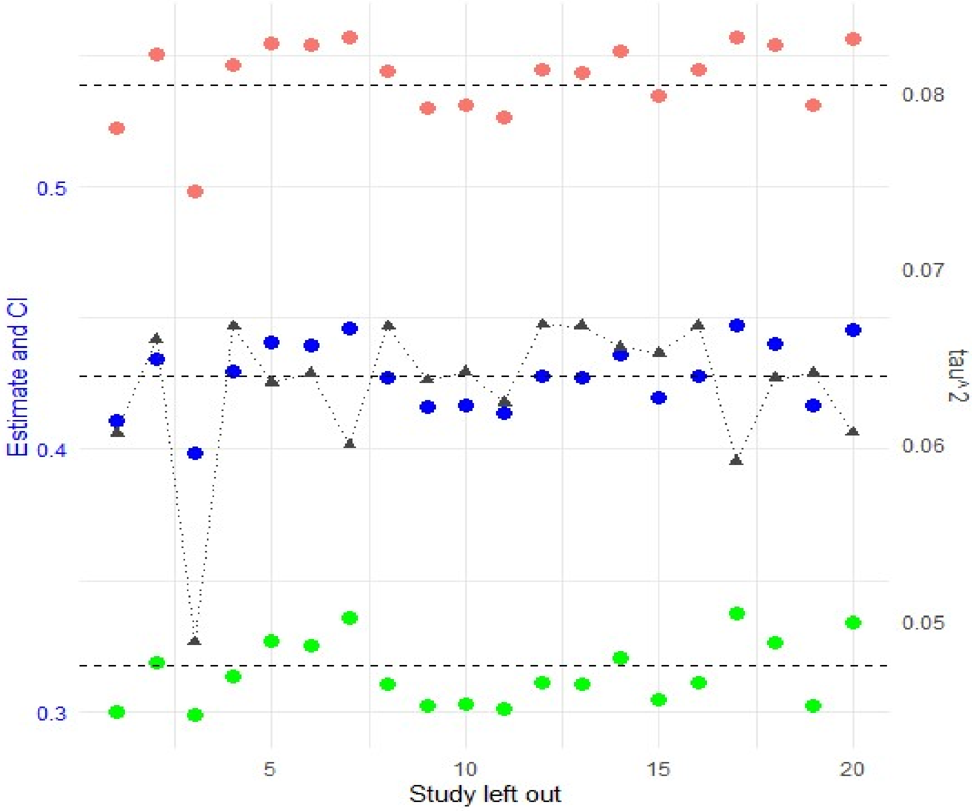


Figure 11: Leave-1-Out Results for Response Rates and Tau^2 Estimates

To detect potentially influential studies, the meta-analytic estimation of the response rate and the between-study variance tau^2 is conducted, leaving out one study at a time. Figure 11 provides information on the resulting mean response rates and confidence interval limits for each estimation. The black dashed lines represent the corresponding estimates in the full dataset with 20 studies. The deviation of the colored points from these estimates, due to leaving out one study each time, shows the influence of this study on the overall results. Obviously, Study 3 can be classified as an outlier. Leaving out only this study results in a reduction of the estimate for the response rate of about 2.5%. In other words, the response rate in Study 3 is markedly higher than in the other studies and, thus, has a considerable impact on the overall estimation.

The scale for the tau^2 estimates is on the right-hand side of Figure 11. Lower values for tau^2 indicate that the study left out contributes massively to the between-study variance. Therefore, leaving out Study 3, which differs considerably from the rest of the studies, results in a lower between-study variance. The influence of single studies on the response rate estimates detects deviating studies with a sufficient sample size to have an impact on the overall estimates. On the contrary, the examination of the tau^2 values also helps to find smaller studies that differ from the remaining studies. This is important, as extreme values in the response rate of small studies might have an impact on meta-regression results, even if the sample size is too low to have a considerable effect on the estimation of the mean response rate.

The examination of tau^2 estimates suggests that Studies 1, 3, 7, 17, and 20 cause the greatest between-study variance. Tau^2 is considerably lower when leaving out these studies suggesting a strong impact on the overall heterogeneity. This finding indicates that the response rates in these studies differ markedly from the other studies. Even if they are too small to influence the estimation of mean response rates, a detrimental impact on moderator analyses is possible. Therefore, the impact of these studies on the results of a meta-regression with publication year and number of items as moderator variables is examined. Mixed-effects models are computed without the potentially influential studies detected in Figure 11 and the results concerning the moderator analyses are compared to the model estimates with all 20 studies, represented again with horizontal black dashed lines in Figure 12.

The amount of variance explained by the model with the two moderators varies markedly depending on which study is excluded from the data. The significance level in the upper panel of Figure 12 refers to the test of the moderators accounting for the variance in response rates. In the full model, this test is significant at the .05 level with an R^2 (amount of heterogeneity accounted for by the model) of about 19%. Leaving out Study 7 or Study 20 results in either an even lower p-value or a slight increase of the share of variance explained. This suggests that the response rate in these studies deviates from the expectations based on the moderator values. Leaving out this seemingly contradictory evidence increases the explanatory value of the model.

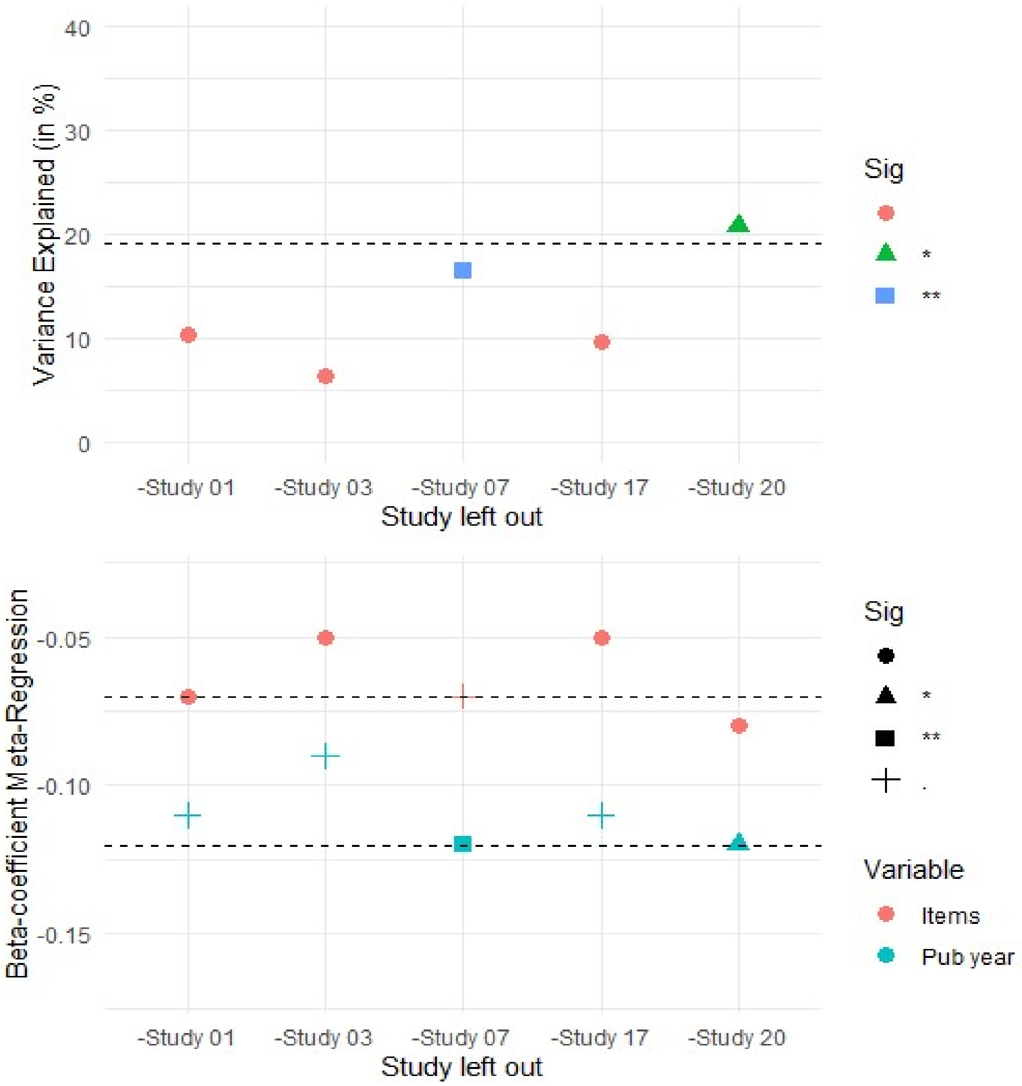


Figure 12: Meta-Regression Results Leaving Out Influential Studies;
*Significance levels: .=p<0.1, *=p<0.05, **=p<0.01*

Leaving out Studies 1, 3, or 17 results in a low explained variance of the model and the moderator test is no longer significant. In line with this finding, it is evident in the lower panel of Figure 12, that excluding these studies reduces the strength of the estimated effects of the moderator variables on the response rates. Leaving out Study 20 has no impact on the estimates of the beta coefficients in the meta-regression. The direction of both moderator effects is the same in all dataset constellations, and the effect of publication year on response rates is significant regardless of which study is excluded. Therefore, the result of declining response rates over time proves to be robust. To draw clear conclusions regarding further potential determinants of response rates in psychological online surveys, more evidence is needed.

4.4. Discussion

To conclude, the hypothesized influences on response rates were mainly confirmed. The mean response rate of 43% is rather high compared to the mean response rates of 34% and 39.6% for online samples found in the meta-analyses of Shih and Fan (2008) and Cook et al. (2000), respectively. This may, however, be due to the restriction to samples of respondents with depression or anxiety disorder. In these studies, (1) many samples were personally recruited from patient lists in hospitals, and (2) the topics of the surveys surrounded the affective disorders the participants were suffering from and therefore, had a high personal relevance to them.

The restrictions to the study sample as well as the necessary information requirements to compute the response rates resulted in a small pool of studies available for the meta-analysis. Due to these limitations concerning the generalizability of the results, the existing evidence on response rates in online surveys for other populations would be of great importance and should be meta-analyzed in the future.

There are several conclusions that can be drawn from the meta-analysis to guide researchers to optimally implement online psychology surveys and achieve high response rates. First of all, clear evidence for the expected decrease in response rates is found, despite the small sample of studies and the short time interval examined. A power analysis, as well as sensitivity analyses prove the robustness of this finding. This result also corroborates existing research of numerous studies on response rates (Brick & Williams, 2013; Krosnick, 1999; Van Horn, Green, & Martinussen, 2009).

Second, the increasing number of items in a survey significantly reduces the response rates. Thus, researchers should strive to keep the burden of the survey rather small. This is in line with previous research showing an effect of survey length on the initial response rate (Galesic & Bošnjak, 2009). Moreover, Mavletova and Couper's (2015) meta-analysis revealed a similar relationship for survey length and breakoff during the survey. However, the effect size of the number of items is rather low, and sensitivity analyses revealed that the effect was not robust when excluding single studies.

Third, when sending invitations to participate in online surveys, the meta-regressions clearly indicate that it is more effective to approach potential participants using more personal forms of contact, such as face-to-face or phone contact. Cook et al.'s (2000) earlier meta-analysis provided evidence for the importance of personal forms of contact to achieve higher response rates. The more recent studies investigated in this meta-analysis support this finding. Thus, to attain high response rates in surveys conducted online, contacting and personally inviting respondents to participate in an offline mode before sending them the survey or the link to the survey is recommended.

A potentially highly relevant moderator is the use of incentives. Previous research has shown, however, that the effectiveness of incentives for increasing response rates depends on the timing and type of incentive (Church, 1993; Pforr et al., 2015). With only four studies reporting the use of incentives in this meta-analysis, the effectiveness of incentives could not be evaluated in detail. More evidence on the use of incentives in online surveys is needed.

Further study design factors that could be included in a future meta-analysis are contact protocols, such as the use of prenotifications and reminders (Bošnjak, Neubarth, Couper, Bandilla, & Kaczmire, 2008; Cook et al., 2000), or the use and design of an advance letter or e-mail (for a meta-analysis on advance letters in telephone surveys, see de Leeuw et al., 2007). These study characteristics are reported less frequently than, for example, the use of incentives or the contact mode for invitation. Hence, the small number of studies (i.e., 20) in this meta-analysis did not allow the examination of these characteristics as moderators. Nonetheless, they may be highly relevant for achieving high response rates and should be included in studies of online surveys in the future.

A more recent research trend that also requires further examination in the context of web surveys is the increase in mobile web surveys. Findings suggest that their breakoff rates are significantly higher than those rates found in surveys that are completed via PC (Mavletova & Couper, 2015). Since the future of web surveys appears to be moving towards implementation via mobile devices, it is vital that research focuses on the optimization of web response rates for these devices.

The presented meta-analysis replicates previous findings and provides recommendations for the initial contact to potential respondents. However, relevant open questions remain, calling for the collection of further evidence. Above all, the inclusion criteria for the target population need to be expanded to achieve more generalizability and more studies to enable further sugroup analyses. This would also enable to collect information on further potential moderators, as contact protocols or the type of device used for the survey. Ideally, the existing dataset could be re-used and simply updated with further evidence. A technical solution to facilitate this process in presented in chapter 6.

5. Meta-Analysis 2: Panel Conditioning in Sensitive Questions

Burgard, T., Wedderhoff, N., & Bošnjak, M. (2020). Konditionierungseffekte in Panel-Untersuchungen: Systematische Übersichtsarbeit und Meta-Analyse am Beispiel sensitiver Fragen [Conditioning Effects in Panel Studies. Systematic Review and Meta-Analysis for Sensitive Items]. *Psychologische Rundschau, 71*, 89-95. https://doi.org/10.1026/0033-3042/a000479

5.1. Methods

In the second meta-analysis, panel conditioning effects for sensitive items are examined. To draw causal conclusions on conditioning effects, (quasi-)experimental studies on the response behavior in panel surveys are relevant for this meta-analysis. For this purpose, the responses of a previously interviewed experimental group and a control group not yet conditioned by interviewing must be compared at the same time and on the same sensitive items assessing behavior or attitudes. Information was extracted from the relevant studies on three levels: (1) general information on the reported study, (2) description of the intervention, and (3) quantitative results of both groups. A complete overview of all utilized coding categories is documented in Table D1.

An initial literature search was conducted in December 2017 with the meta-search engine, CLICsearch. A list of all databases included in CLICsearch is available as an electronic supplement. In addition to "panel conditioning," 15 synonymous search terms were used (see Table D2). Using the relevant articles identified in the initial screening, a manual forward and backward search was performed that reviewed all cited and all referring literature entries.

The calculated effect sizes are standardized mean differences (SMD). During coding and calculation these are arranged in such a way that positive values indicate that experienced panelists respond in a less socially desirable way and negative values reflect the higher social desirability of the responses of participants in the experimental group (i.e., experienced panel participants).

Multilevel meta-regressions, which account for the hierarchical structure of the data (e.g., when several effect sizes from the same study are included), are used to test the hypotheses (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). For the present meta-analysis, a three-level model was chosen. Likelihood ratio tests were conducted to check whether the consideration of the individual levels of analysis in the multilevel model improves the model. For this purpose, the model with all four levels was compared with a reduced model in which one variance component was set to 0 in each case. If the reduced model is significantly worse, the corresponding analysis level should be taken into account.

As the random effects model with four levels of analysis (sample variance, effect sizes, reported samples, study reports) did not improve the model fit, a model with three levels of analysis (sample variance of effect sizes, variance within samples, variance between samples) should be used. This is also plausible since in about half of the 19 studies only results of one sample are reported, so that there are only few study reports in which there can be any variation at all between samples.

The distribution of the variance in the three-level model was determined according to the model of Assink and Wibbelink (2016). Around 5% of the variance was due to sampling error. Variance within samples accounted for 80% of the variance, whereas 14% of the variance was due to differences between samples. Overall, almost 95% of the variance was therefore heterogeneity in the true effects and could be explained by differences between studies. All analyses were performed using the R-package metafor, version 2.0-0 (Viechtbauer, 2010).

5.2. Results

From a total of 2,355 articles initially retrieved, 19 reports were selected for inclusion in the meta-analysis, and these included 85 samples and 154 effect sizes. The corresponding literature selection steps can be found in the form of a PRISMA flowchart (Moher et al., 2009) in Figure D1.

Table D3 provides an overview of some characteristics of the 19 study reports included in the meta-analysis. In addition to author and year of publication, the table also reports the distribution of the total number of samples and group comparisons in the publications. The effects of behavioral questions were examined more frequently (n = 116) than those of attitude questions (n = 38). The mean SMD at the study report level are mostly close to zero, indicating that the difference between control and experimental group is small. This finding speaks for no or only minor conditioning effects. However, there are some studies that suggest medium to strong effects, both in the direction of higher (negative SMD) and lower social desirability (positive SMD).

Table 4: Overall Effects and Effects by Question Type

	Overall effect	Hypothesis 1	Hypothesis 2
Intercept	-0.028*** [-0.042; -0.013]; *p* < .001	-	-
Attitudes (*n* = 38)	-	0.026 [-0.025; 0.078]; *p* = .302	-
Behavior (*n* = 116)	-	-	-0.048*** [-0.056; -0.040]; *p* <.001

Note: n = 154 effect sizes, x=85 samples, effect size: directional SMD (Cohen's d*)*
*Significance levels: .=p<0.1, *=p<0.05, **=p<0.01, ***=p<0.001*

To quantify possible conditioning effects across all studies, meta-analyses were calculated with three levels of analysis. In Table 4, the results of the overall effect, as well as the impact of panel conditioning on social desirability bias are shown. Moreover, the group differences in attitudinal and behavioral items are illustrated in Figure 13 using a caterpillar plot. These are especially useful to display results from meta-analyses containing many effect sizes and are produced using the orchaRd package in R (Nakagawa et al., 2021).

In Hypothesis 1, it was expected that experienced panel participants respond to attitudinal questions in a less socially desirable way. In fact, as expected, the sign of the estimated effect for attitudinal questions is positive albeit nonsignificant. Figure 13 shows, that more than half of the effect sizes from attitudinal items are positive. The confidence interval is displayed by the red summary polygon and including 0. The prediction interval around the point estimate is even wider, showing no clear evidence for the direction of the effect.

Hypothesis 2 postulated higher conformity with social norms for behavioral questions in the experimental group, assuming negative standardized mean differences. In Figure 13, the majority of effect sizes indicates more socially desirable answers to behavior questions. The meta-analytic model in Table 4 reveals a significant effect, confirming hypothesis 2. However, the prediction interval for the mean effect in behavior questions shows a wider range of expected effect sizes, demonstrating substantive heterogeneity (IntHout, Ioannidis, Rovers, & Goeman, 2016).

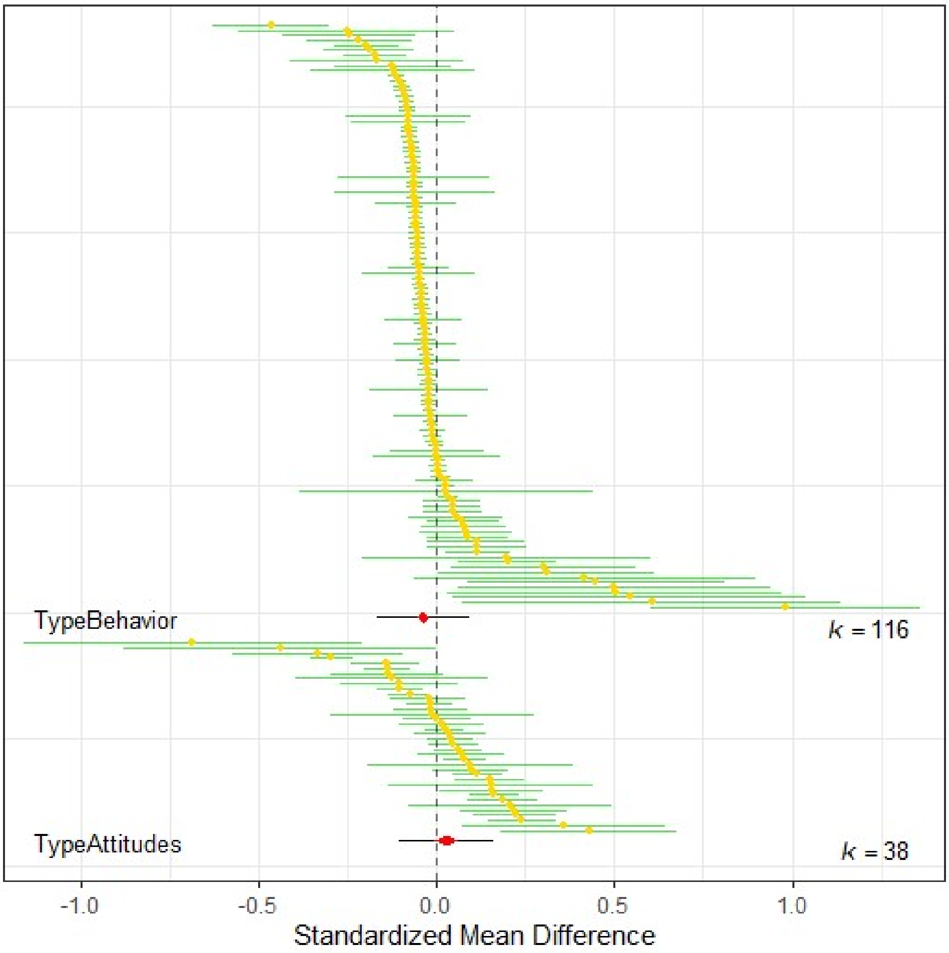


Figure 13: Caterpillar Plot of the Impact of Question Type on Panel Conditioning

Overall, all estimated effects are very small (Ferguson, 2009). No serious conditioning effects are to be expected in panel surveys for both behavioral and attitudinal questions. Thus, serious restrictions in data quality are not expected.

With Hypotheses 3 and 4, dose effects were established. On the one hand, conditioning effects should become stronger through more frequent questioning (Hypothesis 3). On the other hand, it was assumed that the influence of previous interviews on the new measurement should decrease with increasing distance between waves (Hypothesis 4). To test these effects, the absolute SMD is used as the effect size of interest because only the absolute differences between the groups is of significance here and not the direction. The stronger the differences between the groups, the stronger the conditioning effect and vice versa.

Thus, for Hypothesis 3, a positive sign for the effect of the frequency of questioning is expected. As the results in Table 5 show, the effect in the univariate random effects model is close to 0 and not significant. Even in the full model, which also takes into account the timing of the survey and the type of question, the effect of the frequency of previous surveys remains negligible. The coefficient for the interaction effect of attitudes and survey frequency is even negative and thus contradictory to the expectations of the dose effect. Due to the small effect sizes and in light of partly small sample sizes for specific subgroups, especially for attitudinal questions, this finding should not be overestimated.

Table 5: Dose Effects on Absolute SMD (Cohen's d)

	Hypothesis 3	Hypothesis 4	Complete Model	Complete Model with interactions
Intercept	0.061*** [0.048; 0.074]; *p* < .001	0.054*** [0.038; 0.070]; *p* < .001	-	-
Frequency (log)	0.001 [-0.009; 0.010]; *p* = .842	-	0.008. [-0.001; 0.017], *p* = .092	-
Distance between waves (log)	-	0.004** [-0.003; 0.011]; *p* = .285	-0.002** [-0.009; 0.005]; *p* = .595	-
Attitudes	-	-	0.117*** [0.082; 0.153]; *p* < .001	0.421*** [0.183; 0.658]; *p* < .001
Behavior	-	-	0.050*** [0.029; 0.071]; *p* < .001	0.048*** [0.027; 0.069]; *p* < .001
Interactions of the frequency of questioning with the type of question				
Attitudes	-	-	-	-0.077* [-0.148; -0.006]; *p* = .034
Behavior	-	-	-	0.009. [-0.001; 0.018]; *p* = .071
Interactions of the time distances between the interviews with the type of question				
Attitudes	-	-	-	-0.075* [-0.132; -0.018]; *p* = .011
Behavior	-	-	-	-0.001 [-0.008; 0.006]; *p* = .713
Share of heterogeneity accounted for	0%	0%	30.9%	28.6%

Note: n = 154 effect sizes, x = 85 samples, effect size: absolute SMD (Cohen’s d)
*Significance levels: .=p<0.1, *=p<0.05, **=p<0.01, ***=p<0.001*

Hypothesis 4 predicts a weakening of the conditioning effect with a greater time interval between the survey waves. The results of the meta-regressions on the direct effect of the time interval on effect sizes do not support this hypothesis. Only in the model considering interaction effects, a significant reduction of the conditioning effect for longer time intervals in case of attitudinal questions can be stated.

5.3. Sufficiency and Robustness of Results

Statistical Power and Sufficiency of Evidence

A classical power analysis for the the meta-analysis on panel conditioning is depicted in Figure 14. Here, there are three lines representing the different levels of heterogeneity between the studies. Due to the inclusion of 154 effect sizes in the meta-analysis, the overall sample size is large and the power reaches the desired level of 80% for a significance level of 0.01 already at a tiny effect size of less than 0.05. Assuming the absolute value of the meta-analytic estimate of the effect size as the true effect, power reaches almost 100%, although the assumed effect size is only 0.023 and thus negligible from a substantial point of view. In a single study, an effect of this size would not reach sufficient power, even with a large sample.

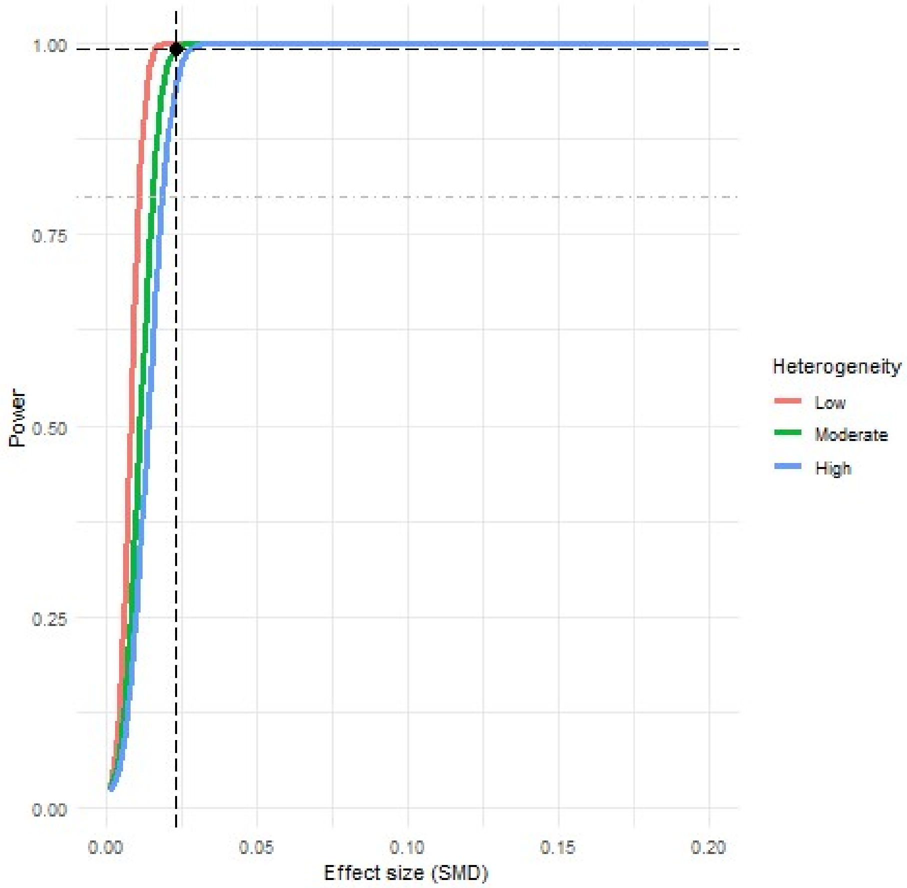


Figure 14: Power of the Meta-Analytic Estimation of Panel Conditioning

Limited power in meta-analyses is typically rather a problem in subgroup analyses. For the relatively few attitude questions (i.e. 38 effect sizes), a true effect size of about the same size as that of behavioral items, might not be sufficient to detect such a small effect. Indeed, a power estimation (as the one shown in Figure 14) for only the subset of effect sizes related to attitudinal items reveals a power estimation of about 33%. This serves as a good example of how quickly meta-analytic moderator analyses can be underpowered, even if the overall synthesized evidence is large and the power for the estimation of overall effects is close to 1. This is in line with theoretical calculations (Hedges & Pigott, 2004) as well as with empirical investigations (Cafri, Kromrey, & Brannick, 2010) on the power of moderator analyses.

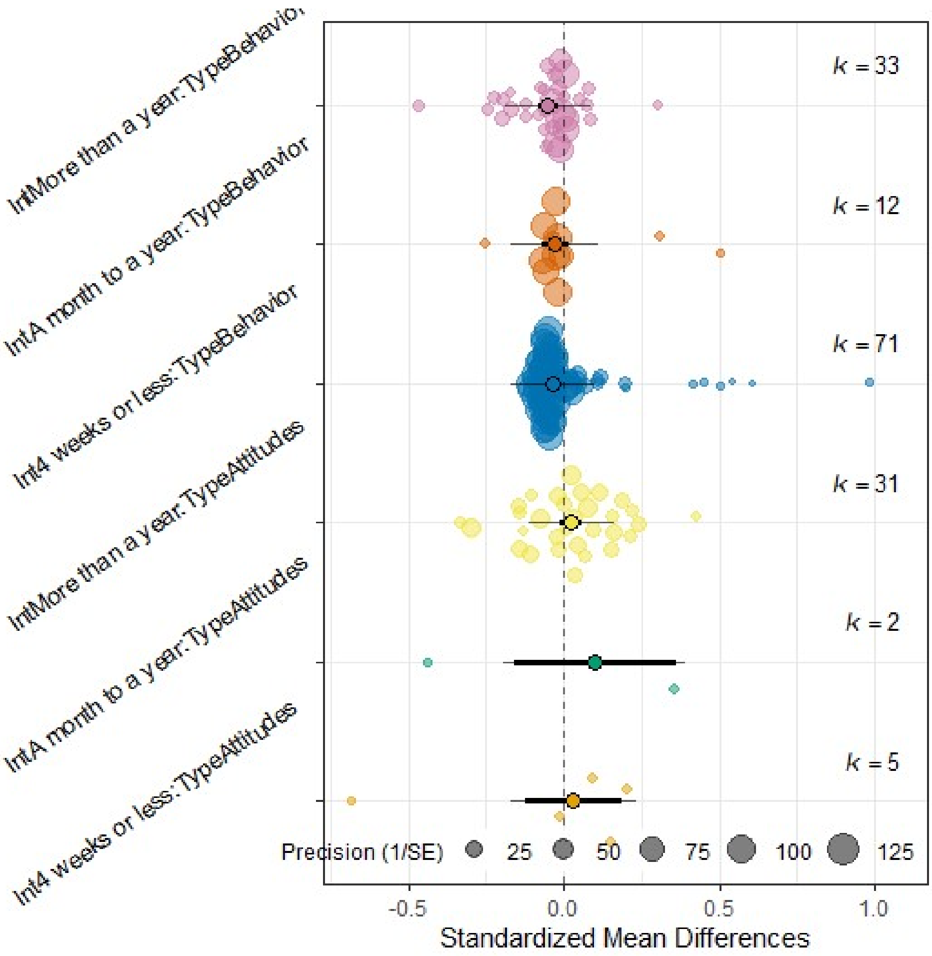


Figure 15: Orchard Plot of Panel Conditioning Effects Grouped by Interval Between Waves and Type of Question

An interesting graphical display for subgroup analyses in meta-analyses with a large number of effect sizes, are orchard plots. As the caterpillar plots, they can be produced with the orchaRd package (Nakagawa et al., 2021). The orchard plot in Figure 15 presents the effect sizes grouped by type of question and time passed since the last surveying for the treatment group. Overall, the tendency of more socially desirable answers to behavior questions and less socially desirable answers for attitudinal questions is confirmed for all subgroups. A relationship between the time interval since the last survey experience and the degree of the differences between experienced and new respondents cannot be detected.

Next to these substantively interesting findings, Figure 15 reveals a lack of evidence for some subgroups. For almost all attitudinal items, the last survey of the experienced panelists was at least one year ago. There are only few observations of time intervals between a month and a year, certainly owing to the lack of planned experiments in the field (Struminskaya, 2016). Most studies use panel refreshments to compare new participants with previously interviewed participants (Warren & Halpern-Manners, 2012). This results in limited variation in frequency of and time intervals between the waves. For example, about half of the 154 group comparisons included in the meta-analysis are based on experimental groups that were previously interviewed only one time. Frequent intervals between the panel waves in the available comparisons for the meta-analysis are one week, one month, or one year. Under these circumstances, evidence is not sufficient for a robust estimation of dose effects.

Publication Bias and p-hacking

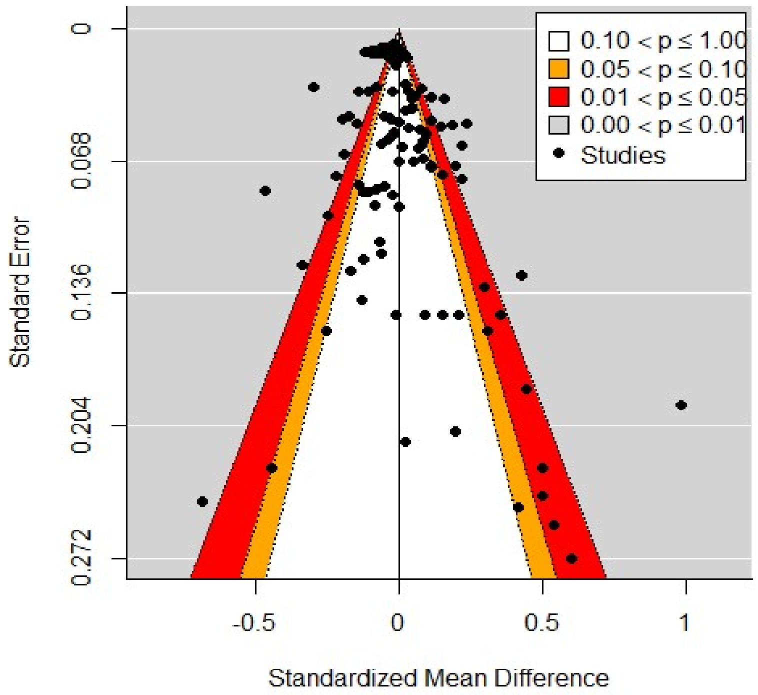
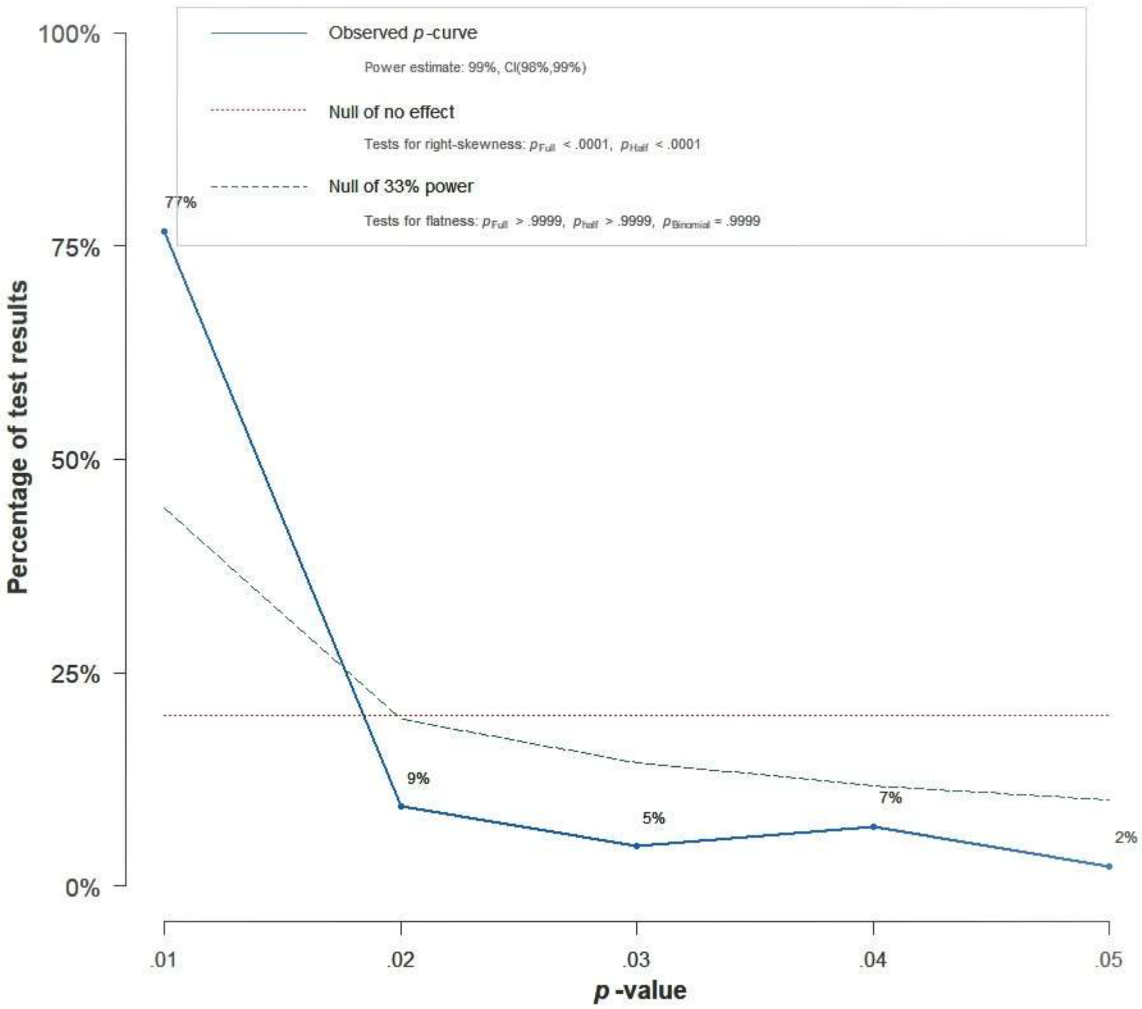


Figure 16: Contour-Enhanced Funnel Plot For Panel Condioning Effects

Concerning potential detrimental effects as publication bias or p-hacking on the reliability of the meta-analytic conclusions, the contour-enhanced funnel plot in Figure 16 shows that many of the effect sizes suggested panel conditioning within the primary studies.

This might suggest, that panel conditioning really occurs under certain circumstances. As there indeed is a lack of non-significant findings within the smaller studies and the Egger's test is significant with z = 5.06 ($p < .0001$), this could also be an indication of publication bias.

A funnel plot using the trim and fill method (Duval & Tweedie, 2000) to augment the observed data for a more symmetric distribution of the effect sizes is depicted in Figure D2. The filled studies are also significant. This suggests that effect sizes speaking for less social desirability bias are observed to a greater extent than those speaking for more social desirable responses. With the four effect sizes added, the test statistic of the Egger's test reduces to z = 2.88 (p=0.0039). The trim and fill method does not only detect asymmetry due to publication bias. The asymmetry can as well reflect the distribution of the true effects.



Note: The observed *p*-curve includes 86 statistically significant ($p < .05$) results, of which 75 are $p < .025$. There were 68 additional results entered but excluded from *p*-curve because they were $p > .05$.

Figure 17: P-Curve of Significant Findings of Panel Conditioning

Another diagnostic for the evidential value of the findings from primary studies is p-curve analysis to detect potential p-hacking. Therefore, the distribution of statistically significant p-values ($p < 0.05$) is examined. It is assumed that this distribution is a function of the real underlying effect. If there is no real effect, the p-values are expected to be uniformly distributed, as the red dotted line in Figure 17 indicates. A left-skewed curve would indicate p-hacking, as researchers may stop collecting more data as soon as findings achieve statistical significance. This would result in mainly large significant p-values.

If there really is an effect, smaller p-values are more likely to be observed, resulting in right-skewed p-curves. The green dashed line shows the shape of a hypothetical curve of 33% power. The curve observed for the meta-analysis on panel conditioning is the blue one. It is even more skewed to the right, indicating the higher statistical power of the underlying results. As the information reported in the box in Figure 17 indicates, the estimated power, based on the observed significant p-values in the meta-analysis, is 99%. This finding substantiates the result of the classical power analysis illustrated in Figure 14.

Influential Studies and Robustness of Findings

To detect potentially influential studies, leave-one-out analyses were conducted. In Figure 18, the estimates and confidence interval limits of conditioning effects are illustrated, leaving out one study at a time in each estimation. The black dashed lines represent the corresponding estimates in the full dataset of 154 effect sizes. The more the colored points deviate from these estimates due to leaving out a study, the greater is the influence of this study on the overall results.

The tau^2 estimates are also depicted with black triangles and the scale on the right-hand side of Figure 18. Lower values indicate that the study left out contributes significantly to the between-study variance. This implies that this study differs considerably from the other studies. Even if the impact on the meta-analytic point estimates for panel conditioning is low for all effect sizes, outlier studies might have a considerable effect on the results of a meta-regression.

Therefore, the most influential studies, with the lowest tau^2 values and the highest deviation from the results of the complete dataset when left out, are the studies 6, 11, 43, 44, 49, 57, and 82. The sensitivity of leaving out these studie on the coefficients of meta-regressions was examined.

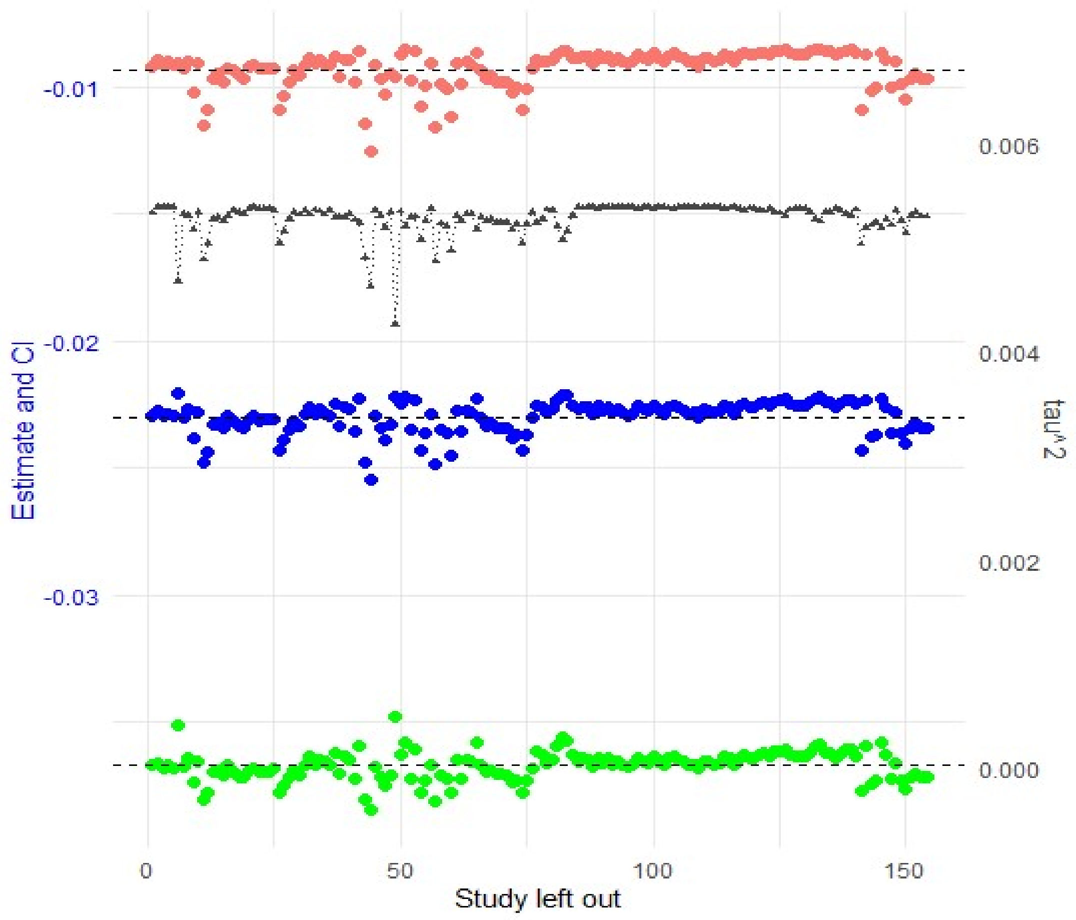


Figure 18: Leave-1-Out Results for Panel Conditioning and Tau^2 Estimates

In Figure D3, the influence of leaving out simultaneously the four outliers biased upwards (11, 43, 44, 57), and respectively the three outliers bias downwards (6, 49, 82) on the estimate of the panel conditioning effects on behavior and on attitudes are examined. A meta-regression model with the type of question as moderator was conducted with the complete dataset. The dashed lines represent the estimate, as well as the lower and upper bounds of the confidence interval. Then the same meta-regression was conducted leaving out the lower, and respectively the upper effect sizes. For attitudes, the positive conditioning effect was significant, when leaving the lower outliers. However, all effect sizes are negligibly small and leaving out influential studies has no relevant influence on the coefficients.

In Figure D4, the influence of all influential studies simultaneously on the effect of frequency on conditioning in attitudes and behavior was examined using a meta-regression with the interaction effect of frequency and type of question. The weak effect of survey frequency is robust and significant in all cases. Overall, no relevant influence of outlier studies on the results of the meta-analysis was detected.

5.4. Discussion

The results of the meta-analysis investigating conditioning effects on sensitive attitudinal and behavioral questions allow the conclusion, that participation in previous survey waves has only a very limited effect on the response behavior in subsequent survey waves. Panel-based surveys are and remain an important data source for psychology. However, as hypothesized, longer time intervals between the survey waves do indeed appear to weaken the conditioning effects, at least for attitudinal items. In general, especially when evaluating dose effects, the limitations of the available database must be taken into account.

The minor effects of panel conditioning found in the present analysis do not yet allow for final conclusions regarding the effects on the quality of panel data. Only the effects on sensitive attitudinal and behavioral items were considered. The database on attitudinal items is already relatively limited, with 38 effect sizes from seven studies. Other types of questions, such as demographic data that could be considered confidential, or nonsensitive filter questions that could be answered incorrectly for strategic reasons to shorten the survey duration, were not examined here.

Panel conditioning is diverse and can have different causes and effects. To make far-reaching statements on the quality and limitations of panel data, the various mechanisms must each be considered individually. This requires further meta-analyses. Potential research questions in the area of panel conditioning are for example differences in the proportion of "Don't know" answers in complex attitudinal questions (Binswanger et al., 2013), the accuracy of responses on demographic items (Fisher, 2019), or the strategic negative answering of filter questions to avoid follow-up questions (Silber et al., 2019).

Especially for the thorough investigation of dose effects, which are particularly important for frequently surveyed populations such as participant pools of online panels, experimental primary studies are required in addition to meta-analyses. These allow a targeted variation of the survey dose in order to close the gaps of previous research and to investigate timing effects more precisely. Further evidence should be used to continuously updated the existing dataset. A technical solution to facilitate this process in presented in the following.

6. Publication of Meta-Analyses for Cumulative Research and Robust Evidence

Burgard, T., Bosnjak, M., & Studtrucker, R. (2021). Community-augmented meta-analyses (CAMAs) in psychology: Potentials and current systems. *Zeitschrift für Psychologie, 229*(1), 15-23. https://doi.org/10.1027/2151-2604 /a000431

Typically, meta-analyses are published exclusively as static snapshots, depicting the evidence in a specific area up to a certain point in time. Moreover, in psychology, published meta-analyses rarely meet common reporting standards, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement, which was conceptualized more than a decade ago, or the more recently suggested MARS (Meta-Analysis Reporting Standards) (Lakens et al., 2017). This practice leads to serious limitations with regard to the reusability of meta-analytic data and the currentness of evidence.

To facilitate and simplify cumulative research and to strengthen the evidence, for example, if practical challenges call for clear recommendations and decisions, we need to think about how to effectively publish our meta-analyses to make knowledge production more efficient. The key challenges for the publication of meta-analyses, therefore, are to make the preexisting research reproducible and to allow the updating of meta-analyses by reusing the information that has been collected up to the point of the most recent meta-analysis.

In the two meta-analyses presented in the previous chapters, a lack of evidence in relation to certain moderator variables was detected and relevant open questions remained. It is, however, crucial that the datasets are usable for their inclusion in further analyses. Having analyzable datasets enables the research community to close research gaps by making use of existing resources rather than investing time and effort into starting an investigation from scratch with literature selection and data extraction. To promote cumulative research of meta-analyses, an appropriate publication format for sustainably usable meta-analytic datasets is needed. In the following, the requirements of an infrastructure for the publication of dynamic meta-analyses are outlined, and a generic tool for psychology, namely PsychOpen CAMA, is presented.

6.1. Challenges and Requirements for the Publication of Meta-Analyses

The first problem often encountered by researchers is the lack of information needed to replicate the results of a meta-analysis. As a response to this problem, Lakens, Hilgard, and Staaks (2016) argue for open meta-analytic data to make meta-analyses dynamic and reproducible. This is important for several reasons. First, having open access meta-analytic data would enable researchers the possibility of examining the sensitivity of the results to subjective decisions that were made in the original process of synthesizing the data, such as the underlying inclusion criteria, statistical models, or use of moderators. Second, an open access meta-analyses register would enable the application of new statistical procedures to existing data and allow testing the effects that these have on the meta-analytic results. Third, open access to existing meta-analyses provides other researchers with special research questions the opportunity to use subsets of the preexisting meta-analytic data (Bergmann et al., 2018).

To overcome the challenge of making meta-analyses reproducible, the four principles (open access, open methodology, open data, and open source) of the open science movement advocated by Kraker, Leony, Reinhardt, and Beham (2011) may be applied. Based on these principles, we derive the following requirements:

1. The transparent documentation of all steps and decisions along the meta-analytic process as presented in Table 1 enables the assessment of possible biases.
2. Common standards for interoperable and usable open data and scripts allow the verification of the results of a review. Subjective decisions may be modified, and new procedures may be applied with minimal effort to check the robustness of the results.

The second problem of static snapshot meta-analyses is the fact that they are only valid for a specific cutoff date (Créquit, Trinquart, Yavchitz, & Ravaud, 2016). Without additional electronic material, a meta-analysis represents the cumulative evidence on a research question up to a certain point in time and may quickly become outdated as soon as new findings from primary studies are published or new methodological or statistical procedures are developed (Shojania et al., 2007). If the data are no longer accessible, the time-consuming process of conducting a meta-analysis starts from scratch.

As requirements to overcome the challenge of updating meta-analyses, we can derive:

1. There is a need for infrastructures that are able to monitor the currentness and validity of meta-analytic evidence and to provide and apply decision rules for the necessity of updates.

2. Open access to data and metadata provides preexisting research a usable and sustainable future. Extracted metadata and coding can be used to update a meta-analysis or even to conduct another meta-analysis on a similar subject with an overlap in the relevant literature.
3. Accumulating science and keeping evidence updated is a cooperative task and the participation in this task has to be supported and incentivized, for example, as proof of achievement instead of, or in addition to, the classical single publication.

6.2. Community-Augmented Meta-Analysis (CAMA) as a Publication Format

Openly available and regularly updated meta-analyses support the efficiency of science. Researchers can get a quick overview on a research field, can use the latest evidence for power analyses and study planning, and may make use of curated information and data to identify research gaps, as understudied moderator variables. As a solution for comprehensive, dynamic, and up-to-date evidence synthesis, Créquit et al. (2016) call for living systematic reviews, that is, high-quality online summaries, that are continuously updated. Similarly, Haddaway (2018) proposes open synthesis.

Actually, a concept for a publication format for meta-analyses that meets these requirements already exists. There are slightly differing forms that have been suggested for this meta-analytical concept including living (Elliott et al., 2017), dynamic (Bergmann et al., 2018), or cloud-based meta-analysis (Bosco, Steel, Oswald, Uggerslev, & Field, 2015). Braver, Thoemmes, and Rosenthal (2014) describe an approach called continuously cumulating meta-analysis (CCMA) to incorporate and evaluate new replication attempts to existing meta-analyses. In the following, the term community-augmented meta-analysis, CAMA for short, is used (Tsuji, Bergmann, & Cristia, 2014). A CAMA is a combination of an open repository for meta-analytic data and an interface offering meta-analytic analysis tools.

The core of a CAMA, as shown in Figure 19, is the data repository, where meta-analytic data contributions from researchers in specific research areas are stored. It serves as a dynamic resource and can be used and augmented by the research community to keep the state of research updated and accumulate knowledge continuously. Tools to replicate and modify analyses with these data are accessible via an open web-based platform, usually encompassing a graphical user interface. For example, examining moderator effects beyond the analyses presented in the original meta-analysis may be conducted. The available evidence from the meta-analyses archived in a CAMA can also be used to improve study planning. Estimates of the expected size of an effect can serve as input for power analyses. The examination of possible relevant moderators can help to identify research gaps and guide the design of new studies (Tsuji et al., 2014).

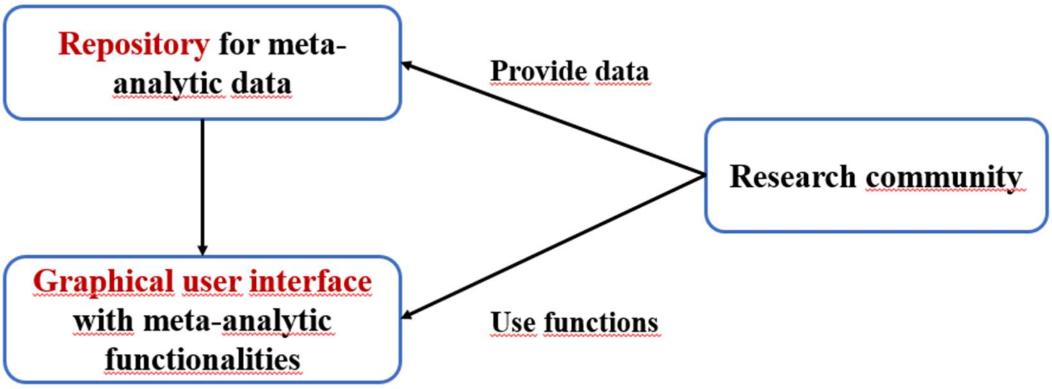

Repository for meta-analytic data
Provide data
Research community
Graphical user interface with meta-analytic functionalities
Use functions

Figure 19: The Basic Structure of a CAMA

6.3. PsychOpen CAMA – A Platform for Cumulative Meta-Analysis in Psychology

Burgard, T., Bosnjak, M., & Studtrucker, R. (under review). PsychOpen CAMA: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods, xx.*

In the following, the focus will be on a system for the publication of dynamic meta-analyses in psychology, a discipline consisting of various domains suffering from small sample sizes such as, for example, developmental research (Bergmann et al., 2018) or neuroscience (Button et al., 2013). There are already several systems and initiatives in psychology, as metalab (Bergmann et al., 2018) or metaBUS (Bosco et al., 2015), aiming at developing an infrastructure for the continuous curation and updating of meta-analytic evidence and, thereby, fulfilling the call to make meta-analyses reproducible and dynamic. For a review and presentation of these systems, see Burgard, Bosnjak, and Studtrucker (2021).

At the Leibniz Institute for Psychology (ZPID), PsychOpen CAMA is currently under development with a first version becoming available in 2021. This service aims to serve the psychological research community as a whole by covering a broad scope of potential research domains. Meta-analyses can be published via the platform to become accessible to and expandable by the community. As Figure 20 shows, PsychOpen CAMA relies on a PHP web application with an OpenCPU server for the R calculations. This improves the scalability of the web application according to the number of users compared to commonly used R shiny architectures. This is of special relevance for a service provided by a research infrastructure institute covering a broad scope of potential research domains, with the possibility to reach more users than rather narrowly specified applications targeted to a small research community.

Original meta-analytic data from users is standardized according to a spreadsheet defining the structure and naming of CAMA data. Standardized data becomes part of a self-maintained R package, that also contains all functions needed for analysis requests offered on the GUI. The user can choose a dataset and request meta-analytic outputs for this data on the GUI. The request is then sent to the server, where the computations are executed. The resulting outputs of the analyses are given back to the user via the GUI.

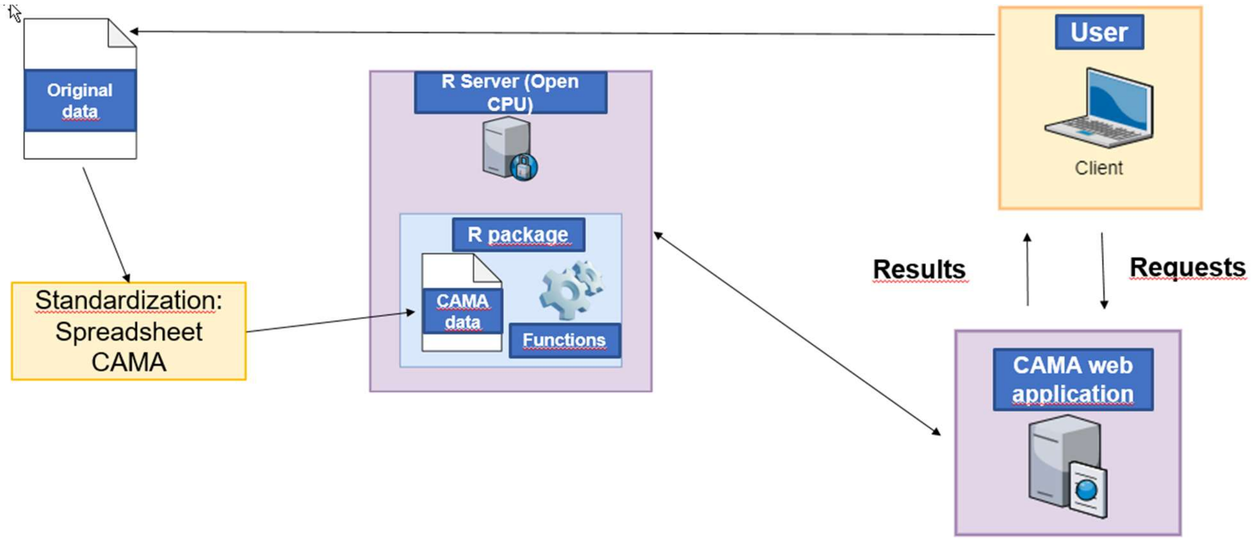


Figure 20: The Architecture of PsychOpen CAMA

Interoperability: Data standardization and metadata

Interoperability enables operational processes and information exchange between different systems. Optimally, standardized identifiers and metadata for all data and digital objects allow for an automated access and use of data by humans and machines (Sansone & Rocca-Serra, 2016). To achieve interoperability of different datasets with the analysis functions used in PsychOpen CAMA, a template for meta-analytic data and machine-readable metadata are used.

As Figure 21 illustrates, the template of PsychOpen CAMA intends the collection of data on different levels. There may be dependencies in the outcome measures of meta-analyses. This might occur, if the effect of an intervention is measured using multiple outcomes, for example competences in different domains. If multiple outcomes are measured using the same study sample, results within a sample might be more similar than between samples. Not accounting for this potential covariance of the outcome measures from the same sample can bias statistical inferences (Van den Noortgate et al., 2013). Data do not have to be nested necessarily. In some meta-analyses, there might only be one outcome measure per sample and report. However, the structure and variable naming enable to distinguish the information levels of the variables and – in case of dependencies – the corresponding information in the metadata automatically triggers the use of a multilevel model in the analysis scripts.

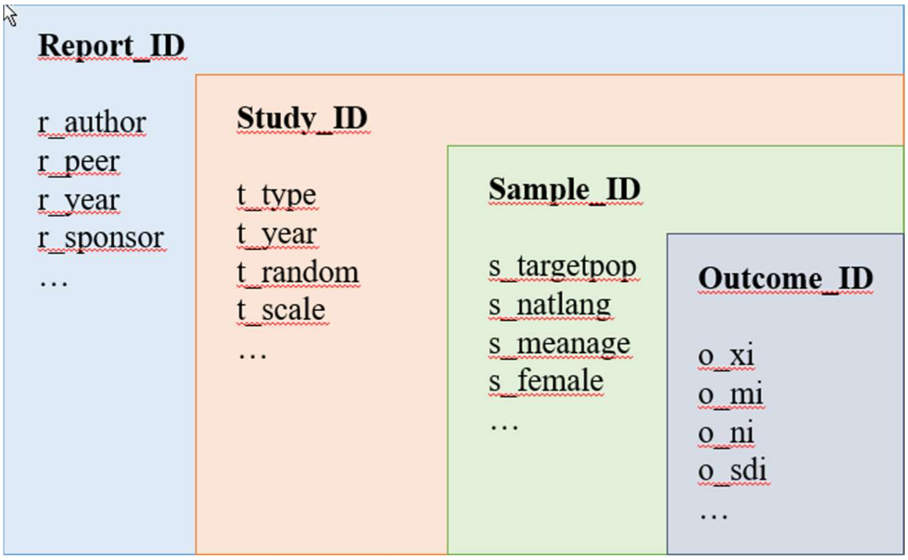


Figure 21: Template for Datasets in PsychOpen CAMA

As a first orientation for a template for basic meta-analytic information on report, study, and sample, the spreadsheet of metalab (Tsuji et al., 2014) served as a starting point. As the meta-analyses for which this spreadsheet serves as a template are all located in the domain of language acquisition and cognitive development, adaptations for other fields of research are necessary. As it is not possible to include each moderator variable that might be relevant in any field, the template is kept rather generic and leaves space for specific adaptations in the form of adding relevant moderators that are not included in the basic template.

The variable names of the outcome data follow the naming of potential measures serving as inputs to compute effect sizes with the escalc() function in metafor (Viechtbauer, 2010). As PsychOpen CAMA operates mainly with this package and effect sizes are computed using the escalc() function, it is convenient to follow the naming and description of metafor for the various outcome information potentially given in a report coded for a meta-analysis.

The naming of metafor is also used in the metadata, where the kind of effect size measure has to be given for each meta-analysis. The options for the ‘measure’ argument are used in various functions in metafor and therefore, following the standards of metafor in this case is also reasonable. Next to the kind of effect size measure, the metadata of each dataset contain the inclusion criteria, relevant moderator variables, research question, nesting of the data, and bibliographic information. The metadata thus serves the purpose of documentation of the methodological conduction of each meta-analysis. Moreover, the metadata are crucial for the automated analyses of the various datasets.

If a user for example selects a certain dataset, the GUI instantly reacts and offers the user the moderators that are available for this dataset. If the user asks for a meta-regression in the next step, the self-maintained R package takes the information on the type of effect size measure and potential dependencies in the data from the metadata to choose the right function and arguments.

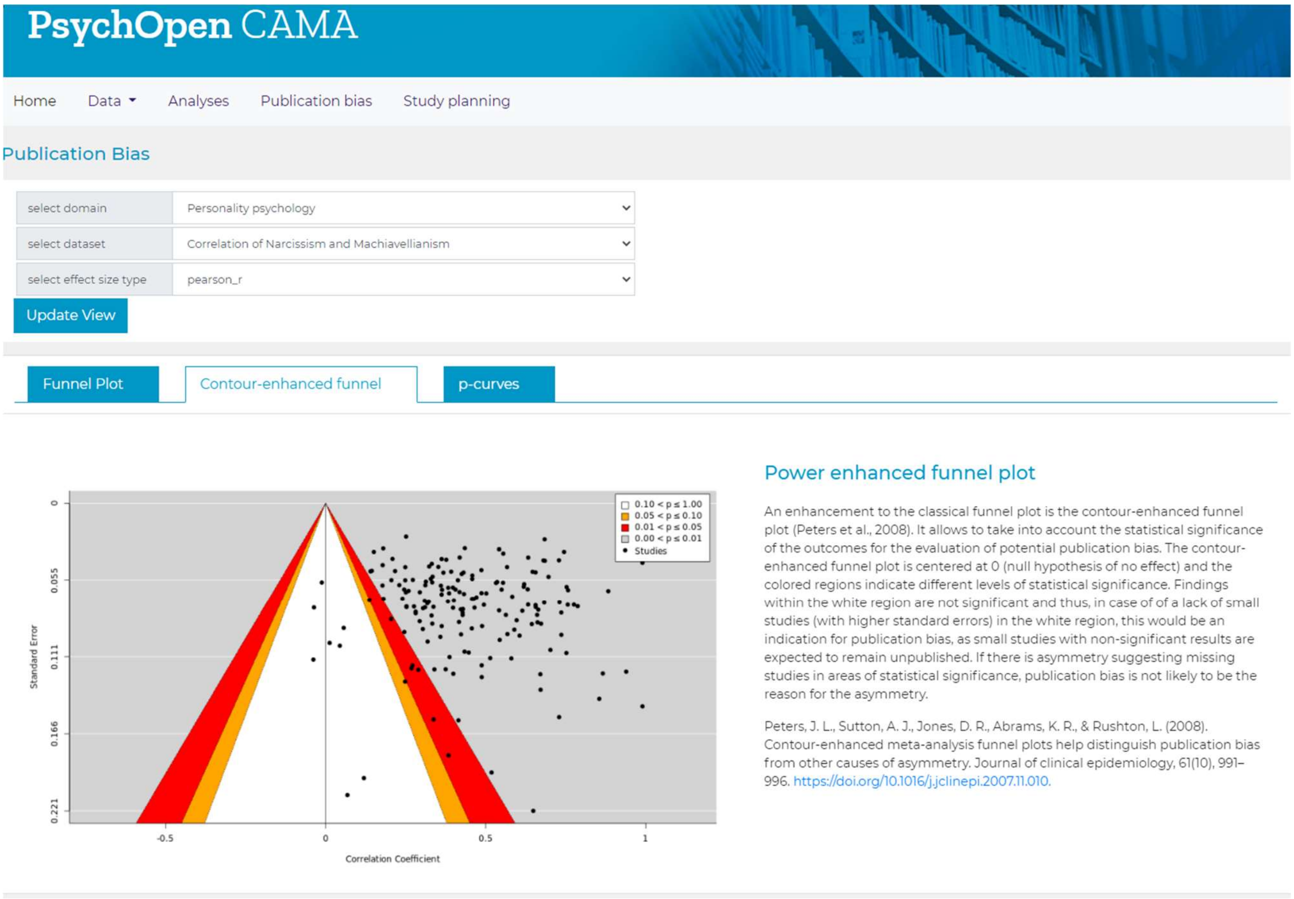
Graphical user interface: Use cases and functionalities

In the first version released, PsychOpen CAMA provides a GUI, offering the user easy access to the results of 14 meta-analytic datasets (February 2021), including the meta-analyses presented in chapters 4 and 5. An intuitive and responsive point-and-click tool makes it easy to explore the data. Interpretation aids to each output make the results comprehensible, even for scientific laypersons. Moreover, these aids are also suited to serve educational purposes.

The menu item “Data” contains a thorough documentation, including bibliographic and methodological information, as well as links to primary studies included in the meta-analyses, and a data table for each dataset. Moreover, a data exploration tool provides a quick overview on the univariate distributions of effect sizes and potentially relevant moderator variables and the corresponding bivariate and trivariate distributions between these variables.

Basic meta-analytic outputs, such as forest plots and meta-analytic estimation, can be found under the item “Analyses”. A dataset and an available effect size type, as well as moderators for inclusion in the meta-regression, can be chosen. If the data are nested, a multilevel model is automatically used to consider dependency in effect sizes. A detailed description of the statistical coefficients is given next to the output to give the user the opportunity to understand the statistics behind and to draw conclusions from the results.

In Figure 22, one of the outputs to assess potential publication bias is displayed, the contour-enhanced funnel plot. A classical funnel plot, the results of an Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997), as well as p-curve analysis (Simonsohn, Nelson, & Simmons, 2014) are also available in this context to give the user the opportunity to assess the evidential value and potential bias of a meta-analysis using different tools.



Publication Bias

select domain	Personality psychology
select dataset	Correlation of Narcissism and Machiavellianism
select effect size type	pearson_r

Power enhanced funnel plot

An enhancement to the classical funnel plot is the contour-enhanced funnel plot (Peters et al., 2008). It allows to take into account the statistical significance of the outcomes for the evaluation of potential publication bias. The contour-enhanced funnel plot is centered at 0 (null hypothesis of no effect) and the colored regions indicate different levels of statistical significance. Findings within the white region are not significant and thus, in case of of a lack of small studies (with higher standard errors) in the white region, this would be an indication for publication bias, as small studies with non-significant results are expected to remain unpublished. If there is asymmetry suggesting missing studies in areas of statistical significance, publication bias is not likely to be the reason for the asymmetry.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. Journal of clinical epidemiology, 61(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010.

Figure 22: Screenshot of the GUI of PsychOpen CAMA

Finally, a study planning tool allows to conduct a prospective power analysis for a potential new study on one of the research questions of the included meta-analyses. Therefore, the meta-analytic estimate of the corresponding meta-analysis is assumed as the true underlying effect size. The sample size and desired significance level are chosen by the user. The tool calculates the expected power of the prospective study, as well as a necessary sample size to achieve a power of 80%. This provides a quick indication of how large a study in a certain domain needs to be to achieve sufficient statistical power and may thus guide researchers in planning new studies.

To conclude, PsychOpen CAMA addresses the open science principles to a great extent by providing data download and open analyses. The risk of bias of meta-analyses is also minimized by giving users the opportunity to include unpublished studies, add unconsidered moderator variables, and modify model specifications.

6.4. Future Challenges for CAMAs and potential solutions for PsychOpen CAMA

With a growing number of publications, efficient accumulation and synthesis of knowledge becomes the key to making scientific results usable and valid thus enabling more informed decisions. The survival time of the synthesized evidence in static meta-analyses, in many cases, is short. To keep this information up-to-date, the publication of meta-analyses in a format allowing the reuse of data already collected and an easy avenue to verify, update, and modify meta-analyses is beneficial for the research community and the public.

A solution to enable dynamic and reusable meta-analyses is CAMA (community-augmented meta-analysis), a new, specialized publication format for meta-analyses. The maintenance of a repository for data in a CAMA, however, is challenging. Depending on the specific domain, a taxonomy for the concepts that are typically assessed, their designations, and the standards for the structure of the collected data has to be defined to allow the combination of research results assessing the same concepts or relations, regardless of how these were originally designated. Standards and taxonomies are needed to ensure that all research results are retrievable and comparable. The workload for the long-term maintenance of the repository can be reduced via crowdsourcing (McCarthy & Chartier, 2017), depending on a research community willing and able to provide relevant data, at best in the desired format.

To support users in the submission of data, the submission assistant of ZPID's archive for digital research objects in psychology, PsychArchives (Weiland, Baier, & Ramthun, 2019) is intended for the submission of data for PsychOpen CAMA. To ensure interoperability of the data with PsychOpen CAMA for the implementation on the platform, manual effort for validity checks will still be needed. Repetitive processes will be automated as far as possible, for example by using notifications in case of new data entries, and scripts for validity checks. But at least for the monitoring of these processes, additional plausibility checks, and necessary corrections in case of erroneous entries, manual effort cannot fully be replaced.

To strategically acquire new data for PsychOpen CAMA, there are various resources that can be used. Research data from primary studies shared in PsychArchives can be used to update corresponding meta-analyses in CAMA. Alternatively, the results of studies or even complete meta-analyses preregistered at ZPID (https://prereg-psych.org/), as well as data collected in ZPID's online or offline laboratory will be used to extend the database for PsychOpen CAMA. For meta-analyses published in one of the journals of PsychOpen (https://www.psychopen.eu/), authors could be asked to share the meta-analytic data of the meta-analysis. The long-term goal of these strategies is to automate these linkages as far as possible to accumulate evidence in PsychOpen CAMA and keep pace with the mass of scientific results produced and published in various domains in psychology.

There are also technical solutions that could be used to automatize processes such as those involved in literature search (e.g., push notifications, database aggregators, automatic retrieval of full texts) and selection (e.g., machine learning classifiers), or extraction of information from published reports (e.g., RobotReviewer for information extraction and risk of bias assessment, Graph2Data for automatic data extraction from graphics) (Thomas et al., 2017). Currently, the software used to carry out these tasks is far from perfect and requires manual supervision. An R package facilitating all the single tasks mentioned at once, from abstract screening to data extraction and reporting of the literature selection process, is 'metagear' (Lajeunesse, 2016). However, the further development of software is a research field in its own right. Algorithms need training data to learn how to decide on the inclusion of studies and extract information from reports. These training data have to be produced by manual effort.

Thus, neither crowdsourcing nor automatization completely solve the problem of the need for continuous curation of cumulative meta-analytic evidence. All relevant processes in the selection, collection, and standardization of research results require human supervision. However, this is an effort providing benefit for the research community as a whole by improving the usability and currentness of existing evidence. As continuously curated meta-analytic evidence also discloses and specifies research gaps, it enables an efficient distribution of research funds for closing these gaps purposefully.

7. Towards Cumulative Meta-Analytic Evidence in the Field of Survey Methodology

The aim of research on survey methods is to guide survey operations for data collection procedures, targeting an optimal combination of survey methods trading off survey errors against survey costs. An overview of the existing meta-analytic evidence on survey errors has shown that only research on nonresponse and measurement errors has been synthesized. There are comprehensible reasons for the lack of meta-analyses on other types of survey errors, such as a lack of primary studies due to nonexistence of relevant information for useful comparisons or a lack of comparability of outcome variables for a quantitative synthesis.

To counteract the current situation and to systematically bridge these research gaps, two meta-analyses focusing on response rates and on panel conditioning have been conducted, taking into account moderator variables lacking meta-analytic evidence so far. For online surveys in psychology, the decrease in response rates found in numerous studies before is confirmed. For the implementation of online surveys, choosing a rather personal approach to contact potential respondents is recommended, for example, by personally addressing them in an invitation mail or by contacting them by phone before sending the online questionnaire. The survey itself should be as short as possible. Mentioning the low survey burden in the invitation letter may motivate potential respondents to participate. Furthermore, low survey burden can reduce breakoff rates during the survey. The main conclusion for panel conditioning is that no significant conditioning effects can be found for sensitive attitudinal and behavioral questions, so that panel-based surveys remain an important data source for research in psychology.

Approaches to assess the sufficiency and robustness of these results have been presented and applied in the context of the two meta-analyses, resulting in the identification of further research needs and evidence gaps. For nonresponse in online surveys, the study sample restrictions in the first meta-analysis were rather strict, limiting the generalizability of the results. Accordingly, evidence for moderators of online survey participation for the general population would be of interest and should be added to the existing database, which to date only includes studies with samples of affective disorder patients. An expansion of the population would also lead to a larger evidence base and improve the power of moderator analyses. For example, more evidence on the effects of incentives in online surveys is needed to have a sufficient pool of studies utilizing incentives to examine the effects of timing and type of incentives. Study characteristics reported less frequently, such as the use of prenotifications or the design of an advance letter, can also be examined in a meta-analysis synthesizing a larger pool of primary studies.

Concerning panel conditioning, there is a noticeable lack of planned experiments to date. Most studies use panel refreshments as a control group. These accidental comparisons do not allow targeted variation of survey frequency and distances between waves. Experimental research on dose effects in panel conditioning would be of interest for a reliable examination of these. Moreover, the second meta-analysis only examined effects on sensitive attitudinal and behavioral items. Other types of items and conditioning effects, such as demographic data or strategic answering of filter questions, were not examined here. The meta-analysis should be augmented accordingly.

New research results should be immediately incorporated into the meta-analytic datasets to enable knowledge accumulation and make use of already existing resources. The recommendations in the field of survey methodology can thus become more reliable and robust, and evidence gaps can be systematically and sustainably closed by adding new evidence to ever-growing knowledge resources.

To publish and curate research findings in meta-analytic datasets and at the same time make analyses replicable and usable by the research community, a publication format for cumulative meta-analyses curated by the community is proposed. The infrastructure intended to serve the psychological discipline as a whole is PsychOpen CAMA, a webapp developed at ZPID and available soon. The results of the meta-analyses presented in this thesis will also be available on the online platform and serve as a starting point for a growing evidence base on survey errors.

Literature

Studies with * were used in the meta-analyses

* Al Atassi, H., Shapiro, M. C., Rao, S. R., Dean, J., & Salama, A. (2018). Oral and maxillofacial surgery resident perception of personal achievement and anxiety: A cross-sectional analysis. *Journal of Oral and Maxillofacial Surgery, 76*, 2532–2539. https://doi.org/10.1016/j.joms.2018.06.018

Alattar, L., Rogofsky, D. & Messel, M. (2018). An introduction to the understanding America study internet panel. *Social Security Bulletin, 78* (2), 13–26. Retrieved from https://www.ssa.gov/policy/docs/ssb/v78n2/v78n2p13.html

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys (9th edition).* AAPOR. Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: a meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, *58*(2), 218-227. https://doi.org/10.2501/JAR-2017-001

* Axinn, W.G., Jennings, E.A., & Copuer, M.P. (2015). Response of sensitive behaviors to frequent measurement. *Social Science Research, 49*, 1-15. https://doi.org/10.1016/j.ssresearch.2014.07.002

* Axisa, C., Nash, L., Kelly, P., & Willcock, S. (2019). Psychiatric morbidity, burnout and distress in Australian physician trainees. *Australian Health Review.* https://doi.org/10.1071/AH18076

Baer, R., Gu, J., Cavanagh, K., & Strauss, C. (2019). Differential sensitivity of mindfulness questionnaires to change with treatment: A systematic review and meta-analysis. *Psychological Assessment, 31*(10), 1247-1263. http://dx.doi.org/10.1037/pas0000744

Banks, G.C., Rogelberg, S.G., Woznyj, H.M., Landis, R.S., & Rupp, D.E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

* Barber, J. S., Gatny, H. H., Kusunoki, Y. & Schulz, P. (2016). Effects of intensive longitudinal data collection on pregnancy and contraceptive use. *International Journal of Social Research Methodology, 19*, 205–222. https://psycnet.apa.org/doi/10.1080/13645579.2014.979717

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine, 7*(9), e1000326. https://doi.org/10.1371/journal.pmed.1000326

Baumeister, R., Vohs, K., DeWall, C., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and*

Social Psychology Review, 11(2), 167-203. https://doi.org/10.1177%2F1088868307301033

Beevers, C., & Meyer, B. (2002). Lack of positive experiences and positive expectancies mediate the relationship between BAS responsiveness and depression. *Cognition and Emotion, 16*, 549–564. https://doi.org/10.1080/02699930143000365

Bergmann, M., & Barth, A. (2018). What was I thinking? A theoretical framework for analysing panel conditioning in attitudes and (response) behaviour. *International Journal of Social Research Methodology, 21*(3), 333–345. https://doi.org/10.1080/13645579.2017.1399622

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Biemer, P.P., & Lyberg, L.E. (2003). Coverage and Nonresponse Error. In Biemer, P.P. & Lyberg, L.E. (Eds.), *Introduction to Survey Quality*, (pp. 63-115). Hoboken, New Jersey: John Wiley & Sons. https://doi.org/10.1002/0471458740.ch3

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Bijttebier, P., Beck, I., Claes, L., & Vandereycken, W. (2009). Gray's reinforcement sensitivity theory as a framework for research on personality-psychopathology associations. *Clinical Psychology Review, 29,* 421–430. https://doi.org/10.1016/j.cpr.2009.04.002

Binswanger, J., Schunk, D., & Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly*, *77*(3), 783–797. https://doi.org/10.1093/poq/nfl030

Bogen, K. (1996). The Effect of Questionnaire Length on Response Rates—A Review of the Literature. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1020–1025.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons.

Bosco, F., Steel, P., Oswald, F., Uggerslev, K., & Field, J. (2015). Cloud-based meta-analysis to bridge science and practice: Welcome to metaBUS. *Personnel Assessment and Decisions, 1*(1). https://doi.org/10.25035/pad.2015.002

Bosco, F. A., Field, J. G., Larsen, K. R., Chang, Y., & Uggerslev, K. L. (2020). Advancing meta-analysis with knowledge-management platforms: Using metaBUS in psychology. *Advances in Methods and Practices in Psychological Science, 3*(1), 124–137. https://doi.org/10.1177/2515245919882693

Bošnjak, M. (2017). Mixed-mode surveys and data quality. Meta-analytic evidence and avenues for future research. In Eifler, S. & Faulbaum, F. (Eds.): *Methodische Probleme* *von Mixed-Mode-Ansätzen in der Umfrageforschung*, (pp. 11–25). Wiesbaden: Springer Fachmedien.

Bošnjak, M. (2018). Evidence-Based Survey Operations: Choosing and Mixing Modes. In Vanette, D. L. & Krosnick, J. A. (Eds.), *The Palgrave Handbook of Survey Research*, (pp. 319–330). Cham: Springer.

Bošnjak, M., & Danner, D. (2015). Survey participation and response. *Psihologija, 48*(4), 307–310. https://doi.org/10.2298/PSI1504307B

Bošnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel. *Social Science Computer Review, 36*(1), 103–115. https://doi.org/10.1177/0894439317697949

Bošnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmire, L. (2008). Prenotification in web-based access panel surveys – The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review, 26*, 213–223. https://doi.org/10.1177/0894439307305895

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics, 29*(3), 329–353. https://doi.org/10.2478/jos-2013-0026

Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *Annals of the American Academy of Political and Social Science, 645*(1), 36–59. https://doi.org/10.1177/0002716212456834

Bricker, J. (2014). Survey incentives, survey effort, and survey costs. *Finance and Economics Discussion Series,* 2014(74), 1–36. https://doi.org/10.17016/feds.2014.074

* Bridge, G. G. (1977). Interviewing Changes Attitudes – Sometimes. *Public Opinion Quarterly, 41* (1), 56–64. https://doi.org/10.1086/268352

Bruder, M., Göritz, A., Reips, U.-D., & Gebhard, R. (2014). Ein national gefördertes Onlinelabor als Infrastruktur für die psychologische Forschung. *Psychologische Rundschau*, *65*, 75–85.

Burgard, T., Bosnjak, M., & Studtrucker, R. (2021). Community-augmented meta-analyses (CAMAs) in Psychology: Potentials and current systems. *Zeitschrift für Psychologie, 229*(1), 15-23. https://doi.org/10.1027/2151-2604 /a000431

Burgard, T., Bosnjak, M., & Studtrucker, R. (under review). PsychOpen CAMA: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods, xx*.

Burgard, T., Bošnjak, M. & Wedderhoff, N. (2020). Response rates in online surveys with affective disorder participants. A meta-analysis of study design and time effects between

2008 and 2019. *Zeitschrift für Psychologie, 228*, 14-24. https://doi.org/10.1027/2151-2604/a000394

Burgard, T., Wedderhoff, N., & Bošnjak, M. (2020). Konditionierungseffekte in Panel-Untersuchungen: Systematische Übersichtsarbeit und Meta-Analyse am Beispiel sensitiver Fragen. *Psychologische Rundschau, 71*, 89-95. https://doi.org/10.1026/0033-3042/a000479

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. & Munafò, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. https://doi.org/10.1038/nrn3475

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*(2), 239–270. https://doi.org/10.1080/00273171003680187

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes–no answers versus check-all in self-administered modes. A systematic review and analyses. *International Journal of Market Research, 57*(2), 203–223. https://doi.org/10.2501%2FIJMR-2015-014a

Card, N. A. (2012). *Applied meta-analysis for social science research*. Guilford, New York.

Cehovin, G., Bošnjak, M., & Lozar Manfreda, K. (2018). Meta-analyses in survey methodology: A systematic review. *Public Opinion Quarterly, 82*(4), 641–660. https://doi.org/10.1093/poq/nfy042

Cho, Y.I., Johnson, T.P., & Vangeest, J.B. (2013). Enhancing surveys of health care professionals. A meta-analysis of techniques to improve response. *Evaluation & the Health Professions, 36*, 382–407. https://doi.org/10.1177/0163278713496425

Cho, S. K., LoCascio, S. P., Lee, K.-O., Jang, D.-H. & Lee, J. M. (2017). Testing the Representativeness of a Multimode Survey in South Korea: Results from KAMOS. *Asian Journal for Public Opinion Research, 4* (2), 73–87. https://doi.org/10.15206/ajpor.2017.4.2.73

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates. A meta-analysis. *Public Opinion Quarterly, 57*, 62–79. https://doi.org/10.1086/269355

Clarke, M., Brice, A., & Chalmers, I. (2014). Accumulating research: A systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. *PLoS ONE, 9*(7), e102670. https://doi.org/10.1371/journal.pone.0102670

* Clinton, J.D. (2001). Panel Bias from Attrition and Conditioning. A Case Study of the Knowledge Networks Panel. *The American Association for Public Opinion Research (AAPOR)*, 56th Annual Conference.

Coffé, H., & Van Den Berg, J. (2017). Understanding shifts in voting behavior away from and towards radical right populist parties: The case of the PVV between 2007 and 2012. *Comparative European Politics, 15*(6), 872–896. https://doi.org/10.1057/s41295-016-0008-3

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying open-ended reports: Factors affecting the reliability of occupation codes. *Journal of Official Statistics, 32*(1), 75–92. https://doi.org/10.1515/JOS-2016-0003

Cook, C., Heath, F. & Thompson, R. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement 60*, 821–36.

* Coombs, Lolagene C. (1973). Problems of contamination in panel surveys: A brief report on an independent sample, Taiwan, 1970. *Studies in Family Planning, 4*(10), 257-261. https://doi.org/10.2307/1964738

Cooper, H. (1998). *Synthesizing research* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Cooper, H. (2017). *Research Synthesis and meta-analysis. A step-by-step approach*. Thousand Oaks, CA: Sage Publications.

Cornesse, C., Bošnjak, M. (2018). Is there an association between survey characteristics and representativeness? A meta-analysis. *Survey Research Methods, 12*(1), 1-13. https://doi.org/10.18148/srm/2018.v12i1.7205

Corr, P. J. (2002). J. A. Gray's reinforcement sensitivity theory: Tests of the joint subsystems hypothesis of anxiety and impulsivity. *Personality and Individual Differences, 33*, 511–532. https://doi.org/10.1016/S0191-8869(01)00170-2

* Crawford, N. M., Hoff, H. S., & Mersereau, J. E. (2017). Infertile women who screen positive for depression are less likely to initiate fertility treatments. *Human Reproduction, 32*, 582–587. https://doi.org/10.1093/humrep/dew351

Créquit, P., Trinquart, L., Yavchitz, A., & Ravaud, P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: The example of lung cancer. BMC Medicine, 14(8). https://doi.org/10.1186/s12916-016-0555-0

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177%2F0956797613504966

Daikeler, J., Bošnjak, M., & Lozar Manfreda, L. (2020). Web surveys versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology, 8*(3), 513-539. http://dx.doi.org/10.1093/jssam/smz008

Das, M., Kapteyn, A., & Bošnjak, M. (2018). Open probability-based panel infrastructures. In Vannette, D.L. & Krosnick, J.A. (Eds.), *The Palgrave Handbook of Survey Research*, (pp. 199-209). Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_25

* Das, M., Toepoel, V., & van Soest, A. (2011). Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys. *Sociological Methods & Research, 40*(1), 32–56. https://doi.org/10.1177/0049124110390765

David, M. C., & Ware, R. S. (2014). Meta-analysis of randomized controlled trials supports the use of incentives for inducing response to electronic health surveys. *Journal of Clinical Epidemiology, 67*(11), 1210–1221. https://doi.org/10.1016/j.jclinepi.2014.08.001

DeBell, M. (2013). Harder than it looks: Coding political knowledge on the ANES. *Political Analysis, 21*(4), 393–406. https://doi.org/10.1093/pan/mpt010

* De Graaff, A. A., Van Lankveld, J., Smits, L. J., Van Beek, J. J., & Dunselman, G. A. (2016). Dyspareunia and depressive symptoms are associated with impaired sexual functioning in women with endometriosis, whereas sexual functioning in their male partners is not affected. *Human Reproduction, 31,* 2577–2586. https://doi.org/10.1093/humrep/dew215

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., & Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys a meta-analysis. *Public Opinion Quarterly, 71*(3), 413–443. https://doi.org/10.1093/poq/nfm014

De Leeuw, E. D., & De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little, R. J. (Eds.), *Survey Nonresponse,* (pp. 41–54). New York: John Wiley & Sons.

Dewaele, A., Caen, M., & Buysse, A. (2014). Comparing survey and sampling methods for reaching sexual minority individuals in Flanders. *Journal of Official Statistics, 30*(2), 251–275. https://doi.org/10.2478/jos-2014-0016

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The Tailored Design Method* (4th ed.). New York: John Wiley & Sons.

Dodou, D., & de Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487–495. https://doi.org/10.1016/j.chb.2014.04.005

Dolders, M., Zeegers, M., Groot, W., & Ament, A. (2006). A meta-analysis demonstrates no significant differences between patient and population preferences. *Journal of Clinical Epidemiology, 59*(7), 653–664. https://doi.org/10.1016/j.jclinepi.2005.07.020

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89−98. https://doi.org/10.2307/2669529

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical research ed.), 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., & Thomas, J. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology, 91*, 23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE, 4*(5). https://doi.org/10.1371/journal.pone.0005738

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538. https://doi.org/10.1037/a0015808

Fiedler, K. (2018). Wo sind die wissenschaftlichen Standards für hochwertige Replikationsforschung? *Psychologische Rundschau, 69*(1), 45–56. https://doi.org/10.1026/0033-3042/a000388

Fisher, P. (2019). Does Repeated Measurement Improve Income Data Quality? *Oxford Bulletin of Economics and Statistics*, *81*(5), 989–1011. https://doi.org/10.1111/obes.12296

* Fitzsimons, G. J., John, D. S. N., Nunes, J. C. & Williams, P. (2007). License to Sin: The Liberating Role of Reporting Expectations. *Journal of Consumer Research, 34* (1), 22–31. https://dx.doi.org/10.1086%2F513043

Fitzsimons, G. J., & Moore, S. G. (2008). Should we ask our children about sex, drugs and rock & roll? Potentially harmful effects of asking questions about risky behaviors. Journal of consumer *psychology: the official journal of the Society for Consumer Psychology, 18*(2), 82–95. https://doi.org/10.1016/j.jcps.2008.01.002.

Fowler, J., Floyd J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Fritz, M., & Koch, M. (2019). Public support for sustainable welfare compared: Links between attitudes towards climate and welfare policies. *Sustainability, 11*(15), 4146. MDPI AG. http://dx.doi.org/10.3390/su11154146

Fuchs, M. (2010). Improving research governance through use of the Total Survey Error Framework. In RatSWD (Ed.), *Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences,* (pp. 471–486). Opladen: Barbara Budrich. https://www.jstor.org/stable/j.ctvbkk43d.29

Galesic, M., & Bošnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349–360. https://doi.org/10.1093/poq/nfp031.

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Akl, E., Beyene, J., Chang, S., Churchill, R., Dearness, K., Guyatt, G., Lefebvre, C., Liles, B., Marshall, R., García, L., Mavergames, C., Nasse, M., Qaseem, A., Sampson, M., Takwoingi, Y., Thabane, L., Trivella, M., Tugwell, P., Welsh, E., Wilson, E., & Schünemann, H. (2016). When and how to update systematic reviews: Consensus and checklist. *BMJ (Online), 354*, 1–10. https://doi.org/10.1136/bmj.i3507

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft: Strategische Ausrichtung von Forschungsprozessen als Weg aus der Replikationskrise. *Psychologische Rundschau, 69*(1), 22-36. https://doi.org/10.1026/0033-3042/a000384

Gnambs, T., & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavior Research Methods, 47*(4), 1237–1259. https://doi.org/10.3758/s13428-014-0533-4

Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the Candor Hypothesis. *Assessment, 24*(6), 746–762. https://doi.org/10.1177/1073191115624547

* Goodwin, G. M., Price, J., De Bodinat, C., & Laredo, J. (2017). Emotional blunting with antidepressant treatments: A survey among depressed patients. *Journal of Affective Disorders, 221,* 31–35. https://doi.org/10.1016/j.jad.2017.05.048

* Gouttebarge, V., Jonkers, R., Moen, M., Verhagen, E., Wylleman, P., & Kerkhoffs, G. (2017). The prevalence and risk indicators of symptoms of common mental disorders among current and former Dutch elite athletes. *Journal of Sports Sciences, 35*, 2148–2156. https://doi.org/10.1080/02640414.2016.1258485

Graham, A.L., Hasking, P., Clarke, D. & Meadows, G. (2015). How people with depression receive and perceive mental illness information: Findings from the Australian National Survey of Mental Health and Wellbeing. *Community Mental Health Journal 51*, 994–1001. https://doi.org/10.1007/s10597-015-9900-6

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*(4), 475–495. https://doi.org/10.1086/269338

Groves, R.M., & Couper, M. (1998). *Household Survey Nonresponse*. New York: John Wiley & Sons.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. New York: John Wiley & Sons.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Haddaway, N. R. (2018). Open synthesis: On the need for evidence synthesis to embrace open science. *Environmental Evidence, 7*(1), 4–8. https://doi.org/10.1186/s13750-018-0140-4

Halmdienst, N., & Radhuber, M. (2018). Quality management in social sciences research. Quality management in the Survey of Health, Ageing and Retirement in Europe (SHARE). *Austrian Journal of Political Science, 47*(2), 49–60. https://doi.org/10.15203/ozp.2021.vol47iss2

* Halpern-Manners, A., Warren, J. R., & Torche, F. (2014). Panel conditioning in a longitudinal study of illicit behaviors. *Public Opinion Quarterly*, *78*(3), 565–590. https://doi.org/10.1093/poq/nfu029.

* Halpern-Manners, A., Warren, J. R., & Torche, F. (2017). Panel Conditioning in the General Social Survey. *Sociological Methods & Research, 46*(1), 103–124. https://doi.org/10.1177/0049124114532445

* Han, K., Bohnen, J., Peponis, T., Martinez, M., Nandan, A., Yeh, D. D., . . . Kaafarani, H. M. (2017). Surgeon as the second victim? Results of the Boston Intraoperative Adverse Events Surgeons' Attitude (BISA) Study. *Journal of the American College of Surgeons, 224,* 1048–1056. https://doi.org/10.1016/j.jamcollsurg.2016.12.039

* Harmark, L., Van Puijenbroek, E., & Van Grootheest, K. (2013). Intensive monitoring of duloxetine: Results of a web-based intensivemonitoring study. *European Journal of Clinical Pharmacology, 69*, 209–215. https://doi.org/10.1007/s00228-012-1313-7

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426–445. https://doi.org/10.1037/1082-989X.9.4.426

* Hoff, H. S., Crawford, N. M., & Mersereau, J. E. (2015). Mental health disorders in infertile women: Prevalence, perceived effect on fertility, and willingness for treatment for anxiety and depression. *Fertility and Sterility, 104*, e357. https://doi.org/10.1016/j.fertnstert.2015.07.1113

Hohn, R. E., Slaney, K. L., & Tafreshi, D. (2019). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology, 9*, 1–15. https://doi.org/10.3389/fpsyg.2018.02667

Hox, J. J., & de Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity, 28*, 329–344. https://doi.org/10.1007/BF01097014

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open, 6*(7), e010247. https://doi.org/10.1136/bmjopen-2015-010247

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods, 8*(3), 290–302. https://doi.org/10.1002/jrsm.1240

Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal, 323*(7), 42–46. https://doi.org/10.1136/bmj.323.7303.42

Kalton, G., Kasprzyk, D., & McMillen, D. B. (1989). Nonsampling errors in panel surveys. In Kasprzyk, D., Duncan, G., Kalton, G. & Singh, M. (Eds.), *Panel surveys,* (pp. 249–270). New York, NY: Wiley.

Katz, D. L., Williams, A. L., Girard, C., & Goodman, J. (2003). The evidence base for complementary and alternative medicine: methods of evidence mapping with application to CAM. *Alternative therapies in health and medicine, 9*(4), 22.

Kepes, S., & McDaniel, M. A. (2015). The Validity of Conscientiousness Is Overestimated in the Prediction of Job Performance. *PloS one, 10*(10), e0141468. https://doi.org/10.1371/journal.pone.0141468

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4.

* Kikuchi, T., Uchida, H., Suzuki, T., Watanabe, K., & Kashima, H. (2011). Patients' attitudes toward side effects of antidepressants: An Internet survey. *European Archives of Psychiatry and Clinical Neuroscience, 261*, 103–109. https://doi.org/10.1007/s00406-010-0124-z

Kopp, J., & Richter, N. (2016). Social mechanisms and empirical research in the field of sociology of the family: The case of separation and divorce. *Analyse & Kritik, 1*, 121–148. https://doi.org/10.1515/auk-2016-0107

Kormos, C., & Gifford, R. (2014). The validity of self-report measures of pro-environmental behavior: A meta-analytic review. *Journal of Environmental Psychology, 40*, 359–371. https://doi.org/10.1016/j.jenvp.2014.09.003

Kraker, P., Leony, D., Reinhardt, W., & Beham, G. (2011). The case for an open science in technology enhanced learning. *International Journal of Technology Enhanced Learning, 3*(6), 643–654. https://doi.org/10.1504/IJTEL.2011.045454

* Kraut, R. E. & McConahay, J. B. (1974). How Being Interviewed Affects Voting: An Experiment. *Public Opinion Quarterly, 37* (3), 398. https://psycnet.apa.org/doi/10.1086/268101

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research, 40*(1), 88–104. https://doi.org/10.1177/0049124110392342

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology, 50*(1), 537–567. doi:10.1146/annurev.psych.50.1.537

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review, 25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P. K. (2016). From social to sale: The effects of firm-generated content in social media on customer behavior. *Journal of Marketing, 80*(1), 7–25. https://doi.org/10.1509/jm.14.0249

Lajeunesse, M. J. (2016). Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution, 7*(3), 323–330. https://doi.org/10.1111/2041-210X.12472

Lakens, D., van Assen, M., Anvari, F., Corker, K., Grange, J., Gerger, H., Hasselman, F., Koyama, J., Locher, C., Miller, I., Page-Gould, E., Schönbrodt, F., Sharples, A., Spellman, B., & Zhou, S. (2017). Examining the reproducibility of meta-analysis in psychology: A preliminary report. https://doi.org/10.31222/osf.io/xfbjf

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*(1), 1–10. https://doi.org/10.1186/s40359-016-0126-3

Lau, J., Schmid, C. H., & Chalmers, T. C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology*, *48*(1), 45–57. https://doi.org/10.1016/0895-4356(94)00106-Z

Lazarsfeld, Paul F. (1940). „Panel“ Studies. *The Public Opinion Quarterly, 4*, 122-128.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpeamel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389–402. https://doi.org/10.1177/251524591878748

Leimu, R., & Koricheva, J. (2004). Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1551), 1961–1966. https://doi.org/10.1098/rspb.2004.2828

Lensvelt-Mulders, G., Hox, J., Van Der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized response research. Thirty-five years of validation. *Sociological Methods and Research, 33*(3), 319–348. https://doi.org/10.1177/0049124104268664

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling Errors in Surveys.* New York: John Wiley & Sons.

Lewis, M. G., & Nair, S. (2015). Review of applications of Bayesian meta-analysis in systematic reviews. *Global Journal of Medicine and Public Health, 4*(1), 1–9. https://www.researchgate.net/publication/274393040

Li, J., & Van den Noortgate, W. (2019). A meta-analysis of the relative effectiveness of the item count technique compared to direct questioning. *Sociological Methods & Research.* https://doi.org/10.1177/0049124119882468

Light, R.J., Pillemer, D.B. (1984). *Summing up. The science of reviewing research*. Harvard University Press, Cambridge, MA.

Linde, K., Berner, M., Egger, M., & Mulrow, C. (2005). St John's wort for depression: Metaanalysis of randomised controlled trials. *British Journal of Psychiatry, 186*, 99–107. https://doi.org/10.1192/bjp.186.2.99

Lensvelt-Mulders, G. J., Hox, J. J., Van Der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research, 33*(3), 319–348. https://doi.org/10.1177/0049124104268664

London, J. E. (2016). *The effect of time period, field, and coding context on rigor, interrater agreement, and interrater reliability in meta-analysis*. Dissertation, North Carolina State University

Lozar Manfreda, K., Bošnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*, 79–104. https://doi.org/10.1177/147078530805000107

Massing, N., Wasmer, M., Wolf, C., & Zuell, C. (2019). How standardized is occupational coding? A comparison of results from different coding agencies in Germany. *Journal of Official Statistics, 35*(1), 167–187. https://doi.org/10.2478/jos-2019-0008

Mavletova, A., & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In Toninelli, D., Pinter, R., & de Pedraza, P. (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies,* (pp. 81–98). London, UK: Ubiquity Press.

McCarley, J. S., & Benjamin, A. S. (2013). Bayesian and signal detection models. In Lee, J., & Kirlik, A. (Eds.), *The Oxford Handbook of Cognitive Engineering*, (pp. 465–475). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0032

McCarthy, R.J., Chartier, C.R. (2017): Collections^2: Using "Crowdsourcing" within Psychological Research. *Collabra: Psychology*, 3 (1): 26. https://doi.org/10.1525/collabra.107

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with Signal Detection Theory. *Annual Review of Psychology, 50*(1), 215–241. https://doi.org/10.1146/annurev.psych.50.1.215

Medway, R. L., & Fulton, J. (2012). When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly, 76*(4), 733–746. https://doi.org/10.1093/poq/nfs047

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly, 79*(1), 105–129. https://doi.org/10.1093/poq/nfu059

Millard, T., Synnot, A., Elliott, J., & Turner, T. (2018). *Results from the evaluation of the pilot living systematic reviews*. Retrieved from: https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201905 LSR_pilot_evaluation_report.pdf

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ (Clinical research ed.)*, 339, b2535. https://doi.org/10.1136/bmj.b2535

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Internet coverage and coverage bias in Europe: Developments across countries and over time. *Journal of Official Statistics, 29*(4), 609–622. https://doi.org/10.2478/jos-2013-0042

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, *27*(11), 1450–1462. https://doi.org/10.1177/01461672012711006

* Murray, M., Swan, A. V., Kiryluk, S. & Clarke, G. C. (1988). The Hawthorne effect in the measurement of adolescent smoking. *Journal of epidemiology and community health, 42*, 304–306. https://doi.org/10.1136/jech.42.3.304

Nakagawa, S., Lagisz, M., O'Dea, R.E., Rutkowska, J., Yang, Y., Noble, D., & Senior, A. (2021). The orchard plot: Cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods, 12*: 4– 12. https://doi.org/10.1002/jrsm.1424

Nancarrow, C., & Cartwright, T. (2007). Online access panels and tracking research: the conditioning issue. *International Journal of Market Research, 49*(5), 573–594. https://doi.org/10.1177/147078530704900505

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137-141. http://dx.doi.org/10.1027/1864-9335/a000192

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH - a graphical display of study heterogeneity. *Research Synthesis Methods, 3*(3), 214–223. https://doi.org/10.1002/jrsm.1053

* Paiva, C. E., Martins, B. P., & Paiva, B. S. R. (2018). Doctor, are you healthy? A cross-sectional investigation of oncologistburnout, depression, and anxiety and an investigation of their associated factors. *BMC Cancer, 18*, 1044. https://doi.org/10.1186/s12885-018-4964-7

Peeters, Y., & Stiggelbout, A. M. (2010). Health state valuations of patients and the general public analytically compared: A meta-analytical comparison of patient and population health state utilities. *Value in Health, 13*(2), 306–309. https://doi.org/10.1111/j.1524-4733.2009.00610.x

Percy, A., McAlister, S., Higgins, K., McCrystal, P. & Thornton, M. (2005), Response consistency in young adolescents' drug use self-reports: a recanting rate analysis. *Addiction, 100*: 189-196. https://doi.org/10.1111/j.1360-0443.2004.00943.x

* Peth, J., Jelinek, L., Nestoriuc, Y., & Moritz, S. (2018). Adverse effects of psychotherapy in depressed patients – First application of the Positive and Negative Effects of Psychotherapy Scale (PANEPS). *Psychotherapie Psychosomatik Medizinische Psychologie, 68*, 391–398. https://doi.org/10.1055/s-0044-101952

Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods, 11*(4), 361–368. https://doi.org/10.18148/srm/2017.v11i4.7145

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, A., Hajek, K., Helmschrott, S., Kleinert, C., Koch, A., Krieger, U., Kroh, M., Martin, S., Saßenroth, D., Schmiedeberg, C., Trüdinger, E., & Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly, 79*, 740–768. https://doi.org/10.1093/poq/nfv014

Phillips, D.L., Clancy, K.J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology 77*:5, 921-940. https://doi.org/10.1086/225231

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence, 38*(3), 314–323. https://doi.org/10.1016/j.intell.2010.03.001

Pigott, T. (2012). *Advances in Meta-Analysis*. New York: Springer.

Pinto-Meza, A., Caseray, X., Soler, J., Puigdemont, D., Perez, V., & Torrubia, R. (2006). Behavioral inhibition and behavioural activation systems in current and recovered major depression participants. *Personality and Individual Differences, 40*, 215–226. https://doi.org/10.1016/j.paid.2005.06.021

Ponto, J. (2015). Understanding and evaluating survey research. *Journal of the Advanced Practitioner in Oncology, 6*(2), 168–171.

* Prinz, B., Dvorak, J., & Junge, A. (2016). Symptoms and risk factors of depression during and after the football career of elite female players. *BMJ Open Sport & Exercise Medicine, 2*, e000124. https://doi.org/10.1136/bmjsem-2016-000124

Pupovac, V., & Fanelli, D. (2015). Scientists admitting to plagiarism: A meta-analysis of surveys. *Science and Engineering Ethics, 21*(5), 1331–1352. https://doi.org/10.1007/s11948-014-9600-6

* Quick, A., Böhnke, J.R., Wright, J., & Pickett, K.E. (2017). Does involvement in a cohort study improve health and affect health inequalities? A natural experiment. *BMC Health Service Research 17*, 79. https://doi.org/10.1186/s12913-017-2016-7

Raudenbush, S.W. (2009). Analyzing effect sizes: Random-effects models. In Cooper, H., Hedges, L., & Valentine, J. (Eds.), *The Handbook of Research Synthesis and Meta-Analysis*, chapter 16, (pp. 295-316). New York: Russell Sage Foundation.

Regan, T. L., & Oaxaca, R. L. (2009). Work experience as a source of specification error in earnings models: Implications for gender wage decompositions. *Journal of Population Economics, 22*(2), 463–499. https://doi.org/10.1007/s00148-007-0180-5

Richman, W. L., Weisband, S., Kiesler, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754–775. https://doi.org/10.1037/0021-9010.84.5.754

Roberts, C., Vandenplas, C., & Stähli, M. E. (2014). Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs. *Survey Research Methods, 8*(2), 67–80. https://doi.org/10.18148/srm/2014.v8i2.5459

Roberts, C., & Vandenplas, C. (2017). Estimating components of mean squared error to evaluate the benefits of mixing data collection modes. *Journal of Official Statistics, 33*(2), 303–334. https://doi.org/10.1515/jos-2017-0016

Robin, X., Turck, N., Hainard, A. et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics 12*, 77. https://doi.org/10.1186/1471-2105-12-77

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*, 1101–1108. https://doi.org/10.1016/j.jval.2011.06.00

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (2005). Publication bias in meta-analyses. In Rothstein, H.R., Sutton, A.J., & Borenstein, M. (eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments,* (p. 1–7). West Sussex, UK: Wiley.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*(3), 638−641. https://psycnet.apa.org/doi/10.1037/0033-2909.86.3.638

* Salles, A., Wright, R. C., Milam, L., Panni, R. Z., Liebert, C. A., Lau, J. N., Lin, D. T., & Mueller, C. M. (2019). Social belonging as a predictor of surgical resident well-being and attrition. *Journal of Surgical Education, 76*, 370–377. https://doi.org/10.1016/j.jsurg.2018.08.022

Sanchez, M. E. (1992). Effects of Questionnaire Design on the Quality of Survey Data. *Public Opinion Quarterly, 56*(2), 206–217. https://www.jstor.org/stable/2749170

Sansone, S.-A. & Rocca-Serra, P. (2016). Interoperability Standards - Digital Objects in Their Own Right. *Wellcome Trust.* https://dx.doi.org/10.6084/m9.figshare.4055496

Santos, H. C., Varnum, M. E. W., & Grossmann, I. (2017). Global increases in individualism. *Psychological Science, 28*, 1228– 1239. https://doi.org/10.1177/0956797617700622

Saran, A., & White, H. (2018). Evidence and gap maps: a comparison of different approaches. *Campbell Systematic Reviews, 14*(1), 1–38. https://doi.org/10.4073/cmdp.2018.2

Saran, A, White, H, Albright, K, Adona, J. Mega map of systematic reviews and evidence and gap maps on the interventions to improve child well-being in low- and middle-income countries. *Campbell Systematic Reviews. 2020*; 16:e1116. https://doi.org/10.1002/cl2.1116

Sayles, H., Belli, R., & Serrano, E. (2010). Interviewer variance between event history calendar and conventional questionnaire interviews. *Public Opinion Quarterly, 74*, 140–153.

Saywitz, K. J., Wells, C. R., Larson, R. P., & Hobbs, S. D. (2019). Effects of interviewer support on children's memory and suggestibility: Systematic review and meta-analyses of experimental research. *Trauma, Violence, & Abuse, 20*(1), 22–39. https://doi.org/10.1177/1524838016683457

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly, 62*, 378–397. https://doi.org/10.1086/297851

Schanze, J.-L., & Zins, S. (2019). Undercoverage of the elderly institutionalized population: The risk of biased estimates and the potentials of weighting. *Survey Methods: Insights from the Field.* https://doi.org/10.13094/SMIF-2019-00017

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 181*(2), 379–407. https://doi.org/10.1111/rssa.12297

Schnell, R., Trappmann, M., & Gramlich, T. (2014). A study of assimilation bias in name-based sampling of migrants. *Journal of Official Statistics, 30*(2), 231–249. https://doi.org/10.2478/jos-2014-0015

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. https://doi.org/10.2139/ssrn.2722435.

Schonlau, M. & Toepoel, V. (2015). Straightlining in Web survey panels over time. *Survey Research Methods, 9* (2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

* Schuring, N., Kerkhoffs, G., Gray, J., & Gouttebarge, V. (2017). The mental wellbeing of current and retired professional cricketers: an observational prospective cohort study. *Physician & Sports Medicine, 45*, 463–469. https://doi.org/10.1080/00913847.2017.1386069

Schwarz, N. (2003). Culture-sensitive context effects: A challenge for cross-cultural surveys. In Harkness, J., Van de Vijver, F.J., & Mohler, P. (Eds.), *Cross-cultural survey methods,* (pp. 93–100). New Jersey: Wiley.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review*, *17*(8), 881–901. https://doi.org/10.1016/S0272-7358(97)00056-1

* Shigemura, J., Sato, Y., Yoshino, A., & Nomura, S. (2008). Patient satisfaction with antidepressants: An Internet-based study. *Journal of Affective Disorders, 107*, 155–160. https://doi.org/10.1016/j.jad.2007.08.019

Shih, T.-H., & Fan, X. (2007). Response rates and mode preferences in web-mail mixed-mode surveys: a meta-analysis. *International Journal of Internet Science, 2*(1), 59–82. Retrieved from https://www.ijis.net/ijis2_1/ijis2_1_shih.pdf

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods, 20*(3), 249–271. https://doi.org/http://dx.doi.org/10.1177/1525822X08317085

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine, 147*(4), 224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179

Silber, H., Schröder, J., Struminskaya, B., Stocké, V., & Bošnjak, M. (2019). Does panel conditioning affect data quality in ego-centered social network questions? *Social Networks*, *56*, 45–54. https://doi.org/10.1016/j.socnet.2018.08.003

Simmonds, M., Salanti, G., McKenzie, J., Elliott, J., On Behalf of the Living Systematic Review Network (2017). Living systematic reviews: 3. Statistical methods for updating meta-analyses. *Journal of Clinical Epidemiology*, *91*, 38–46. https://doi.org/10.1016/j.jclinepi.2017.08.008

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. http://dx.doi.org/10.2139/ssrn.2694998

Slud, E. V., & Bailey, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *Journal of Official Statistics, 26*(1), 127–143. Retrieved from https://www.math.umd.edu/~slud/myr.html/Census/JOS-Slud-Bailey.pdf

Smith, Tom W. (2011). Refining the Total Survey Error Perspective, *International Journal of Public Opinion Research*, 23(4), 464–484. https://doi.org/10.1093/ijpor/edq052

Smyth, J. D., & Olson, K. (2020). How well do interviewers record responses to numeric, interviewer field-code, and open-ended narrative questions in telephone surveys? *Field Methods, 32*(1), 89–104. https://doi.org/10.1177/1525822X19888707

Snilstveit, B, Bhatia, R, Rankin, K and Leach, B. (2017). 3ie evidence gap maps: a starting point for strategic evidence production and use, *3ie Working Paper 28*. New Delhi: International Initiative for Impact Evaluation (3ie)

* Song, Y. (2017). Rotation group bias in current smoking prevalence estimates using TUS-CPS. *Survey Research Methods, 11*, 383–404. https://doi.org/10.18148/srm/2017.v11i4.6262

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

* Strohmeier, H., Scholte, W. F., & Ager, A. (2018). Factors associated with common mental health problems of humanitarian workers in South Sudan. *PLoS One, 13*, e0205333. https://doi.org/10.1371/journal.pone.0205333

* Struminskaya, B. (2016). Respondent conditioning in online panel surveys: Results of two field experiments. *Social Science Computer Review, 34*(1), 95–115. https://doi.org/10.1177/0894439315574022

Sturgis, P. (2004). The effect of coding error on time use surveys estimates. *Journal of Official Statistics, 20*(3), 467-480. Retrieved from http://eprints.lse.ac.uk/id/eprint/102000

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: the psychology of panel conditioning. In Lynn, P. (ed.), *Methodology of Longitudinal Surveys, (pp. 113-126).* Chichester, GB: Wiley Series in Survey Methodology.

Swets, J.A. (1996) *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers.* Mahwah: Lawrence Erlbaum Associates.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26. https://doi.org/10.1111/1529-1006.001

Swets, J.A. & Pickett, R.M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Synnot, A., Turner, T., & Elliott, J. (2017). *Cochrane Living Systematic Reviews*. Retrieved from https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/LSR Interim guidance_v0.3_20170703.pdf

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., & Elliott, J. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology, 91*, 31–37. https://doi.org/10.1016/j.jclinepi.2017.08.011

* Toepoel, V., Das, M., & van Soest, A. (2009). Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents. *Survey Research Methods, 3*(2), 73-80. https://doi.org/10.18148/srm/2009.v3i2.874

* Torche, F., Warren, J. R., Halpern-Manners, A., & Valenzuela, E. (2012). Panel conditioning in a longitudinal study of adolescents’ substance use: Evidence from an experiment. *Social Forces, 90*(3), 891–918. https://doi.org/10.1093/sf/sor006

Tourangeau, R., & Hanover, L. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education, 26*(2), 169–181.

Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response.* Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Tran, U. S., Hofer, A. A., & Voracek, M. (2014). Sex differences in general knowledge: Meta-analysis and new data on the contribution of school-related moderators among high-school students. *PLoS One, 9*(10), Article e110391. https://doi.org/10.1371/journal.pone.0110391

Trappmann, M., Gundert, S., Wenzig, C., & Gebhardt, D. (2010). PASS: A Household Panel Survey for Research on Unemployment and Poverty. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 130*, 609–622. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-429755

Trudel-Fitzgerald, C., Millstein, R.A., von Hippel, C., Howe, C.J., Tomasso, L., Wagner, G.R., & VanderWeele, T.J. (2019). Psychological well-being as part of the public health debate? Insight into dimensions, interventions, and policy. *BMC Public Health 19*, 1712. https://doi.org/10.1186/s12889-019-8029-x

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Tsuji, S., Bergmann, C., Lewis, M., Braginsky, M., Piccinini, P., Frank, M. C., & Cristia, A. (2018). MetaLab: A repository for meta-analyses on language development, and more. In International Speech Communication Association (Ed.), *Proceedings of the Annual*

Conference of the International Speech Communication Association (INTERSPEECH, 2017). Retrieved from https://www.isca-speech.org/archive/Interspeech_2017/pdfs/2053.PDF.

Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Bassem H, Mueller M, Lisacek F, et al. (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine, 36*. 107–115. 10.1007/s00134-009-1641-y

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PLoS ONE*, *8*(3), 1–8. https://doi.org/10.1371/journal.pone.0059202

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105–110. https://doi.org/10.1037/h0031322

van Aert, R., Wicherts, J., & van Assen, M. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*(5), 713–729. https://doi.org/10.1177/1745691616650874

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6.

Van Horn, P. S., Green, K. E., & Martinussen, M. (2009). Survey response rates and survey administration in counseling and clinical psychology. *Educational and Psychological Measurement, 69*, 389–403. https://doi.org/10.1177/0013164408324462

* Van Overveld, M., De Jong, P. J., Peters, M. L., Van Hout, W. J. P. J., & Bouman, T. (2008). An internet-based study on the relation between disgust sensitivity and emetophobia. *Journal of Anxiety Disorders, 22*, 524–531. https://doi.org/10.1016/j.janxdis.2007.04.001

Vehovar, V., Lozar Manfreda, K., & Batagelj, Z. (1999): Web Surveys: Can the Weighting Solve the Problem? *Proceedings of the American Statistical Association, 1999,* 962-967. Retrieved from http://www.asasrms.org/Proceedings/papers/1999_168.pdf

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Journal of Psychology, 215*(2), 104–121. https://doi.org/10.1027/0044-3409.215.2.104

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package, *Journal of Statistical Software, 36* (3), 1-48. http://dx.doi.org/10.18637/jss.v036.i03

Villar, A., Callegaro, M., & Yang, Y. (2013). Where am I? A meta-analysis of experiments on the effects of progress indicators for web surveys. *Social Science Computer Review, 31*(6), 744–762. https://doi.org/10.1177/0894439313497468

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Journal of Psychology, 227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Voutilainen, A., Pitkäaho, T., Vehviläinen-Julkunen, K., & Sherwood, P. R. (2015). Meta-analysis: Methodological confounders in measuring patient satisfaction. *Journal of Research in Nursing, 20*(8), 698–714. https://doi.org/10.1177/1744987115619209

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics, 20*(2), 165–205. https://doi.org/10.1214/aoms/1177730030

* Warren, J. R., & Halpern-Manners, A. (2012). Panel conditioning in longitudinal social science surveys. *Sociological Methods and Research, 41*(4), 491–534. https://doi.org/10.1177/0049124112460374

* Waterton, J. & Lievesley, D. (1989). Evidence of conditioning effects in the British social attitudes panel survey. In: D. Kasprzyk (Ed.). *Panel surveys* (pp. 319–339). New York, NY: Wiley.

Weiland, P., Baier, C., & Ramthun, R. (2019). PsychArchives – Unterstützung des Forschungszyklus in der Psychologie mit DSpace. (Vortrag, 11.04.2019). *Bamberg: DSpace Anwendertreffen*. http://dx.doi.org/10.23668/psycharchives.2416

Weiner, S. P., & Dalessio, A. T. (2006). Oversurveying: Causes, consequences, and cures. In Kraut, A.I. (Ed.), *Getting action from organizational surveys: New concepts, technologies, and applications*, (pp. 294–311). Chapter 12. San Francisco, CA: Jossey-Bass.

Weisberg, H. (2005). *The Total Survey Error Approach. A Guide to the New Science of Survey Research*. University of Chicago Press.

Weiß, B., Das, M., Kapteyn, A., Bošnjak, M. & Schaurer, I. (2020). *Open Probability-based Panels.* New York: Wiley. http://dx.doi.org/10.1002/9781118445112.stat07988.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

World Bank, World Development Indicators. (2018). *Individuals using the Internet (% of population)* [Data file]. Washington, DC: The World Bank. Retrieved from https://data.worldbank.org/indicator/IT.NET.USER.ZS

WHO. (2017). *Depression and other common mental disorders: Global health estimates.* Geneva, Switzerland: World Health Organization. Retrieved from https://apps.who.int/iris/handle/10665/254610

* Williams, P., Block, L. G., & Fitzsimons, G. J. (2006). Simply asking questions about health behaviors increases both healthy and unhealthy behaviours: Erratum. *Social Influence, 1*(3), 247. https://doi.org/10.1080/15534510600958251

* Williams, E., Martin, S. L., Fabrikan, A., Wang, A., & Pojasek, M. (2018). Rates of depressive symptoms among pharmacy residents. *American Journal of Health-System Pharmacy, 75*, 292–297. https://doi.org/10.2146/ajhp161008

Wood, W., & Eagly, A.H. (2009). Advantages of certainty and uncertainty. In Cooper, H., Hedges, L., & Valentine, J. (Eds.). *The Handbook of Research Synthesis and Meta-Analysis,* chapter 24, (pp. 455-472). New York: Russell Sage Foundation.

Yarger, J., James, T., Ashikaga, T., Hayanga, A., Takyi, V., Lum, Y., Kaiser, H., & Mammen, J. (2013). Characteristics in response rates for surveys administered to surgery residents. *Surgery 154*:38–45. https://doi.org/10.1016/j.surg.2013.04.060

Zhang, C., Antoun, C., Yan, H. Y., & Conrad, F. G. (2020). Professional Respondents in Opt-in Online Panels: What Do We Really Know? *Social Science Computer Review*, *38*(6), 703–719. https://doi.org/10.1177/0894439319845102

Zhang, L. C., Thomsen, I. B., & Kleven, O. (2013). On the use of auxiliary and paradata for dealing with non-sampling errors in household surveys. *International Statistical Review, 81*(2), 270–288. https://doi.org/10.1111/insr.12009

Zhu, Z. N., Jiang, Y. F., & Ding, T. (2014). Risk of fracture with thiazolidinediones: An updated meta-analysis of randomized clinical trials. *Bone, 68*, 115–123. https://doi.org/10.1016/j.bone.2014.08.010

* Zimmerman, M., Chelminski, I., Young, D., & Dalrymple, K. (2011). Using outcome measures to promote better outcomes. *Clinical Neuropsychiatry, 8*, 28–36. https://psycnet.apa.org/record/2011-16491-003

Appendix

Appendix A: Original Publication of Study 1

Response Rates in Online Surveys With Affective Disorder Participants

A Meta-Analysis of Study Design and Time Effects Between 2008 and 2019

Tanja Burgard[1], Michael Bošnjak[1,2], and Nadine Wedderhoff[2]

[1] Research Synthesis Unit, Leibniz Institute for Psychology Information (ZPID), Trier, Germany
[2] Department of Psychology, University of Trier, Germany

Abstract: A meta-analysis was performed to determine whether response rates to online psychology surveys have decreased over time and the effect of specific design characteristics (contact mode, burden of participation, and incentives) on response rates. The meta-analysis is restricted to samples of adults with depression or general anxiety disorder. Time and study design effects are tested using mixed-effects meta-regressions as implemented in the metafor package in R. The mean response rate of the 20 studies fulfilling our meta-analytic inclusion criteria is approximately 43%. Response rates are lower in more recently conducted surveys and in surveys employing longer questionnaires. Furthermore, we found that personal invitations, for example, via telephone or face-to-face contacts, yielded higher response rates compared to e-mail invitations. As predicted by sensitivity reinforcement theory, no effect of incentives on survey participation in this specific group (scoring high on neuroticism) could be observed.

Keywords: response rates, online survey, meta-analysis, affective disorders

Declining Survey Response Rates and Oversurveying

Nonresponse is one of the most severe problems in social and behavioral research challenging both the internal and the external validity of surveys (Hox & de Leeuw, 1994). There are different forms of nonresponse. The dependent outcome of this meta-analysis is the response rate, which is defined here as the number of complete interviews divided by the number of interview attempts (interviews plus the number of refusals and breakoffs plus all cases of unknown eligibility; American Association for Public Opinion Research, 2016).

If the causes for missingness are independent to any other (observed or unobserved) parameter (i.e., data are missing completely at random; Little & Rubin, 2019), nonresponse reduces the amount of data collected. A smaller sample size leads to a larger sampling variance, resulting in less precise estimates and lower statistical power. However, if the reason for nonresponse is nonrandom, missing data can cause biased results and invalid conclusions, as the final respondents are no longer representative for the population of interest (Groves & Peytcheva, 2008).

There is ample evidence on declining response rates to household surveys in the social and political sciences (Brick & Williams, 2013; Krosnick, 1999), and to surveys in counseling and clinical psychology (Van Horn, Green, & Martinussen, 2009). This trend can aggravate the possible bias due to nonresponse.

To explain this decline, participation in a scientific study can be regarded as a culturally shaped decision problem (Haunberger, 2011a). Evidence indicating cultural differences in response patterns, including the extent of nonresponse, is found in cross-cultural survey methodology (Baur, 2014). In cultures emphasizing individualism, individuals are mainly responsible for themselves and decisions tend to be based on an individual cost-benefit analysis. In this context, value-expectancy theories (Esser, 2001), such as the theory of planned behavior (Ajzen, 1991), are especially suitable to explain participation in surveys (Bošnjak, Tuten, & Wittmann, 2005; Haunberger, 2011b).

https://doi.org/10.1027/2151-2604/a000394

In Western societies, a shift from collectivist values toward individualism has been observed (Greenfield, 2013; Hofstede, 2001). For instance, substantial increases in individualistic tendencies in word use and naming of children have been detected (for China between 1975 and 2015: Zeng & Greenfield, 2015; for Japan: Ogihara, 2017; for the United States between 2004 and 2015: Twenge, Dawson, & Campbell, 2016). There is also evidence of changes in relational and cultural practices in both the United States and Japan across several decades until 2015 (Grossmann & Varnum, 2015; Hamamura, 2011; Ogihara, 2018). Furthermore, increases in individualistic behavioral choices, practices, and values were observed by Santos, Varnum, and Grossmann (2017) for 37 of the 51 countries they examined. Taken together, these findings substantiate a global shift toward individualistic values and behaviors.

This shift serves as a rationale for the overall decline in participation rates, as participants feel less socially obliged to help the interviewer, for instance, if the survey does not provide a benefit for themselves. In contrast, it also nurtures the assumption that characteristics of the study design related to the individual costs and benefits of participants, such as incentives and interest in the topic, might have gained in importance for motivating people to participate in a survey (Esser, 1986).

As survey participation is interrelated with culture and communication (Schwarz, 2003), another factor that might have caused changes in response rates is the increase of Internet usage in recent years. In the European Union, the share of individuals using the Internet has increased from 60% in 2007 to 84% in 2018 (World Bank, 2018).

The Internet has also become a more popular platform for conducting surveys in recent years, due to the fast and easy implementation and low costs of online surveys. Yet they are thought to suffer even more from issues of nonresponse and a lack of representativeness (Cook, Heath, & Thompson, 2000). In their meta-analysis, Lozar Manfreda, Bošnjak, Berzelak, Haas, and Vehovar (2008) concluded that web surveys yield lower response rates than other survey modes. Interestingly, the ever-increasing growth of the Internet and the increase in web surveys in general have not changed the willingness to participate in these types of surveys relative to other survey modes (Daikeler, Bošnjak, & Lozar Manfreda, 2019). More than a decade ago, Shi and Fan's (2008) meta-analytic comparison revealed that web surveys yielded an average response rate of 34% in contrast to 45% for paper surveys. Following the general trend of increased nonresponse rates, the absolute level of these response rates might have decreased in the 11 years since this meta-analysis.

A decrease in the participation in online psychology surveys may be a result of oversurveying (Groves et al., 2004; Weiner & Dalessio, 2006), reflecting the research trends in other scientific branches and for other modes of data collection. This is due to less attention to single communication requests, because of the amount of information to be processed. As a consequence, potential survey participants may not be interested in taking part in single studies (Groves, Cialdini, & Couper, 1992). Oversurveying may also influence the perception of social exchange, in the sense of giving participants the feeling to have done their part after having participated in a few studies, reducing the willingness to participate in the following (Groves & Magilavy, 1981).

Given the severe consequences of nonresponse on external validity, it should be the ambition of every scientist to keep survey nonresponse to a minimum. Therefore, it is essential to know the possible effects of a study's design on people's willingness to participate in the study. This knowledge may serve as a guide when determining, for example, the use of incentives or the contact mode of the invitation.

In this meta-analysis, we will examine if the trend of declining response rates holds for online surveys in psychology, specifically focusing on participants with depression or anxiety disorders. From an epidemiological perspective, this is an important population that may be hard to reach and difficult to motivate to participate in studies. The moderating effects of time and survey design will be tested using study characteristics (contact mode, number of items, and use of incentives). The results of the meta-analysis should guide researchers in how to optimally implement online psychology surveys that yield high response rates. Thus, our first hypothesis focuses on the time effect:

> *Hypothesis 1 (H1)*: The response rates in online psychology surveys have decreased over time.

Effects of Study Design Characteristics on Response Rates

In times of oversurveying, one method to draw attention to studies in order to achieve higher response rates is contacting the potential participants personally. Participants can be invited to access online surveys in various ways that differ in the extent of personal contact. For example, contacting potential respondents by phone is a more personal invitation than sending an e-mail invitation to participate via a mailing list. Schaefer and Dillman (1998) stress the importance of a personal contact to potential respondents, an act which conveys their importance for the survey institution. In a study of student engagement in a university survey (Nair, Adams, & Mertova, 2008), about half of the nonrespondents recontacted by telephone were convinced by the personal contact to complete the online survey. The meta-analysis of Cook et al. (2000) also shows that more personalized contacts yield higher response rates in online surveys.

Examining the type of contact to deliver the invitation to participate in a survey, we assume:

> *Hypothesis 2 (H2)*: Personal or phone contact as an invitation mode yields higher response rates in online psychology surveys than e-mail invitations.

The influence of survey length on response rates was examined meta-analytically by Rolstad, Adler, and Rydén (2011): They found a clear association between questionnaire length and response rates. Yet it is not clear whether the difference in response rates is directly attributable to the length of the questionnaires. For the association between questionnaire length and experienced response burden, only weak support is found. In Mercer, Caporaso, Cantor, and Townsend's (2015) meta-analysis, multiple criteria were used to classify surveys as burdensome, and findings indicated that a survey classified as burdensome led to response rates more than 20% lower than for low-burden surveys. Galesic and Bošnjak (2009) conducted an experiment in which the announced length of the survey, incentives, and the order of thematic blocks were randomly assigned to participants. Findings revealed that the respondents were more likely to start the survey when the stated length was shorter. However, many surveys do not provide information on the length of the survey, with the consequence that a longer survey may lead to higher breakoff rates (e.g., Mavletova & Couper, 2015, in their meta-analysis of mobile web surveys) and thus incomplete datasets. As we are interested in the response rate as the share of completed interviews related to all interview attempts, breakoff during the survey also means lower survey response in this case.

In the context of the higher importance of the cost-benefit analysis due to cultural individualization, over time it can be expected that longer studies suffer more from the decrease in participation than shorter ones. Thus:

> *Hypothesis 3 (H3)*: The higher the number of items in an online survey questionnaire, the lower is the response rate.

An intensively researched topic in the area of survey participation is the effect of incentives. An early meta-analysis showed that prepaid monetary incentives were the most effective, with an average increase in participation of 19.1 percentage points (Church, 1993). The meta-analysis moreover revealed that only initial incentives had an effect on response rates. Incentives contingent on the return of the questionnaire did not provide significant benefits, independent of the type of incentive. In general, cash incentives have a stronger effect on response rates than lottery tickets or other nonmonetary incentives (Pforr et al., 2015). This difference between prepaid and promised incentives was also corroborated by Mercer et al. (2015), but only for telephone and mail surveys. For in-person interviews, the timing of the incentive had no significant impact on the response rates. These findings from cross-sectional research indicate that incentives, under certain conditions, may have an effect on response rates.

However, in the present research we are considering a special population, namely samples with a considerable share of respondents suffering from depressive or anxiety disorders. Following the reinforcement sensitivity theory (Corr, 2002), we can expect that this population, scoring high on neuroticism, will be less sensitive to rewards (Beevers & Meyer, 2002; Bijttebier, Beck, Claes, & Vandereycken, 2009; Pinto-Meza et al., 2006). This would also imply that the effect of incentives for survey participation will be lower than expected for the general population. Thus, we hypothesize:

> *Hypothesis 4 (H4)*: Response rates in online psychology surveys in a group scoring high on neuroticism are not affected by incentives awarded for participation.

Method

Inclusion and Exclusion Criteria

This review has been reported in accordance with the PRISMA statement[1] (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). Of interest are psychological studies reporting response rates from online surveys. Studies reporting on mixed survey types (e.g., online with telephone reminders) that do not report online survey-only rates or studies where the type of survey is not explicitly reported were excluded. Moreover, to be useful for hypothesis testing, at least one of the study design characteristics of interest has to be reported: number of items in the questionnaire, use of incentives, or contact mode of the invitation.

Student samples were excluded due to differing motivation structure and incentives. Especially psychology students are often obliged to take part in psychology surveys as part of their studies. Their motivation therefore

[1] The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement consists of a checklist with 27 items that have to be reported in a research synthesis and a diagram for reporting the flow of information through the four phases of the literature selection. It helps authors to improve their reporting.

is not comparable to other populations participating voluntarily. Moreover, the meta-analysis is restricted to samples of adults with general anxiety disorder or depression, as this population is of growing epidemiological importance (WHO, 2017) and, therefore, of special interest in the domain of psychology. Beyond this, the restriction on a focus population kept the number of primary studies manageable. In the case of panel studies, only the first wave is of interest due to panel mortality in later waves. In longitudinal studies with multiple cross-sectional samples, there is a new sample for each wave and, thus, all samples are coded separately. There is no restriction concerning the year or language of publication. Relevant systematic reviews identified during screening were tagged for reference checking. An overview of the inclusion and exclusion criteria for study selection is presented in Table S1 (available at: https://doi.org/10.23668/psycharchives.2626).

Moderator Analyses

In the meta-analysis, we test moderating effects on survey response rates. As potential moderators, basic information from the report, such as name of the first author, publication year, type of report, and funding, are coded, as well as information on the sample and potentially relevant study design characteristics. In the present meta-analysis, we focus on a specific population, that is, adults with depression or anxiety disorders; thus, relevant descriptive information includes the specific diagnosed disorders in the population and the percentages of participants diagnosed with each disorder. Moreover, the mean age of the population, the percentage of women in the sample, the year, and the country of data collection are of interest.

Characteristics of the study design that could have an effect on the willingness to participate in an online survey are the contact mode of the invitation (mail, e-mail, phone, or personal contact), the burden of the survey (measured with the number of items), the use of incentives, and the topic of the survey.

Finally, a response rate is either given in the report or calculated using the formula defined by the American Association for Public Opinion Research (2016): the number of complete interviews divided by the number of interview attempts (interviews plus the number of refusals and breakoffs plus all cases of unknown eligibility).

Search Strategies

To search for relevant records, 10 databases were used: PsycInfo, Embase, Medline and In-Process, Medline Ahead of Print/Daily Update, the Campbell Library, Science Citation Index, SocINDEX and the Cochrane Central Register of Controlled Trials (CENTRAL), PubPsych containing PSYNDEX, and ReStore. The results of these database searches are shown in Table S2 (available at: https://doi.org/10.23668/psycharchives.2626). In addition, the following conference proceedings were searched manually for potentially relevant records: European Survey Research Association (conference year 2017) and American Association for Public Opinion Research (conference years 2016, 2017, 2018).

The main search terms utilized for the literature search were: (participat* OR respons* OR respond*) AND ("online" OR "internet" OR "web" OR "electronic" OR "world wide web" OR "computer" OR "email") AND ("interview" OR "survey" OR "questionnaire") AND (depress* OR anxi*). The exact search strategies differed slightly between the respective databases and are reported in detail in Supplement S5 (available at: https://doi.org/10.23668/psycharchives.2626).

The retrieved records were screened for eligibility by three independent coders. In the first step, literature that definitely did not meet the inclusion criteria was identified via abstract screening by two screeners. An agreement of 92% was reached for this initial screening. To achieve full agreement, the first two screeners discussed all disagreements and a third screener was consulted concerning the remaining discrepancies. In the end, full agreement was reached. Potentially relevant records were then assessed in a full-text screening to conclusively identify the eligible literature. The full-text screening was conducted by two of the screeners and full consensus was achieved.

Coding Procedures

Half of the included studies were coded by two coders to detect possible discrepancies or sources of misunderstanding in the coding guide. An interrater agreement rate of more than 80% was found between coders for the majority of the coded moderators. All discrepancies could be solved by discussion. Information on the flow of participants, which is crucial for the calculation of the response rate, differed slightly in some cases, such that initial agreement was only 55%. These discrepancies were reevaluated and discussed until consensus was finally achieved.

Statistical Methods

The outcome is the response rate for each treatment. The treatment is an invitation to participate in an online survey. The response rate is a relative measure and thus restricted to values between 0 and 1. It is calculated by dividing the returned, usable questionnaires (equivalent to the sample size of the study) by the number of potential respondents contacted, that would have been eligible or for whom eligibility is unknown.

Our data were collected on several levels (report, study, sample, outcomes), but as we only have one response rate per study and sample, and in each report there is only one usable sample reported, we actually do not have a multilevel data structure. Using the metafor package in R (Viechtbauer, 2010), the time effect is assessed by calculating mixed-effects meta-regressions to test the influence of the year of data collection and the characteristics of the survey design on the response rate.

In the mixed-effects model we assume that the true effect sizes may vary between studies. The observed variance in effect sizes is then comprised of the variance of the true effect sizes (the heterogeneity) and the random error. Since we want to examine what proportion of the observed variance reflects real differences in effect sizes, we will calculate the I^2 statistic. To assess the proportion of true variance in response rates explained by the model used, we will calculate an index analogous to the R^2 index for primary studies. This index, defined as true variance explained (Borenstein, Hedges, Higgins, & Rothstein, 2009), computes the percent reduction in true variance by comparing T^2 of the model including the moderators of interest versus the null model without moderators. The index is restricted to values between 0 and 1. The higher the percent reduction, the more variance in response rates is explained with the corresponding model.

Publication Bias and Selective Reporting

Assessment of publication bias is crucial in this context because studies with a low response rate may be less likely to be published. Therefore, the response rate will be plotted against the standard error, following Sterne et al.'s (2011) recommendations. The symmetry of the resulting funnel plot will provide a qualitative indication of the existence of a publication bias. Moreover, the performance of Egger's test will provide a *p*-value for a formal test of publication bias.

Results

The literature search yielded 2,874 potentially relevant records for screening. Of these, 2,769 records could be excluded due to obviously not meeting inclusion criteria. 105 articles were screened as full text. Of these records, 20 were found to be relevant for coding. The main reasons for exclusion of full text articles was missing information on the flow of participants, thus that no response rate could be computed. Figure S3 (available at https://doi.org/10.23668/psycharchives.2627) shows the selection process of literature in detail. Table S4 (available at https://doi.org/10.23668/psycharchives.2626) gives an overview of the characteristics of the studies finally included. The major drawback resulting from the small sample of included studies is the lack of studies published before 2008. This was not intended, but a result of the restriction on samples with anxiety disorder or depression and the requirement of information for the calculation of the response rate.

Table 1 reports the means and standard deviations of the variables examined in our meta-analysis. For example, the mean publication year is 2016. Eighty-five percent of the samples (n = 17) were invited to participate via e-mail. The mean number of items in the studies was 54. Only four studies reported the use of incentives. Therefore, the timing and kind of incentives, both characteristics potentially highly relevant for the effectiveness of incentives, could not be distinguished in this meta-analysis.

The mean response rate over the 20 studies is 43%. Table 1 also reports the interrelations between the variables. As expected, we found lower response rates for newer studies and for questionnaires containing more items.

Figure 1 displays the cumulative forest plot of the 20 studies included in the meta-analysis. The studies are sorted chronologically, and the evidence is summed up from study to study. At the beginning (studies published in 2008), the confidence intervals are broad and the overall mean response rates are volatile. From about 2018 on, the overall effect remains stable and is hardly affected by new evidence. This suggests that our evidence to estimate the mean response rate for the studies in this meta-analysis is satisfactory at this point. The mixed-effects meta-analysis conducted in R reveals an overall response rate of 42.8% for the 20 studies, with a 95% confidence interval between 31.7% and 53.9%. Almost all of the variance is between the studies (I^2=99.92%).

The funnel plot in Figure 2 depicts a relationship between the response rates and the standard errors. It seems that the response rates in smaller studies are higher than in larger

Table 1. Univariate and bivariate distributions of study characteristics (n = 20 Studies)

Variable	Mean	*SD*	E-mail invitation	Number of items	Incentive	Response rate
Publication year	2016	3.44	−0.313	−0.176	−0.037	−0.444
E-mail invitation	0.85	0.37	–	−0.485	0.210	−0.018
Number of items	54	42.92	–	–	−0.192	−0.191
Incentive	0.20	0.41	–	–	–	−0.059
Response rate	0.43	0.25	–	–	–	–

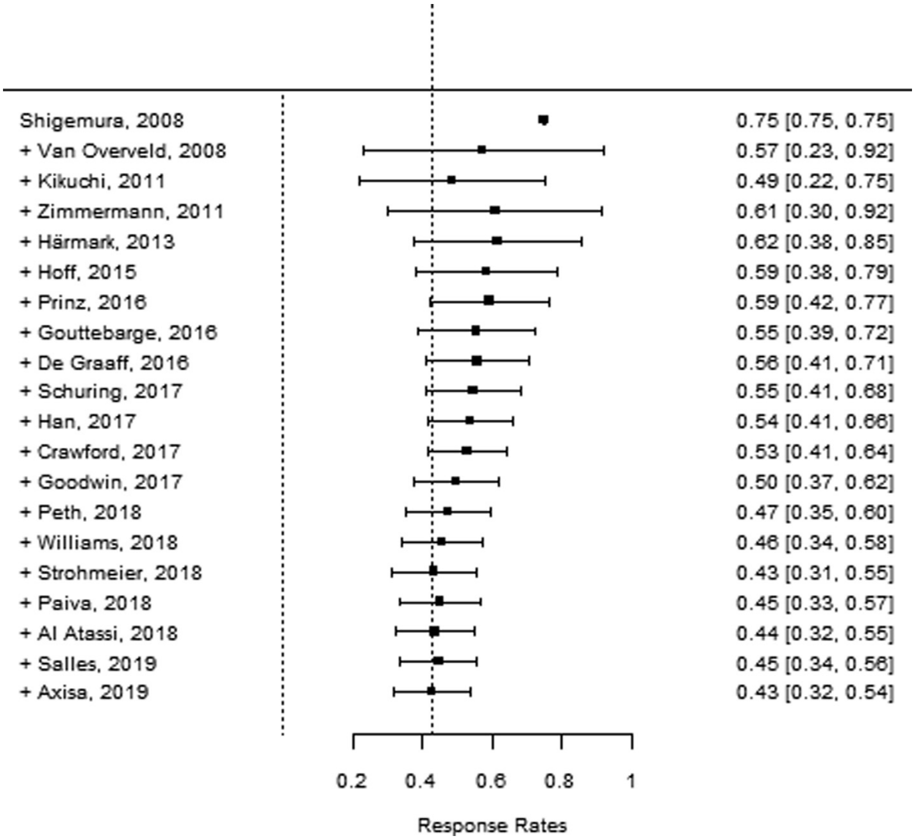


Figure 1. Cumulative forest plot.

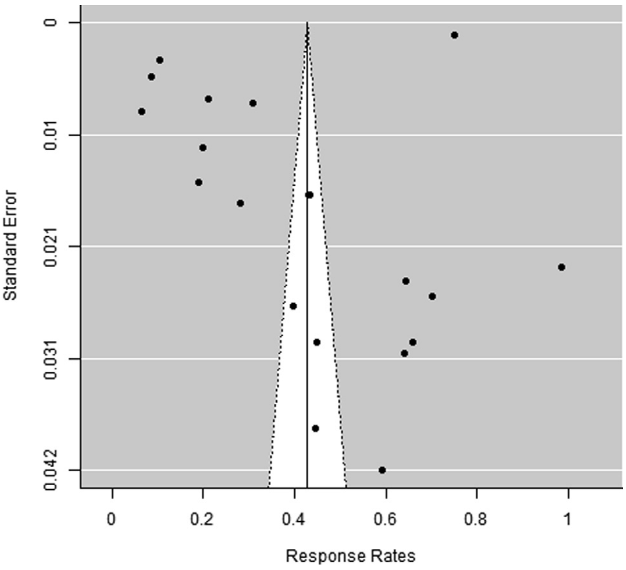


Figure 2. Funnel plot.

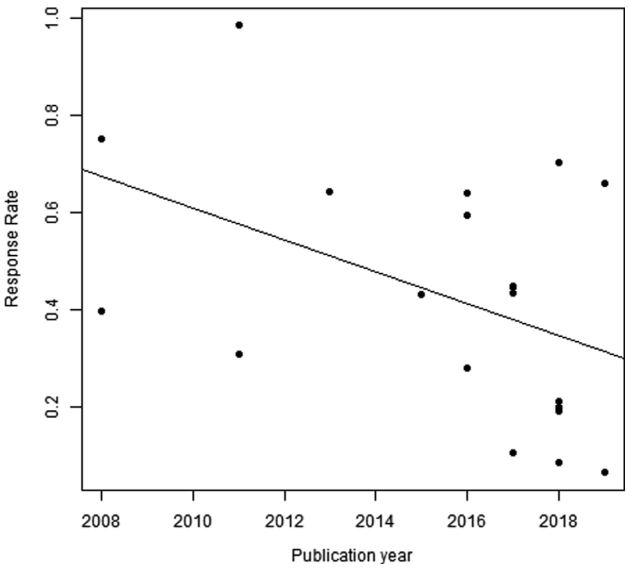


Figure 3. Meta-regression plot of response rates over time.

studies. The result of Egger's test confirms that the relationship is significant ($z = 2.29$, $p = .0219$). As the response rate is not the outcome of interest in the studies, a publication bias is not the most plausible explanation for this finding. Taking into account the assumptions of positive effects of personal contact to potential participants, an alternative rationale for this relationship might be that the participants in smaller studies were more likely to be contacted personally and that this contact might have resulted in higher response rates.

In Figure 3, the bivariate distribution of publication year and response rate is plotted. The linear regression line

Table 2. Results of meta-regressions

Moderator	Full model of study design characteristics	Full model study design + additional controls
Intercept	0.729*** [0.426; 1.033], p < .001	0.698** [0.271; 1.124], p = .001
Publication year (H1)	−0.177**[−0.287; −0.068], p = .002	−0.172** [−0.301; −0.043], p = .009
Contact mode of invitation (H2) (e-mail vs. other)	−0.342* [−0.683; −0.000], p = .050	−0.172** [−0.301; −0.043], p = .009
Number of items (H3)	−0.144*[−0.264; −0.024], p = .018	−0.137. [−0.296; 0.022], p = .092
Incentives (H4)	−0.055 [−0.295; 0.185], p = .654	−0.052 [−0.313; 0.201], p = .695
Funds	–	0.030 [−0.226; 0.286], p = .819
Mean age sample	–	0.001 [−0.120; 0.121], p = .991
I^2	99.72%	99.59%
R^2	28.79%	18.08%

Note. N = 20 studies. Significance levels: ***p < .001, **p < .01, *p < .05, p < .1.

Table 3. Selected predictions for response rates from the full model

Publication year	Number of items	Incentives	Contact mode: e-mail	Actual response rate	Predicted response rate	Difference actual vs. predicted response rates
2008	**187**	0	1	0.3966	**0.3315**	0.07
2011	**18**	0	1	0.9839	**0.7219**	0.26
2018	52	**1**	1	0.1904	**0.2117**	−0.02
2018	54	**0**	1	0.0872	**0.2865**	−0.2
2017	56	0	**1**	0.4457	**0.2993**	0.15
2019	75	0	**0**	0.6602	**0.4862**	0.17
2008	34	1	1	0.7496	**0.7973**	−0.05
2018	38	1	1	0.2119	**0.2544**	−0.04

Note. Bold values highlight the relevant characteristic in the respective comparison and the corresponding results.

shows the negative relationship between both variables. This relationship is also significant, and the corresponding R^2 is 15.75%. That means that almost 16% of the variance in the response rates of the studies in the meta-analysis can be explained by the publication year. As can be seen in Figure 3, we obviously have one study without nonresponse. Omitting this study from the analysis, as its effect size deviates significantly from the other studies, does not change the conclusions presented in Table 2. Neither the size nor the significance level of effects is affected. This may be due to the small sample size ($n = 30$) of this outlier study and speaks for the robustness of our results.

In Table 2, the results of the meta-regressions conducted are reported. The inclusion of additional information about the funding and the mean age of the sample does not change the overall conclusions of the hypothesis testing. There is evidence for an overall decrease in response rates over time. The mode of contacting participants is also relevant. The least personal contact mode was via e-mail. Samples contacted this way showed less willingness to participate in an online survey than samples approached personally, by phone, or mail. A higher number of items in the questionnaire is also related to lower response rates. Corroborating our expectations, an effect of incentives is not supported for the population considered in this meta-analysis. With only four studies reporting incentives in our sample, we were unable to distinguish different types of incentives or take into account the timing of incentives, yet this finding does not necessarily mean that incentives have no effect at all. On the contrary, it might also be possible, as previous research indicates, that incentives only have an effect on response rates under certain conditions.

To illustrate the influence of the moderators investigated in this meta-analysis, Table 3 shows predicted response rates depending on the values of the study design characteristics. For each relevant characteristic, two similar studies were matched that only differed substantially in the expression of this characteristic. The first comparison of this kind is at the top of the table: two somewhat dated studies (from the years 2008 and 2011) with samples contacted via e-mail and not given incentives for participation are compared. The burden of participation measured with the number of items is extremely high in the first study and very low in the second study. The difference in predicted response rates is about 40%, the actual response rates differ even more.

The third study and the fourth study (both from 2018) only differ with respect to incentives. This difference hardly influences the prediction of response rates from the model. Moreover, the model does not predict the actual response rate well. This is plausible because, as the meta-regression demonstrated, incentives are not useful for the explanation of the response rates. The third comparison is between two rather similar studies (from the years 2017 and 2019) differing in mode of contact, with one study contacting participants via e-mail and the other study utilizing a different mode of contact. For these two studies, the actual as well as the predicted response rate values were approximately 20% higher when participants were contacted by means of a more personal invitation to participate. Finally, a clear difference in response rates is also found for the two similar studies from 2018 and 2008. The response rate in the more recent study is about 50% lower than that in the older study, and this finding is also predicted by the model.

Discussion

To conclude, the hypothesized influences on response rates were mainly confirmed. The mean response rate of 43% is rather high compared to the mean response rates of 34% and 39.6% for online samples found in the meta-analyses of Shi and Fan (2008) and Cook et al. (2000), respectively. This may, however, be due to our restriction to include only samples of respondents with depression or anxiety disorder. First, because many samples were recruited personally from patient lists in hospitals and second, due to the personal relevance of the topics of the surveys, that all surrounded the affective disorders the participants were suffering from.

These restrictions to our study sample as well as the necessary information requirements to compute the response rates resulted in a small pool of studies available for our meta-analysis. Moreover, the lack of studies published before 2008 was another factor contributing to the low number of studies available for the analyses. Due to these limitations concerning the generalizability of the results, the existing evidence on response rates in online surveys for other populations would be of great importance and should be meta-analyzed in the future.

There are several conclusions that can be drawn from the meta-analysis to guide researchers to optimally implement online psychology surveys and achieve high response rates. First of all, we found clear evidence for the expected decrease in response rates, despite our small sample of studies and the short time interval examined. This result corroborates existing findings of the numerous studies on response rates (Brick & Williams, 2013; Krosnick, 1999; Van Horn et al., 2009).

Second, the increasing number of items in a survey significantly reduces the response rates. Thus, researchers should strive to keep the burden of the survey rather small. This is in line with previous research showing an effect of length of survey on the initial response rate (Galesic & Bošnjak, 2009). Moreover, Mavletova and Couper's (2015) meta-analysis revealed a similar relationship for survey length and breakoff during the survey. Thus, to keep the burden for the respondent low, researchers should aim to design their surveys to be as brief as possible. If the survey can be responded to within a few minutes' time, it might be helpful to mention this in the invitation to the survey.

Third, when sending invitations to participate in online surveys, the meta-regressions clearly indicate that it is more effective to approach potential participants using more personal forms of contact, such as face-to-face or phone contact. Cook et al.'s (2000) earlier meta-analysis provided evidence for the importance of personal forms of contact to achieve higher response rates. The more recent studies investigated in this meta-analysis support Cook et al.'s (2000) finding. Thus, to attain high response rates in surveys conducted online, we recommend contacting and personally inviting respondents to participate in an offline mode before sending them the survey or the link to the survey.

A potentially highly relevant moderator is the use of incentives. In the present study, our search strategy uncovered only a small number (i.e., four) of studies utilizing incentives to include in the meta-analysis, and no effect of incentives was found. Previous research has shown, however, that the effectiveness of incentives for increasing response rates depends on the timing and type of incentive (Church, 1993; Pforr et al., 2015). Here, we were unable to differentiate the type or timing of incentives of the four studies reporting the use of incentives in our sample; consequently, the effectiveness of incentives could not be evaluated in detail. More evidence on the use of incentives in online surveys is needed. A potential strategy could be to increase the population of interest in the meta-analysis to include more, diverse groups, with the consequence of an expanded evidence base allowing for more detailed analyses. Moreover, experimental primary studies examining this effect in detail, for example, by varying timing and type of incentives, would also be of interest.

Further study design factors that could be included in a future meta-analysis are contact protocols, such as the use of prenotifications and reminders (Bošnjak, Neubarth, Couper, Bandilla, & Kaczmire, 2008; Cook et al., 2000), or the use and design of an advance letter or e-mail (for a meta-analysis on advance letters in telephone surveys, see De Leeuw, Callegaro, Hox, Korendijk, & Lensvelt-Mulders, 2007). These study characteristics are reported less frequently than, for example, the use of incentives or the contact mode for invitation. Hence, the small number of studies (i.e., 20) in this meta-analysis did not allow us to examine these characteristics as moderators. Nonetheless they may be highly relevant for achieving high response rates and should be included in studies of online surveys in the future.

The results of our meta-analysis does not allow conclusions to be made concerning nonresponse bias, although we know that this is crucial for drawing conclusions on the generalizability of survey results (Groves & Peytcheva, 2008). Depending on the target outcome of the respective study, we could argue that patients or former patients participating in a survey are also more willing to deal with their psychological problems. This may result in differences between responders and nonresponders, and thus to nonresponse bias, if the treatment of and dealing with a disorder is the topic of the study.

Focusing on online surveys, to examine nonresponse bias and update Groves and Peytcheva's meta-analysis (2008), we need research that empirically examines the characteristics of nonrespondents and their reasons for rejecting the participation (as, e.g., in the study of Sax, Gilmartin, & Bryant, 2003). This can be accomplished, for example, by reviewing administrative records, performing screening interviews before the main interview, or conducting follow-up surveys with nonrespondents. Sample characteristics known to influence response decisions, such as gender or education, can then be compared between responders and nonresponders. A larger difference between the groups of responders and nonresponders would suggest more nonresponse bias.

A more recent research trend that also requires further examination in the context of web surveys is the increase in mobile web surveys. Findings suggest that their breakoff rates are significantly higher than those rates found in surveys that are completed via PC (Mavletova & Couper, 2015). Since the future of web surveys appears to be moving toward implementation via mobile devices, it is vital that research focuses on the optimization of web response rates by investigating the effects of design factors on survey participation and breakoff.

References

* Studies included in the meta-analysis.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 171–211. https://doi.org/10.1016/0749-5978(91)90020-T

*Al Atassi, H., Shapiro, M. C., Rao, S. R., Dean, J., & Salama, A. (2018). Oral and maxillofacial surgery resident perception of personal achievement and anxiety: A cross-sectional analysis. *Journal of Oral and Maxillofacial Surgery, 76*, 2532–2539. https://doi.org/10.1016/j.joms.2018.06.018

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th edition). AAPOR

*Axisa, C., Nash, L., Kelly, P., & Willcock, S. (2019). Psychiatric morbidity, burnout and distress in Australian physician trainees. *Australian Health Review*. https://doi.org/10.1071/AH18076

Baur, N. (2014). Comparing societies and cultures: challenges of cross-cultural survey research as an approach to spatial analysis. *Historical Social Research, 39*(2), 257–291. https://doi.org/10.12759/hsr.39.2014.2.257-291

Beevers, C., & Meyer, B. (2002). Lack of positive experiences and positive expectancies mediate the relationship between BAS responsiveness and depression. *Cognition and Emotion, 16*, 549–564. https://doi.org/10.1080/02699930143000365

Bijttebier, P., Beck, I., Claes, L., & Vandereycken, W. (2009). Gray's reinforcement sensitivity theory as a framework for research on personality-psychopathology associations. *Clinical Psychology Review, 29*, 421–430. https://doi.org/10.1016/j.cpr.2009.04.002

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science, 645*, 36–59. https://doi.org/10.1177/0002716212456834

Bošnjak, M., Tuten, T. L., & Wittmann, W. W. (2005). Unit (non) response in web-based access panel surveys: An extended planned-behavior approach. *Psychology and Marketing, 22*, 489–505. https://doi.org/10.1002/mar.20070

Bošnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., & Kaczmire, L. (2008). Prenotification in Web-based access panel surveys – The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review, 26*, 213–223. https://doi.org/10.1177/0894439307305895

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates. *Public Opinion Quarterly, 57*, 62–79. https://doi.org/10.1086/269355

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or Internet-based surveys. *Educational and Psychological Measurement, 60*, 821–836. https://doi.org/10.1177/00131640021970934

Corr, P. J. (2002). J. A. Gray's reinforcement sensitivity theory: Tests of the joint subsystems hypothesis of anxiety and impulsivity. *Personality and Individual Differences, 33*, 511–532. https://doi.org/10.1016/S0191-8869(01)00170-2

*Crawford, N. M., Hoff, H. S., & Mersereau, J. E. (2017). Infertile women who screen positive for depression are less likely to initiate fertility treatments. *Human Reproduction, 32*, 582–587. https://doi.org/10.1093/humrep/dew351

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2019). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology, smz008*, 1–27. https://doi.org/10.1093/jssam/smz008

*De Graaff, A. A., Van Lankveld, J., Smits, L. J., Van Beek, J. J., & Dunselman, G. A. (2016). Dyspareunia and depressive symptoms are associated with impaired sexual functioning in women with endometriosis, whereas sexual functioning in their male partners is not affected. *Human Reproduction, 31*, 2577–2586. https://doi.org/10.1093/humrep/dew215

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., & Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys. *Public Opinion Quarterly, 71*, 413–443. https://doi.org/10.1093/poq/nfm014

Esser, H. (1986). Über die Teilnahme an Befragungen [Participation in opinion polls]. *ZUMA Nachrichten, 10*, 38–47. SSOAR – Social Science Open Access Repository.

Esser, H. (2001). *Soziologie. Sinn und Kultur* (Vol. 6). Frankfurt am Main, New York: Campus.

Galesic, M., & Bošnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*, 349–360. https://doi.org/10.1093/poq/nfp0313

*Goodwin, G. M., Price, J., De Bodinat, C., & Laredo, J. (2017). Emotional blunting with antidepressant treatments: A survey among depressed patients. *Journal of Affective Disorders, 221*, 31–35. https://doi.org/10.1016/j.jad.2017.05.048

*Gouttebarge, V., Jonkers, R., Moen, M., Verhagen, E., Wylleman, P., & Kerkhoffs, G. (2017). The prevalence and risk indicators of symptoms of common mental disorders among current and former Dutch elite athletes. *Journal of Sports Sciences, 35*, 2148–2156. https://doi.org/10.1080/02640414.2016.1258485

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science, 24*, 1722–1731. https://doi.org/10.1177/0956797613479387

Grossmann, I., & Varnum, M. E. W. (2015). Social structure, infectious diseases, disasters, secularism, and cultural change in America. *Psychological Science, 26*, 311–324. https://doi.org/10.1177%2F0956797614563765

Groves, R. M., & Magilavy, L. J. (1981). Increasing response rates to telephone surveys: A door in the face for foot-in-the-door? *Public Opinion Quarterly, 45*, 346–358. https://doi.org/10.1086/268669

Groves, R. M., Cialdini, R. B., & Couper, M. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*, 475–495. https://www.jstor.org/stable/2749203

Groves, R. M., Fowler, F. J., Couper, M. P., Lepowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Wiley Series in Survey Methodology. Hoboken: Wiley.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly, 72*, 167–189. https://doi.org/10.1093/poq/nfn011

Hamamura, T. (2011). Are cultures becoming individualistic? A cross-temporal comparison of individualism-collectivism in the United States and Japan. *Personality and Social Psychology Review, 16*, 3–24. https://doi.org/10.1177/1088868311411587

*Han, K., Bohnen, J., Peponis, T., Martinez, M., Nandan, A., Yeh, D. D., ... Kaafarani, H. M. (2017). Surgeon as the second victim? Results of the Boston Intraoperative Adverse Events Surgeons' Attitude (BISA) Study. *Journal of the American College of Surgeons, 224*, 1048–1056. https://doi.org/10.1016/j.jamcollsurg.2016.12.039

*Harmark, L., Van Puijenbroek, E., & Van Grootheest, K. (2013). Intensive monitoring of duloxetine: Results of a web-based intensive monitoring study. *European Journal of Clinical Pharmacology, 69*, 209–215. https://doi.org/10.1007/s00228-012-1313-7

Haunberger, S. (2011a). To participate or not to participate: Decision processes related to survey non-response. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 109*, 39–55. https://doi.org/10.1177/0759106310387721

Haunberger, S. (2011b). Explaining unit nonresponse in online panel surveys: An application of the extended theory of planned behavior. *Journal of Applied Social Psychology, 41*, 2999–3025. https://doi.org/10.1111/j.1559-1816.2011.00856.x

*Hoff, H. S., Crawford, N. M., & Mersereau, J. E. (2015). Mental health disorders in infertile women: Prevalence, perceived effect on fertility, and willingness for treatment for anxiety and depression. *Fertility and Sterility, 104*, e357. https://doi.org/10.1016/j.fertnstert.2015.07.1113

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Hox, J. J., & de Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity, 28*, 329–344. https://doi.org/10.1007/BF01097014

*Kikuchi, T., Uchida, H., Suzuki, T., Watanabe, K., & Kashima, H. (2011). Patients' attitudes toward side effects of antidepressants: An Internet survey. *European Archives of Psychiatry and Clinical Neuroscience, 261*, 103–109. https://doi.org/10.1007/s00406-010-0124-z

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). New York, NY: Wiley.

Lozar Manfreda, K., Bošnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*, 79–104. https://doi.org/10.1177/147078530805000107

Mavletova, A., & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies* (pp. 81–98). London, UK: Ubiquity Press.

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly, 79*, 105–129. https://doi.org/10.1093/poq/nfu059

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*, e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nair, C. S., Adams, P., & Mertova, P. (2008). Student engagement: The key to improving survey response rates. *Quality in Higher Education, 14*, 225–232. https://doi.org/10.1080/13538320802507505

Ogihara, Y. (2017). Temporal changes in individualism and their ramification in Japan: Rising individualism and conflicts with persisting collectivism. *Frontiers in Psychology, 8*, 695. https://doi.org/10.3389/fpsyg.2017.00695

Ogihara, Y. (2018). The rise in individualism in Japan: Temporal changes in family structure, 1947–2015. *Journal of Cross-Cultural Psychology, 49*, 1219–1226. https://doi.org/10.1177/0022022118781504

*Paiva, C. E., Martins, B. P., & Paiva, B. S. R. (2018). Doctor, are you healthy? A cross-sectional investigation of oncologist burnout, depression, and anxiety and an investigation of their associated factors. *BMC Cancer, 18*, 1044. https://doi.org/10.1186/s12885-018-4964-7

*Peth, J., Jelinek, L., Nestoriuc, Y., & Moritz, S. (2018). Adverse effects of psychotherapy in depressed patients – First application of the Positive and Negative Effects of Psychotherapy Scale (PANEPS). *Psychotherapie Psychosomatik Medizinische Psychologie, 68*, 391–398. https://doi.org/10.1055/s-0044-101952

*Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fra□ßdorf, A., ... Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly, 79*, 740–768. https://doi.org/10.1093/poq/nfv014

Pinto-Meza, A., Caseray, X., Soler, J., Puigdemont, D., Perez, V., & Torrubia, R. (2006). Behavioral inhibition and behavioural activation systems in current and recovered major depression participants. *Personality and Individual Differences, 40*, 215–226. https://doi.org/10.1016/j.paid.2005.06.021

*Prinz, B., Dvorak, J., & Junge, A. (2016). Symptoms and risk factors of depression during and after the football career of elite female players. *BMJ Open Sport & Exercise Medicine, 2*, e000124. https://doi.org/10.1136/bmjsem-2016-000124

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health, 14*, 1101–1108. https://doi.org/10.1016/j.jval.2011.06.00

*Salles, A., Wright, R. C., Milam, L., Panni, R. Z., Liebert, C. A., Lau, J. N., Lin, D. T., & Mueller, C. M. (2019). Social belonging as a predictor of surgical resident well-being and attrition. *Journal of Surgical Education, 76*, 370–377. https://doi.org/10.1016/j.jsurg.2018.08.022

Santos, H. C., Varnum, M. E. W., & Grossmann, I. (2017). Global increases in individualism. *Psychological Science, 28*, 1228–1239. https://doi.org/10.1177/0956797617700622

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44*, 409–432. https://doi.org/10.1023/A:1024232915870

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard E-mail methodology: Results of an experiment. *Public Opinion Quarterly, 62*, 378–397. https://doi.org/10.1086/297851

*Schuring, N., Kerkhoffs, G., Gray, J., & Gouttebarge, V. (2017). The mental wellbeing of current and retired professional cricketers: an observational prospective cohort study. *Physician & Sports Medicine, 45*, 463–469. https://doi.org/10.1080/00913847.2017.1386069

Schwarz, N. (2003). Culture-sensitive context effects: A challenge for cross-cultural surveys. In J. Harkness, F. J. R. Van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 93–100). New Jersey: Wiley.

Shi, T., & Fan, X. (2008). Comparing response rates from web and mail surveys. *Field Methods, 20*, 249–271. https://doi.org/10.1177/1525822X08317085

*Shigemura, J., Sato, Y., Yoshino, A., & Nomura, S. (2008). Patient satisfaction with antidepressants: An Internet-based study. *Journal of Affective Disorders, 107*, 155–160. https://doi.org/10.1016/j.jad.2007.08.019

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ, 343*, d4002. https://doi.org/10.1136/bmj.d4002

*Strohmeier, H., Scholte, W. F., & Ager, A. (2018). Factors associated with common mental health problems of humanitarian workers in South Sudan. *PLoS One, 13*, e0205333. https://doi.org/10.1371/journal.pone.0205333

Twenge, J. M., Dawson, L., & Campbell, W. K. (2016). Still standing out: children's names in the United States during the Great Recession and correlations with economic indicators. *Journal of Applied Social Psychology, 46*, 663–670. https://doi.org/10.1111/jasp.12409

Van Horn, P. S., Green, K. E., & Martinussen, M. (2009). Survey Response Rates and Survey Administration in Counseling and Clinical Psychology. *Educational and Psychological Measurement, 69*, 389–403. https://doi.org/10.1177/0013164408324462

*Van Overveld, M., De Jong, P. J., Peters, M. L., Van Hout, W. J. P. J., & Bouman, T. (2008). An internet-based study on the relation between disgust sensitivity and emetophobia. *Journal of Anxiety Disorders, 22*, 524–531. https://doi.org/10.1016/j.janxdis.2007.04.001

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Weiner, S. P., & Dalessio, A. T. (2006). Oversurveying: Causes, consequences, and cures. In A. I. Kraut (Ed.), *Getting action from organizational surveys: New concepts, technologies, and applications* (pp. 294–311). Chapter 12. San Francisco, CA: Jossey-Bass.

*Williams, E., Martin, S. L., Fabrikan, A., Wang, A., & Pojasek, M. (2018). Rates of depressive symptoms among pharmacy residents. *American Journal of Health-System Pharmacy, 75*, 292–297. https://doi.org/10.2146/ajhp161008

World Bank, World Development Indicators. (2018). *Individuals using the Internet (% of population)* [Data file]. Washington, DC: The World Bank. Retrieved from https://data.worldbank.org/indicator/IT.NET.USER.ZS

WHO. (2017). *Depression and other common mental disorders: Global health estimates*. Geneva, Switzerland: World Health Organization. Retrieved from https://apps.who.int/iris/handle/10665/254610

Zeng, R., & Greenfield, P. M. (2015). Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. *International Journal of Psychology, 50*, 47–55. https://doi.org/10.1002/ijop.12125

*Zimmerman, M., Chelminski, I., Young, D., & Dalrymple, K. (2011). Using outcome measures to promote better outcomes. *Clinical Neuropsychiatry, 8*, 28–36.

History

Received August 13, 2019
Revision received December 4, 2019
Accepted December 4, 2019
Published online March 31, 2020

Open Data

Table S1: Inclusion and exclusion criteria for study selection
Table S2: Results from database searches
Table S4: Descriptive information for each included study
Supplement S5: Search strategies
Available at https://doi.org/10.23668/psycharchives.2626
Figure S3: Flow chart for the selection process of records for the meta-analysis
Available at https://doi.org/10.23668/psycharchives.2627

ORCID

Tanja Burgard
https://orcid.org/0000-0001-9194-4821

Michael Bošnjak
https://orcid.org/0000-0002-1431-8461

Nadine Wedderhoff
https://orcid.org/0000-0002-4460-4995

Tanja Burgard

Leibniz Institute for Psychology Information (ZPID)
Universitätsring 15
54296 Trier
Germany
burgard@leibniz-psychology.org

Appendix B: Supplementary Material of Study 1

Table B1: Inclusion and Exclusion Criteria for Study Selection

Item	Included	Excluded
Population	Studies consisting of at least one sample or subgroup with at least 30% of adults (> 18 years) with depression or anxiety disorder	Student participants Studies reporting on children or adolescents < 18 years of age Individuals with postpartum depression Individuals with bipolar disorder
Outcomes	% response rate*	-
Study type	Experimental psychological studies of any design that report the results of ***online surveys*** only. And at least one of the following: o **Design of the invitation letter** *(personalized, nonpersonalized, or none)* o **Burden of participation** *(time spent, effort required, cognitive complexity, frequency of participation, amount of stress)* o **Incentives for participation** *(monetary, nonmonetary)*	Studies reporting on any survey type other than online surveys, including face to face interviews, telephone interviews, or mail surveys. Studies reporting on mixed survey types that do not explicitly report on an online survey subgroup. Case reports and case studies reporting on <20 participants. Panel studies that do not report results from the first wave. Review articles and editorials.

Studies were not restricted based on publication date, language, or publication format.

Figure B1: PRISMA Flow Chart of the Literature Selection Process

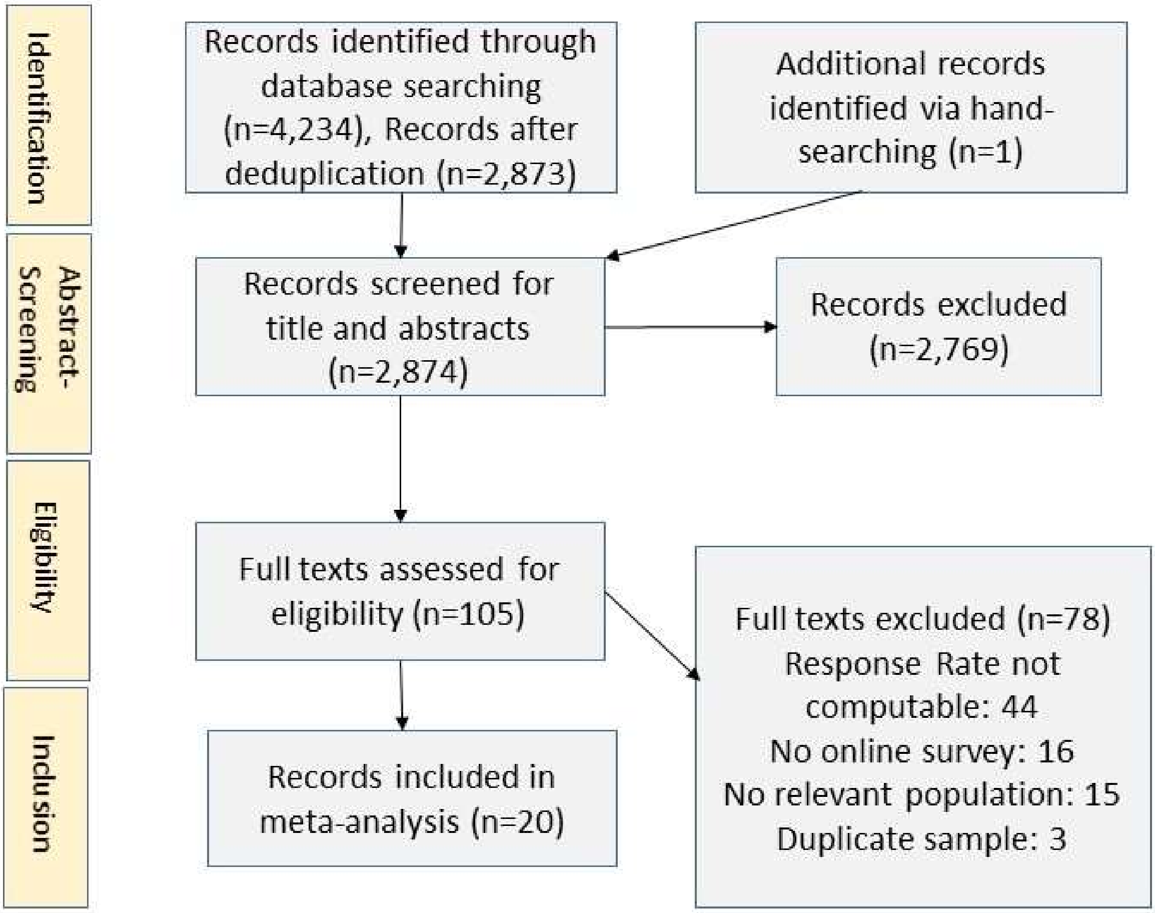

Identification
Abstract-Screening
Eligibility
Inclusion
Records identified through database searching (n=4,234), Records after deduplication (n=2,873)
Additional records identified via hand-searching (n=1)
Records screened for title and abstracts (n=2,874)
Records excluded (n=2,769)
Full texts assessed for eligibility (n=105)
Full texts excluded (n=78)
Response Rate not computable: 44
No online survey: 16
No relevant population: 15
Duplicate sample: 3
Records included in meta-analysis (n=20)

Table B2: Descriptive Information for Each Included Study

Author	Year	Country	Mean age	Sample size	RR	SD (RR)
Al Atassi	2018	USA	30	238	0.2	0.0116
Axisa	2019	Australia	30.4	59	0.07	0.0082
Crawford	2017	USA	35.1	416	0.43	0.0160
De Graaff	2016	Netherlands	34.3	83	0.59	0.0415
Goodwin	2017	CAN, USA, UK	49.5	819	0.11	0.0035
Gouttebarge	2016	Netherlands	27.3	203	0.28	0.0167
Han	2017	USA	49.4	126	0.45	0.0297
Härmark	2013	Netherlands	47	256	0.64	0.0240
Hoff	2015	USA	36	414	0.43	0.0160
Kikuchi	2011	Japan	37.2	1,187	0.31	0.0074
Paiva	2018	Brazil	34	227	0.7	0.0254
Peth	2018	Germany	46.9	135	0.19	0.0147
Prinz	2016	Germany	33	157	0.64	0.0306
Shigemura	2008	Japan	34.6	1,199	0.75	0.0010
Salles	2019	USA	30.4	169	0.66	0.0296
Schuring	2017	South Africa	27	78	0.45	0.0376
Strohmeier	2018	South Sudan	37	277	0.09	0.0050
Van Overveld	2008	Netherlands	25.4	138	0.4	0.0262
Williams	2018	USA	26.9	701	0.21	0.0071
Zimmermann	2011	USA	47	30	0.98	0.0226

Appendix C: Original Publication of Study 2

Originalarbeit

Konditionierungseffekte in Panel-Untersuchungen

Systematische Übersichtsarbeit und Meta-Analyse am Beispiel sensitiver Fragen

Tanja Burgard[1], Michael Bosnjak[1,2] und Nadine Wedderhoff[2]

[1]Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID), Trier
[2]Universität Trier

Zusammenfassung: Paneldaten sind für die Untersuchung kausaler Zusammenhänge und die Beantwortung längsschnittlicher Fragestellungen unverzichtbar. Es ist allerdings umstritten, welchen Effekt die wiederholte Befragung von Panelteilnehmern auf die Qualität von Paneldaten hat. Der zu erwartende Lerneffekt der Teilnehmer bei wiederholter Teilnahme wird als Panelkonditionierung bezeichnet und kann sowohl positive als auch negative Folgen für die Validität der Paneldaten aufweisen. Insbesondere bei sensitiven Items werden Auswirkungen auf die soziale Erwünschtheit der gemachten Angaben erwartet. Die verfügbare Evidenz zu Konditionierungseffekten bei sensitiven Fragen legt unterschiedliche Effekte je nach Art der Frage nahe und wurde bisher lediglich in Form narrativer Reviews aufgearbeitet. In der vorliegenden Meta-Analyse werden anhand der verfügbaren experimentellen Evidenz (154 Effektstärken aus 19 Berichten) Konditionierungseffekte in Abhängigkeit von der Art der Frage, sowie der Häufigkeit und der Abstände zwischen den Erhebungen (Dosiseffekte) untersucht. Standardisierte Mittelwertunterschiede zwischen wiederholt teilnehmenden und erstmalig teilnehmenden Probanden werden mittels Mehrebenen-Meta-Regressionen analysiert. Dabei zeigen sich nur geringe Effekte vorheriger Befragungen auf das Antwortverhalten in Folgewellen. Nach aktuellem Stand kann daher davon ausgegangen werden, dass die Qualität von Paneldaten nicht in relevantem Maße von Konditionierungseffekten beeinflusst wird. Grenzen der vorliegenden Meta-Analyse und relevante Forschungslücken werden diskutiert.

Schlüsselwörter: Konditionierungseffekte, Meta-Analyse, Panel-Erhebungen, Sensitive Items, Soziale Erwünschtheit

Conditioning Effects in Panel Studies. Systematic Review and Meta-Analysis for Sensitive Items

Abstract: Panel data are indispensable for investigating causal relationships and answering longitudinal questions. However, it is controversial how the repeated survey of panel participants affects the quality of panel data. The expected learning effect of repeated participation is called panel conditioning and can have both positive and negative consequences for the validity of panel data. Sensitive items in particular are expected to have an impact on the social desirability of the information provided. The available evidence on conditioning effects for sensitive questions suggests different effects depending on the type of question and has so far only been processed in the form of narrative reviews. In the present meta-analysis, conditioning effects are examined on the basis of the available experimental evidence (154 effect strengths from 19 reports), depending on the type of question, as well as the frequency and intervals between surveys (dosage effects). Standardized mean differences between experienced and fresh participants are analyzed by multi-level meta-regressions. The effects of previous surveys on the response behaviour in subsequent waves are only minor. At present, it can therefore be assumed that the quality of panel data is not influenced to a relevant extent by conditioning effects. Limits of the present meta-analysis and relevant research gaps are discussed.

Keywords: conditioning effects, meta-analysis, panel data, sensitive items, social desirability

Relevanz und Effekte von Panelkonditionierung

Um längsschnittliche Fragestellungen zu beantworten und kausale Schlüsse ziehen zu können, sind Paneldaten unverzichtbar. Da ihre Erhebung und die langfristige Pflege eines Teilnehmerpools aufwändig und teuer sind, gibt es in vielen Disziplinen offene Panel-Infrastrukturen, die dies übernehmen und der Forschungsgemeinschaft zur Verfügung stehen. Beispiele sind das GESIS Panel (Bosnjak et al., 2018), die Understanding America Study (Alattar, Rogofsky & Messel, 2018), KAMOS (Cho, LoCascio, Lee, Jang & Lee, 2017) und das LISS Panel (Beschreibung dieser Infrastrukturen: Das, Kapteyn & Bosnjak, 2018; Weiß et al., 2020). Dadurch werden Ressourcen gebündelt und

https://doi.org/10.1026/0033-3042/a000479

die Objektivität der Erhebung erhöht. Auch für die Psychologie wird eine Stärkung solcher Infrastrukturen gefordert (Bruder, Göritz, Reips & Gebhard, 2014).

Um den Aufbau und die Nutzung eines solchen Labors am ZPID – Leibniz Institut zu begleiten, stellt sich auch die Frage, welche Faktoren die Qualität von Paneldaten beeinflussen. Ein bekanntes Risiko für die Validität von Paneldaten ist Panelmortalität (Sobol, 1959). Es ist in vielen Fällen davon auszugehen, dass das Ausscheiden von Panelteilnehmern in späteren Wellen mit relevanten Untersuchungsvariablen zusammenhängt und dass die Stichprobe somit im Laufe der Zeit immer weniger ein repräsentatives Abbild der Grundgesamtheit darstellt. Ein weiterer Effekt, der neben der Panelmortalität potentiell problematisch sein könnte, ist die sogenannte Panelkonditionierung. Diese kann sowohl positive als auch negative Auswirkungen auf die Datenqualität haben.

Grundsätzlich beschreibt Panelkonditionierung einen auf das Antwortverhalten bezogenen Lerneffekt in Panelstudien. Dabei wird angenommen, dass sich erfahrene Teilnehmer im Vergleich zu erstmals teilnehmenden Panelisten anders verhalten und dass sie anders antworten. Im Gegensatz zur Panelmortalität ist Panelkonditionierung allerdings nicht von Vornherein mit negativen Konsequenzen auf die Datenqualität assoziiert. Ganz im Gegenteil können diesem Effekt unterschiedliche Mechanismen zugrunde liegen, die sich sowohl positiv als auch negativ auf die Validität der Daten auswirken können und die hier beispielhaft entlang des einschlägigen Modells des Befragungsprozesses nach Tourangeau, Rips und Rasinski (2000) skizziert werden. Nach diesem Modell sind für die Beantwortung einer Frage mindestens vier Schritte notwendig: Das Verständnis der Frage, der Abruf von relevanten Informationen, die Verarbeitung und Beurteilung der Informationen und die Auswahl der entsprechenden Antwortmöglichkeit.

Bereits das Verständnis einer Frage könnte durch vorherige Umfrageerfahrungen beeinflusst sein. Ein erfahrener Teilnehmer könnte sowohl die Frage, als auch die Antwortoptionen besser verstehen, die Regeln des Interviews besser kennen und die Instruktionen und deren Zielsetzung genauer identifizieren. Diese kognitive Entlastung im Hinblick auf die Instruktion könnte dazu führen, dass Befragte häufiger Meinungen angeben, anstatt Kategorien wie „Weiß nicht“ auszuwählen, insbesondere bei komplexen Einstellungsfragen (Binswanger, Schunk & Toepoel, 2013). Überdies könnte eine Befragung Reflektionsprozesse über die Befragung hinaus in Gang setzen und zu einer größeren Aufmerksamkeit für Themen der Befragung und damit einhergehend zur Auseinandersetzung mit diesen führen (Sturgis, Allum & Brunton-Smith, 2009). Durch diese kognitive Stimulation könnten sich sowohl Einstellungen, als auch Wissensbestände (z. B. zu demografischen Angaben wie dem Einkommen; Fisher, 2019) ändern und einen Effekt auf Angaben in Folgewellen haben. Neben der kognitiven Stimulation in Bezug auf Themen der Befragung könnten Erinnerungseffekte auch zur Angabe kohärenterer Einstellungen führen (Bergmann & Barth, 2018).

Auch die Verarbeitung und Beurteilung der abgerufenen Informationen in Folgewellen könnte durch vorherige Befragungen beeinflusst sein. Bei einer sich über die Zeit einstellenden ‚Befragungsmüdigkeit‘ könnten Probanden versuchen, zufriedenstellende Antworten zu geben und dabei den kognitiven und zeitlichen Aufwand der Befragung gering zu halten. Krosnick (1991) nennt dieses Verhalten Satisficing. Dazu gehört auch das sogenannte Speeding (Schonlau & Toepoel, 2015), ein möglichst schnelles Beantworten von Fragen, ohne tatsächlich nach der akkuratesten Antwort zu suchen.

Auch bei der Auswahl einer passenden Antwort könnte der Aufwand der Befragung noch reduziert werden. Kennen Befragte die Regeln des Interviews von vorherigen Befragungen, könnten sie zum Beispiel durch die negative Beantwortung von Filterfragen (Kreuter, McCulloch, Presser & Tourangeau, 2011) oder die Angaben eines kleineren sozialen Netzwerks Folgefragen vermeiden und die Befragung so abkürzen (Silber et al., 2019).

Da bei sensitiven Items in besonderer Weise von Konditionierungseffekten, etwa hinsichtlich sozialer Erwünschtheit, auszugehen ist, konzentrieren wir uns in der vorliegenden Meta-Analyse auf diese. Ein Item kann als sensitiv eingestuft werden, wenn es mindestens eine der drei folgenden Eigenschaften besitzt (Tourangeau, Rips & Rasinski, 2000): 1. Frage fordert sozial unerwünschte Antwort (z. B.: Konsumieren Sie regelmäßig illegale Drogen?), 2. Frage wird als zudringlich und privat empfunden (z. B.: Wie viele Sexualpartner hatten Sie in den letzten drei Jahren?), 3. Frage ist in besonderer Weise datenschutzrechtlich relevant (z. B.: Haben Sie im letzten Jahr Einnahmen erzielt, die Sie dem Finanzamt nicht gemeldet haben?).

Die Evidenz zu Panelkonditionierungseffekten bei sensitiven Fragen legt unterschiedliche Wirkungsrichtungen hinsichtlich der Datenqualität je nach Art der Frage nahe. Aufgrund der größeren Vertrautheit mit der Interviewsituation befürchten Befragte weniger Konsequenzen und antworten ehrlicher auf Einstellungsfragen. Somit ist mit weniger sozial erwünschten Antworten bei erfahrenen Teilnehmern im Vergleich zu neuen Teilnehmern zu rechnen (z. B. Phillips & Clancy, 1972; Fowler, 1995; Nancarrow & Cartwright, 2007; Binswanger, Schunk & Toepoel, 2013). Die erste Hypothese geht damit von einer Reduzierung der Verzerrung durch soziale Erwünschtheit bei Einstellungsfragen aus:

H1: Erfahrene Panelteilnehmer antworten bei sensitiven Einstellungsfragen weniger sozial erwünscht als neue Panelteilnehmer.

Im Falle sozial unerwünschter Verhaltensweisen wird argumentiert, dass das Berichten solcher Handlungen negative Emotionen auslöst, wie zum Beispiel Schuld, Scham oder Angst und dadurch ein reflexiver Prozess in Gang gesetzt wird, der zur Anpassung der Antworten in Richtung sozialer Konformität in Folgewellen führt (Baumeister, Vohs, DeWall & Zhang, 2007). Insbesondere in Studien zum Drogenmissbrauch von Jugendlichen und jungen Erwachsenen wurden Hinweise für diesen sogenannten Recanting-Effekt (Percy et al., 2005) gefunden. Dabei wird bereits berichteter Drogenmissbrauch in Folgewellen bestritten (Torche, Valenzuela, Warren & Halpern-Manners, 2012). Ähnliche Effekte wurden auch im Kontext anderer sensitiver Verhaltensfragen festgestellt (Williams, Block & Fitzsimons, 2006; Fitzsimons & Moore, 2008; Halpern-Manners, Warren & Torche, 2014). Daraus folgt die Annahme einer Verstärkung sozialer Erwünschtheitseffekte bei sensitiven Verhaltensfragen bei zunehmender Befragungshäufigkeit:

H2: Erfahrene Panelteilnehmer antworten bei sensitiven Verhaltensfragen eher sozial erwünscht als neue Panelteilnehmer.

Eine generelle Annahme bei Konditionierungseffekten, die sowohl für Einstellungs- als auch für Verhaltensfragen gilt, ist die Existenz von Dosiseffekten. Das heißt:

H3: Je häufiger die Versuchsgruppe bereits befragt wurde, desto stärker der Konditionierungseffekt, also die Differenz der standardisierten Mittelwerte von erfahrenen und neuen Teilnehmern.

H4: Je größer der zeitliche Abstand zur vorherigen Erhebung, desto schwächer der Konditionierungseffekt.

Methoden

Um kausale Rückschlüsse auf Konditionierungseffekte ziehen zu können, sind für diese Meta-Analyse (quasi-)experimentelle Studien des Antwortverhaltens in Panelbefragungen relevant. Dazu müssen die Antworten einer bereits zuvor befragten Versuchsgruppe und einer noch nicht durch Befragung konditionierten Kontrollgruppe zum selben Zeitpunkt verglichen werden. Beide Gruppen sollen dabei nach denselben sensitiven Items befragt worden sein, so dass die korrespondierenden Verhaltensweisen oder Einstellungen miteinander verglichen werden können.

Aus den relevanten Studien wurden Informationen auf drei Ebenen extrahiert: 1. Generelle Angaben zum Studienbericht (Autor, Publikationsjahr), 2. Beschreibung der Intervention (Art der Frage, Häufigkeit der Befragung), 3. Quantitative Ergebnisse beider Gruppen (Mittelwerte, Anteilswerte, Teststatistiken, Standardfehler). Eine vollständige Übersicht aller verwendeten Codierungskategorien sind in OD 1: Anhang 1 dokumentiert.

Eine erste Literatursuche wurde im Dezember 2017 mit der Meta-Suchmaschine CLICsearch durchgeführt. In OD 2: Anhang 2 sind alle in CLICsearch enthaltenen Datenbanken aufgelistet. Neben „Panel Conditioning“ wurden 15 synonyme Suchbegriffe verwendet (siehe OD 3: Anhang 3). Mit den nach dem ersten Screening identifizierten relevanten Artikeln wurde zusätzlich eine manuelle Vorwärts- und Rückwärtssuche durchgeführt. Das heißt, alle zitierten und alle verweisenden Literatureinträge wurden geprüft.

Die berechnete Effektstärke sind standardisierte Mittelwertdifferenzen. Diese werden bei der Kodierung und Berechnung so gerichtet, dass positive Werte bedeuten, dass erfahrene Panelisten weniger sozial erwünscht antworten und negative Werte für eine höhere soziale Erwünschtheit der Angaben in der Versuchsgruppe stehen. Zur Prüfung der Hypothesen und vor dem Hintergrund der hierarchischen Datenstruktur (mehrere Effektstärken sind beispielsweise derselben Studie entnommen worden) werden Multilevel-Meta-Regressionen eingesetzt (Van den Noortgate, López-López, Marín-Martínez & Sánchez-Meca, 2013). Ein Multilevel-Modell erlaubt auch, die Verteilung der Varianz in den Effektstärken durch Variablen auf unterschiedlichen Ebenen zu erklären (Assink & Wibbelink, 2016). In OD 4: Anhang 4 ist dokumentiert und begründet, warum bei der vorliegenden Meta-Analyse ein Drei-Ebenen-Modell gewählt wurde.

Alle Analysen wurden mit dem R-Paket metafor, Version 2.0–0 (Viechtbauer, 2010) durchgeführt.

Ergebnisse

Insgesamt wurden 2 355 Artikel daraufhin untersucht, ob sie die Inklusionskriterien erfüllen. Anhand der Abstracts konnten 2 127 Artikel ausgeschlossen werden. Die übrigen wurden genauer geprüft und weitere 209 Artikel wurden aufgrund des Studiendesigns, der Ergebnisdokumentation oder der Nicht-Sensitivität der berichteten Items nicht in die Meta-Analyse einbezogen. Die letztlich ausgewählten 19 Berichte enthalten 85 Stichproben und 154 Effektstärken. Die entsprechenden Literaturselektionsschritte sind in Form eines PRISMA-Flussdiagramms (Moher, Li-

Tabelle 1. Übersicht der Studiencharakteristika der ausgewählten Primärstudien

Erstautor	Erscheinungsjahr	Anzahl Stichproben	Anzahl Vergleiche	Art der Fragen	Ort der Studie
Struminskaya	2016	1	1	Verhalten	DEU
Halpern-Manners	2014	2	8	Verhalten	USA
Das	2011	1	1	Verhalten	Niederlande
Toepoel	2008	1	6	Verhalten	Niederlande
Torche	2012	1	4	Verhalten	Chile
Quick	2017	1	5	Verhalten	England
Fitzsimons	2007	2	3	Verhalten	USA
Axinn	2015	2	6	Verhalten	USA
Clinton	2001	4	4	Verhalten	USA
Murray	1988	2	6	Verhalten	England
Song	2017	56	56	Verhalten	USA
Williams	2006	1	2	Verhalten	USA
Barber	2016	1	4	Einstellungen	USA
Halpern-Manners	2017	1	2	Einstellungen	USA
Bridge	1977	1	1	Einstellungen	USA
Kraut	1973	2	2	Einstellungen	USA
Warren	2012	4	33	Beides	USA / DEU
Waterton	1989	1	5	Beides	GB
Coombs	1973	1	5	Beides	Taiwan
		85	154		

berati, Tetzlaff & Altman, 2009) in OD 5: Anhang 5 zu finden.

Tabelle 1 liefert eine Übersicht einiger Merkmale der 19 Studienberichte, die in die Meta-Analyse einbezogen wurden. Neben Autor und Erscheinungsjahr ist hier auch ersichtlich, wie sich die Anzahl der Stichproben und der Gruppenvergleiche insgesamt auf die Publikationen aufteilt. Die meisten Studien berichten Ergebnisse von höchstens zwei Stichproben. Die Effekte von Verhaltensfragen wurden häufiger untersucht (n = 116) als die von Einstellungsfragen (n = 38). Die meisten Studien wurden in den USA durchgeführt. Die mittleren standardisierten Mittelwertdifferenzen auf Ebene der Studienberichte sind überwiegend nahe Null, der Unterschied zwischen Kontroll- und Versuchsgruppe ist also gering. Dieser Befund spricht für keine oder geringe Konditionierungseffekte. Es gibt allerdings einige Studien, die mittlere bis starke Effekte nahelegen, sowohl in Richtung höherer (negative SMD), als auch niedrigerer sozialer Erwünschtheit (positive SMD). Einen Überblick über die Verteilung aller Effektstärken bietet auch der Funnelplot in OD 6: Anhang 6.

Um mögliche Konditionierungseffekte über alle Studien hinweg zu quantifizieren, wurden Meta-Analysen mit drei Analyseebenen (Stichprobenvarianz der Effektstärken, Varianz zwischen den Effektstärken, Varianz zwischen den Studienberichten) berechnet.

Alle geschätzten Effekte sind sehr klein (Ferguson, 2009) und nicht signifikant, wie in Tabelle 2 zu sehen ist. Über alle 154 Effektstärken hinweg ist die mittlere standardisierte Mittelwertdifferenz positiv. Erfahrene Panelisten antworten demnach über alle Studien und Fragetypen hinweg weniger sozial erwünscht als neue Teilnehmer. In Hypothese 1 wurde für Einstellungsfragen genau dies angenommen. Tatsächlich ist das Vorzeichen des geschätzten Effekts für Einstellungsfragen wie erwartet positiv, allerdings ist der Effekt nicht signifikant. Hypothese 2, in der für Verhaltensfragen eine höhere Konformität mit sozialen Normen in der Versuchsgruppe erwartet wurde, findet ebenfalls keine Bestätigung. Insgesamt lässt sich daraus ableiten, dass für Verhaltens- und Einstellungsfragen nicht mit erheblichen Konditionierungseffekten in Panelerhebungen zu rechnen ist. Die Unterschiede zwischen den Gruppen und somit auch der Einfluss vorheriger Befragungen auf die Datenqualität sind sehr gering.

Mit den Hypothesen 3 und 4 wurden Dosiseffekte konstatiert. Zum einen sollten Konditionierungseffekte durch häufigere Befragung stärker werden (Hypothese 3). Zum anderen wurde angenommen, dass mit zunehmendem Abstand zwischen den Wellen der Einfluss vorheriger Befragungen auf die erneute Messung abnehmen sollte (Hypothese 4). Zum Testen der Hypothesen 3 und 4 wird als Effektstärke die absolute standardisierte Mittelwertdifferenz verwendet. Während bei den Hypothesen 1 und 2 die

Tabelle 2. Gesamteffekt und Effekte nach Fragetyp

	Gesamteffekt	Hypothese 1	Hypothese 2
Intercept	0.040 [-0.044; 0.124]; p = .343	–	–
Einstellungen (n = 38)	–	0.065 [-0.028; 0.159]; p = .170	–
Verhalten (n = 116)	–	–	0.032 [-0.053; 0.117]; p = .462

Anmerkungen: n = 154 Effektstärken, *k* = 19 Studienberichte, Effektstärke: gerichtete standardisierte Mittelwertdifferenzen (Cohen's *d*)

Tabelle 3. Dosiseffekte auf absolute standardisierte Mittelwertdifferenzen (Cohen's *d*)

	Hypothese 3	Hypothese 4	Vollständiges Modell	Vollständiges Modell mit Interaktionen
Intercept	0.134*** [0.076; 0.192]; p < .001	0.167*** [0.108; 0.226]; p < .001	–	–
Häufigkeit (log)	0.008 [-0.003; 0.018]; p = .151	–	0.008. [-0.002; 0.018], p = .097	–
Abstand zwischen Wellen (log)	–	-0.010** [-0.017; -0.003]; p = .007	-0.010**[-0.017; -0.003]; p = .007	–
Einstellungen	–	–	0.192***[0.124;0.260]; p < .001	0.455***[0.219; 0.692]; p < .001
Verhalten	–	–	0.147*** [0.086; 0.208]; p < .001	0.144*** [0.083; 0.205]; p < .001
Interaktionen der Häufigkeit der Befragung mit der Art der Frage				
Einstellungen	–	–	–	-0.086. [-0.173; 0.002]; p = .055
Verhalten	–	–	–	0.009. [-0.001; 0.019]; p = .079
Interaktionen der Abstände zwischen den Befragungen mit der Art der Frage				
Einstellungen	–	–	–	-0.074* [-0.131; -0.017]; p = .011
Verhalten	–	–	–	-0.009* [-0.016; -0.002]; p = .012
Anteil erklärter Heterogenität	59.7 %	60.8 %	61.1 %	63.9 %

Anmerkungen: n = 154 Effektstärken, *k* = 19 Studienberichte

Richtung der Unterschiede zwischen den Gruppen von Interesse war, beziehen sich die Hypothesen 3 und 4 nur auf die Stärke der Effekte, also auf die absoluten Unterschiede zwischen den Gruppen. Je stärker sich die Gruppen unterscheiden, desto stärker der Konditionierungseffekt und umgekehrt. Für Hypothese 3 ist demnach ein positives Vorzeichen für den Effekt der Häufigkeit der Befragung zu erwarten. Der Effekt ist im univariaten Random-Effects-Modell nahe 0 und nicht signifikant. Auch im vollständigen Modell, welches auch das Timing der Befragung und die Art der Frage berücksichtigt, und beim Modell mit Interaktion bleibt der Effekt der Häufigkeit vorheriger Befragungen vernachlässigbar.

Bei Hypothese 4 wird eine Abschwächung des Konditionierungseffektes bei einem größeren Abstand der Erhebungswellen vorhergesagt. Die Ergebnisse der Meta-Regressionen unterstützen diese Hypothese. Bei größerem Abstand zwischen den Wellen ist die Differenz zwischen Kontroll- und Versuchsgruppe kleiner, was das negative Vorzeichen des Effekts zeigt. Auch im vollständigen Modell zeigt sich dieser Effekt, im Falle von Einstellungsfragen etwas stärker ausgeprägt als bei Fragen zum Verhalten.

Diskussion und Schlussfolgerungen

Insgesamt lässt sich aus der vorliegenden Meta-Analyse zu Konditionierungseffekten bei sensitiven Einstellungs- und Verhaltensfragen schließen, dass vorherige Befragungen durchweg nur sehr geringe Effekte auf das Antwortverhalten in Folgewellen haben. Die vorliegende Evidenz reicht nicht aus, um eindeutige Aussagen über die Richtung der Effekte in Bezug auf sozial erwünschtes Antwortverhalten treffen zu können. Auch die Annahme, dass die Häufigkeit der Befragung Konditionierungseffekte verstärkt, konnte nicht bestätigt werden. Größere Abstände zwischen den Befragungen scheinen aber tatsächlich die vermutete abschwächende Wirkung auf Konditionierungseffekte zu haben.

Generell muss gerade bei der Bewertung der Dosiseffekte die Begrenztheit der verfügbaren Datenbasis berücksichtigt werden. Es gibt kaum geplante Experimente zu Konditionierungseffekten in Panels (Struminskaya, 2016). Die meisten Studien nutzen Panel Refreshments, um neue mit bereits befragten Teilnehmern vergleichen zu können (Warren & Halpern-Manners, 2012). Dabei können Häufigkeit und Abstände zwischen den Wellen jedoch nicht gezielt variiert werden. So beruht zum Beispiel etwa die Hälfte der 154 Gruppenvergleiche der Meta-Analyse auf Versuchsgruppen, die vorher lediglich einmal befragt wurden. Häufige Intervalle zwischen den Panels sind eine Woche, ein Monat und ein Jahr. Dazwischen gibt es kaum Abstufungen.

Nach aktuellem Stand kann nicht von bedeutsamen Konditionierungseffekten in Panelerhebungen ausgegangen werden. Panelbasierte Erhebungen sind und bleiben demnach als bedeutende Datenquelle für die Psychologie interessant. Allerdings lassen die geringen in der vorliegenden Analyse gefundenen Effekte von Panelkonditionierung noch keine endgültigen Schlüsse hinsichtlich der Wirkungen auf die Qualität von Paneldaten zu. Es wurden nur die Auswirkungen bei sensitiven Items zu Einstellungen und Verhalten betrachtet. Die Datenbasis zu Einstellungsitems ist mit 38 Effektstärken aus sieben Studien schon relativ begrenzt. Andere Arten von Fragen, wie zum Beispiel demografische Angaben, die als vertraulich gesehen werden könnten oder nicht-sensitive Filterfragen, die aus strategischen Gründen falsch beantwortet werden könnten, um die Befragungsdauer zu verkürzen, sind hier nicht untersucht worden.

Panelkonditionierung ist vielfältig und kann unterschiedliche Ursachen und Effekte haben. Um weitreichende Aussagen über die Qualität und die Grenzen von Paneldaten treffen zu können, müssen die verschiedenen Mechanismen jeweils einzeln betrachtet werden. Dazu sind weitere Meta-Analysen notwendig, die unterschiedliche Arten von Fragen und auch verschiedene Konditionierungseffekte untersuchen sollten. Gerade für die gründliche Untersuchung der Dosiseffekte, die insbesondere für häufig befragte Populationen wie Teilnehmerpools von Online-Panels von großer Bedeutung sind, werden zusätzlich zu Meta-Analysen experimentelle Primärstudien benötigt. Diese erlauben eine gezielte Variation der Befragungsdosis, um die Lücken bisheriger Forschung schließen und Timing-Effekte genauer untersuchen zu können.

Literatur

Studien mit * wurden in der Meta-Analyse verwendet.

Alattar, L., Rogofsky, D. & Messel, M. (2018). An introduction to the understanding America study internet panel. *Social Security Bulletin,* 78 (2), 13 – 26.

Assink, M. & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology,* 12, 154 – 174.

* Axinn, W. G., Jennings, E. A. & Couper, M. P. (2015). Response of sensitive behaviors to frequent measurement. *Social Science Research, 49,* 1 – 15.

* Barber, J. S., Gatny, H. H., Kusunoki, Y. & Schulz, P. (2016). Effects of intensive longitudinal data collection on pregnancy and contraceptive use. *International Journal of Social Research Methodology, 19,* 205 – 222.

Baumeister, R., Vohs, K. D., DeWall, N. & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review,* 8 (1), 1 – 20.

Bergmann, M. & Barth, A. (2018). What was I thinking? A theoretical framework for analysing panel conditioning in attitudes and (response) behaviour. *International journal of social research methodology, 21,* 333 – 345.

Binswanger, J., Schunk, D. & Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly,* 77.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A. et al. (2018): Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review, 36* (1), 103 – 115.

* Bridge, G. G. (1977). Interviewing Changes Attitudes – Sometimes. *Public Opinion Quarterly*, 41 (1), 56 – 64.

Bruder, M., Göritz, A. S., Reips, U.-D. & Gebhard, R. K. (2014). Ein national gefördertes Onlinelabor als Infrastruktur für die psychologische Forschung. *Psychologische Rundschau, 65,* 75 – 85.

Cho, S. K., LoCascio, S. P., Lee, K.-O., Jang, D.-H. & Lee, J. M. (2017). Testing the Representativeness of a Multimode Survey in South Korea: Results from KAMOS. *Asian Journal for Public Opinion Research,* 4 (2), 73 – 87.

* Clinton, J.D. (2001). Panel Bias from Attrition and Conditioning. A Case Study of the Knowledge Networks Panel. *The American Association for Public Opinion Research (AAPOR),* 56th Annual Conference.

* Coombs, L. C. (1973). Problems of Contamination in Panel Surveys: A Brief Report on an Independent Sample, Taiwan, 1970. *Studies in Family Planning, 4,* 257.

* Das, M., Toepoel, V. & van Soest, A. (2011). Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys. *Sociological Methods & Research, 40* (1), 32 – 56.

Das, M., Kapteyn, A. & Bosnjak, M. (2018). Open probability-based panel infrastructures (pp. 199 – 209). In D.L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research.* London, UK: Palgrave Macmillan.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40,* 532 – 538.

Fisher, P. (2019). Does Repeated Measurement Improve Income Data Quality? *Oxford Bulletin of Economics and Statistics,* 0305–9049.

* Fitzsimons, G. J., John, D. S. N., Nunes, J. C. & Williams, P. (2007). License to Sin: The Liberating Role of Reporting Expectations. *Journal of Consumer Research, 34* (1), 22 – 31.

Fitzsimons, G. J. & Moore, S. G. (2008). Should we ask our children about sex, drugs and rock & roll? Potentially harmful effects of asking questions about risky behaviors. *Journal of Consumer Psychology, 18* (2), 82 – 95.

Fowler, F. J. (2006). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

* Halpern-Manners, A., Warren, J. R. & Torche, F. (2014). Panel conditioning in a longitudinal study of illicit behaviors. *Public Opinion Quarterly,* 78.

* Halpern-Manners, A., Warren, J. R. & Torche, F. (2017). Panel Conditioning in the General Social Survey. *Sociological Methods & Research, 46* (1), 103 – 124.

* Kraut, R. E. & McConahay, J. B. (1974). How Being Interviewed Affects Voting: An Experiment. *Public Opinion Quarterly, 37* (3), 398.

Kreuter, F., McCulloch, S., Presser, S. & Tourangeau, R. (2011). The Effects of Asking Filter Questions in Interleafed Versus Grouped Format. *Sociological Methods & Research, 40* (1), 88 – 104.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5* (3), 213-236.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: *The PRISMA Statement. PLoS Med 6* (7): e1000097.

*Murray, M., Swan, A. V., Kiryluk, S. & Clarke, G. C. (1988). The Hawthorne effect in the measurement of adolescent smoking. *Journal of epidemiology and community health, 42,* 304 – 306.

Nancarrow, C. & Cartwright, T. (2007). Online access panels and tracking research: the conditioning issue. *International Journal of Market Research, 49,* 573 – 594.

Percy, A., McAlister, S., Higgins, K., McCrystal, P. & Thornton, M. (2005). Response consistency in young adolescents' drug use self-reports: a recanting rate analysis. *Addiction, 100,* 189 – 196.

Phillips, D. L. & Clancy, K. J. (1972). Some Effects of „Social Desirability" in Survey Studies. *American Journal of Sociology, 77,* 921 – 940.

* Quick, A., Bohnke, J. R., Pickett, K. E. & Wright, J. (2017). Does involvement in a cohort study improve health and affect health inequalities? A natural experiment. *BMC Health Services Research,* 17:79.

Schonlau, M. & Toepoel, V. (2015). Straightlining in Web survey panels over time. *Survey Research Methods, 9* (2), 125 – 137.

Silber, H., Schröder, J., Struminskaya, B., Stocké, V. & Bosnjak, M. (2019). Does panel conditioning affect data quality in ego-centered social network questions? *Social Networks, 56,* 45 – 54.

Sobol, M. G. (1959). Panel Mortality and Panel Bias. *Journal of the American Statistical Association, 54* (285), 52 – 68.

* Song, Y. (2017). Rotation group bias in current smoking prevalence estimates using TUS-CPS. *Survey Research Methods, 11,* 383 – 404.

* Struminskaya, B. (2016). Respondent Conditioning in Online Panel Surveys: Results of Two Field Experiments. *Social Science Computer Review, 34* (1), 95 – 115.

Sturgis, P., Allum, N. & Brunton-Smith, I. (2009). Attitudes Over Time: The Psychology of Panel Conditioning. In: Lynn, P. (Ed.). *Methodology of longitudinal surveys* (pp. 113 – 126). Chichester, UK: Wiley.

* Toepoel, V., Das, M. & VanSoest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *The Public Opinion Quarterly, 72,* 985 – 1007.

* Torche, F., Valenzuela, E., Warren, J. R. & Halpern-Manners, A. (2012). Panel conditioning in a longitudinal study of adolescents' substance use: Evidence from an experiment. *Social Forces, 90,* 891 – 918.

Tourangeau, R., Rips, L. J. & Rasinski, K. (Eds.). (2000). *The psychology of survey response.* New York, NY: Cambridge University Press.

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45,* 576 – 594.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software, 36* (3), 1 – 48.

* Warren, J. R. & Halpern-Manners, A. (2012). Panel Conditioning in Longitudinal Social Science Surveys. *Sociological Methods and Research, 41,* 491 – 534.

* Waterton, J. & Lievesley, D. (1989). Evidence of conditioning effects in the British social attitudes panel survey. In: D. Kasprzyk (Ed.). *Panel surveys* (pp. 319 – 339). New York, NY: Wiley.

Weiß, B., Das, M., Kapteyn, A., Bosnjak, M. & Schaurer, I. (2020). Open probability-based panels. *Wiley StatsRef: Statistics Reference Online.*

* Williams, P., Block, L. & Fitzsimons, G. (2006). Simply asking questions about health behaviors increases both healthy and unhealthy behaviors. *Social Influence, 1* (2), 117 – 127.

Open Data

Zusatzmaterialien zu diesem Artikel werden unter PsychArchives.org http//dx.doi.org/10.23668/psycharchives.2684 zur Verfügung gestellt:

OD 1. Anhang 1: Codierungskategorien
OD 2. Anhang 2: Verwendete Datenbanken in CLICsearch
OD 3. Anhang 3: Verwendete Suchbegriffe
OD 4. Anhang 4: Begründung für Verwendung eines Drei-Ebenen-Modells
OD 5. Anhang 5: PRISMA Flow Diagramm (Studiensuche und -inklusion)
OD 6. Anhang 6: Funnelplot

Tanja Burgard, M.Sc., M.A.
Leibniz-Zentrum für Psychologische Information
und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
tb@leibniz-psychology.org

Appendix D: Supplementary Material of Study 2

Table D1: Coding Scheme (German)

Level 3: Bericht	
B1. Erstautor	
B2. Erscheinungsjahr	
B3. Art des Berichts	1 = Zeitschriftenartikel 2 = Buch oder Kapitel 3 = Dissertation 4 = Masterarbeit 5 = Regierungsbericht 6 = Konferenz-Paper 7 = Andere: ___________
B4. Peer-review?	0 = Nein, 1 = Ja
B5. Finanzierung / Sponsoring?	0 = Nein, 1 = Ja
Level 2: Stichprobe	
S1. Zielpopulation	1 = Allgemeine Bevölkerung 2 = Spezifische Bevölkerung
S2. Datensatz (z.B. SOEP)	
S3. Befragungsmodus	1 = Selbst-administriert, 2 = Interviewer
S4. Jahr des Vergleichs	
S5. Durchführungsort der Umfrage	Land
S6. Häufigkeit vorheriger Befragungen	
S7. Abstand zwischen Wellen (in Wochen)	
Level 1: Ergebnis Gruppenvergleich	
E1. Frage verlangt nach sozial erwünschter Antwort (Sensitiv 1)	0 = Nein, 1 = Ja
E2. Frage wird als privat oder zudringlich empfunden (Sensitiv 2)	0 = Nein, 1 = Ja
E3. Art der Frage	1 = Einstellungen, 2 = Verhalten
E4. Inhalt der Frage	Freitext
E5.Wert Kontrollgruppe	
E5b. Sd (Kontrollgruppe)	
E6. Wert Versuchsgruppe	
E6b. Sd (Versuchsgruppe)	
E7. Odds Ratio	
E7b. Varianz (LogOddsRatio)	
E8. P-Wert	
E9. T-Wert	
E10. Stichprobengröße Kontrollgruppe (n1)	
E11. Sichprobengröße Versuchsgruppe (n2)	
E12. Standardisierte Mittelwertsdifferenz (d)	
E12b. sd(d)	

Table D2: Search Terms Used in CLICSearch, December 2017

Search terms	**Number of hits**
“panel conditioning”	**280**
“survey conditioning”	**33**
“mere measurement effect”	**158**
“panel bias”	**520**
“time in sample”	**29**
“repeated survey participation”	**4**
“rotation group bias”	**189**
“re-interview effect”	**1**
“respondent conditioning”	**555**
“survey respondent conditioning”	**0**
“measurement reactivity”	**231**
“panel fatigue”	**130**
reinterview + “response bias”	**74**
re-interview + “response bias”	**51**
“testing effects” + “panel study”	**160**

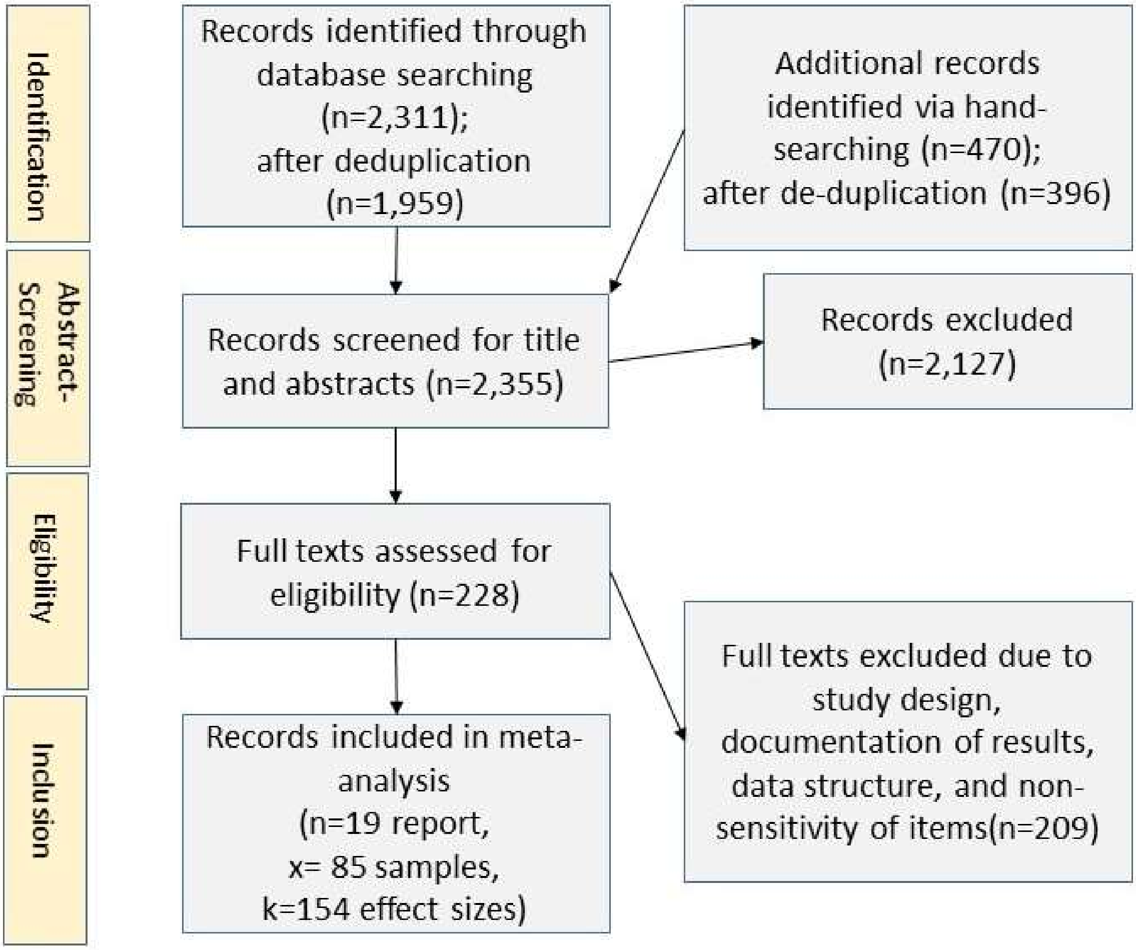

Identification
Abstract-Screening
Eligibility
Inclusion
Records identified through database searching (n=2,311); after deduplication (n=1,959)
Additional records identified via hand-searching (n=470); after de-duplication (n=396)
Records screened for title and abstracts (n=2,355)
Records excluded (n=2,127)
Full texts assessed for eligibility (n=228)
Full texts excluded due to study design, documentation of results, data structure, and non-sensitivity of items(n=209)
Records included in meta-analysis (n=19 report, x= 85 samples, k=154 effect sizes)

Figure D1: PRISMA Flow Chart for the Literature Search Process

Table D3: Overview of the Study Characteristics of the Selected Primary Studies

First Author	Year of Publication	Number of Samples	Number of comparisons	Type of the questions	Location of the study
Struminskaya	2016	1	1	Behavior	Germany
Halpern-Manners	2014	2	8	Behavior	USA
Das	2011	1	1	Behavior	Netherlands
Toepoel	2008	1	6	Behavior	Netherlands
Torche	2012	1	4	Behavior	Chile
Quick	2017	1	5	Behavior	England
Fitzsimons	2007	2	3	Behavior	USA
Axinn	2015	2	6	Behavior	USA
Clinton	2001	4	4	Behavior	USA
Murray	1988	2	6	Behavior	England
Song	2017	56	56	Behavior	USA
Williams	2006	1	2	Behavior	USA
Barber	2016	1	4	Attitudes	USA
Halpern-Manners	2017	1	2	Attitudes	USA
Bridge	1977	1	1	Attitudes	USA
Kraut	1973	2	2	Attitudes	USA
Warren	2012	4	33	Both	USA / Germany
Waterton	1989	1	5	Both	GB
Coombs	1973	1	5	Both	Taiwan
		85	154		

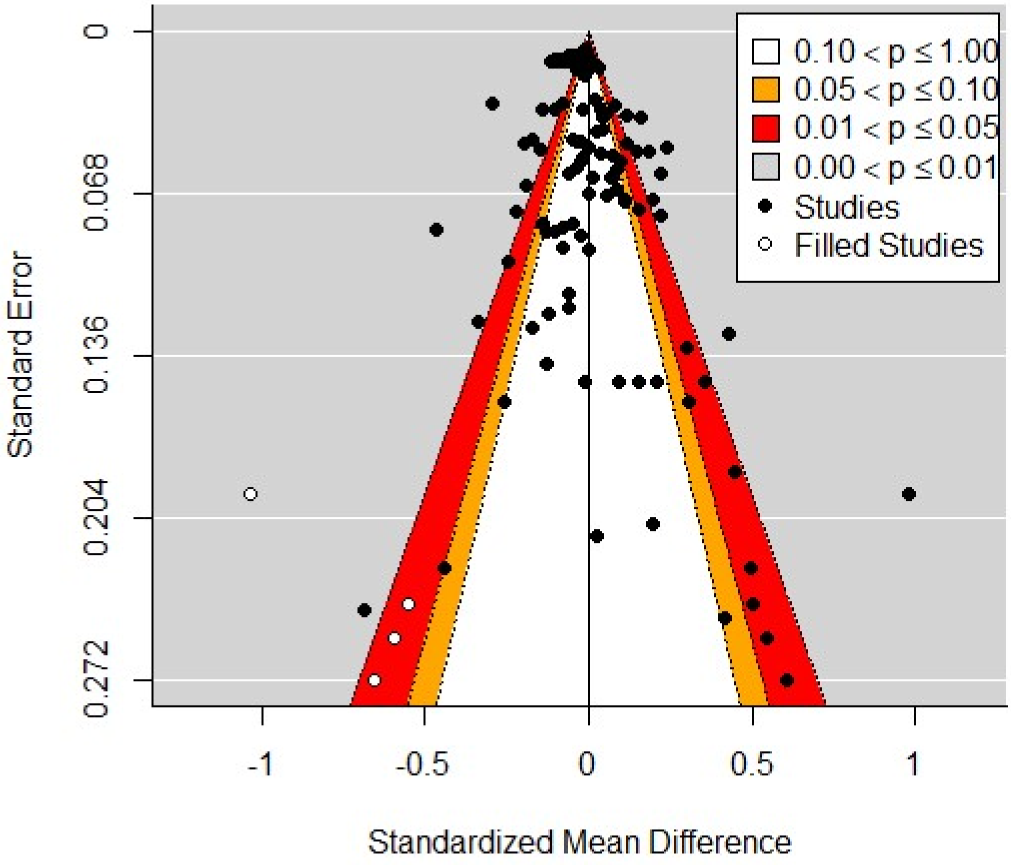


Figure D2: Contour-Enhanced Funnel Plot, Studies Added by the Trim and Fill Method

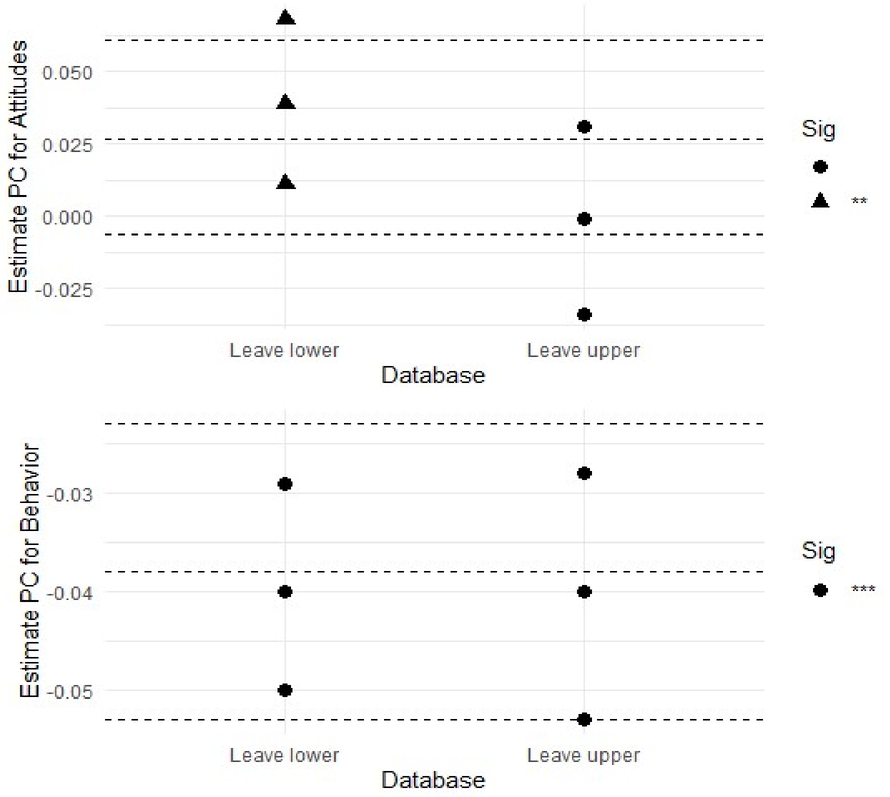


Figure D3: Sensitivity of Conditioning Effects Leaving Out Influential Studies

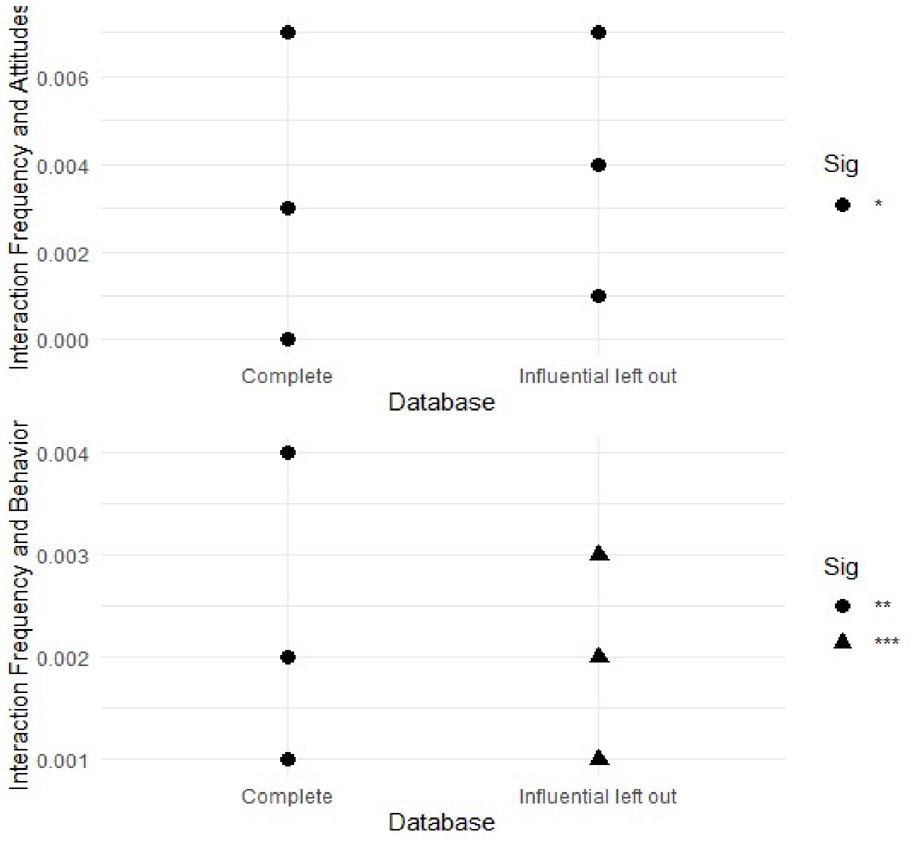


Figure D4: Sensitivity of the Effect of Frequency Leaving Out Influential Studies

Appendix E: Original Publication of Study 3

Review Article

Community-Augmented Meta-Analyses (CAMAs) in Psychology

Potentials and Current Systems

Tanja Burgard, Michael Bošnjak, and Robert Studtrucker

Leibniz Institute for Psychology (ZPID), Trier, Germany

Abstract: The limits of static snapshot meta-analyses and the relevance of reproducibility and data accessibility for cumulative meta-analytic research are outlined. A publication format to meet these requirements is presented: Community-augmented meta-analyses (CAMA). We give an overview of existing systems implementing this approach and compare these in terms of scope, technical implementation, data collection and augmentation, data curation, tools available for analysis, and methodological flexibility.

Keywords: replicability, meta-analysis, cumulative research, open data, repository

Typically, meta-analyses are published exclusively as static snapshots, depicting the evidence in a specific area up to a certain point in time. Moreover, in psychology, published meta-analyses rarely meet common reporting standards, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which was conceptualized more than a decade ago, or the more recently suggested MARS (Meta-analysis reporting standards) (Lakens et al., 2017). This practice leads to serious limitations with regard to the reusability of meta-analytic data and the currency of evidence.

The first problem often encountered by researchers is the lack of information to replicate the results of a meta-analysis. As a response to this problem, Lakens et al. (2016) argue for open meta-analytic data to make meta-analyses dynamic and reproducible. This is important for several reasons. First, having open access meta-analytic data would enable researchers the possibility of examining the sensitivity of the results to subjective decisions that were made in the original process of synthesizing the data, such as the underlying inclusion criteria, statistical models, or use of moderators. Second, an open access meta-analyses register would enable the application of new statistical procedures to existing data and allow testing the effects that these have on the meta-analytic results. Third, open access to existing meta-analyses provides other researchers with special research questions the opportunity to use subsets of the preexisting meta-analytic data (Bergmann et al., 2018).

The second problem of static snapshot meta-analyses is the fact that they are only valid for a specific cut-off date (Créquit et al., 2016). Without additional electronic material, a meta-analysis represents the cumulative evidence on a research question up to a certain point in time and may quickly become outdated as soon as new findings from primary studies are published or new methodological or statistical procedures are developed (Shojania et al., 2007). If the data are no longer accessible, the time-consuming process of conducting a meta-analysis must start from the beginning.

To facilitate and simplify cumulative research and to strengthen the evidence, for example, if practical challenges call for clear recommendations and decisions, we need to think about how to effectively publish our meta-analyses to make knowledge production more efficient. The key challenges for the publication of meta-analyses, therefore, are to make the preexisting research reproducible and to allow the updating of meta-analyses by reusing the information that has been collected up to the point of the most recent meta-analysis. In the following, these challenges will be discussed and the requirements for a publication format that enables reproducible and dynamic meta-analyses will be derived.

https://doi.org/10.1027/2151-2604/a000431

Challenges and Requirements for Meta-Analyses

Reproducibility and Replicability

Scientific findings can be validated on different levels (Stanley et al., 2018). Reproducibility means to produce exactly the same results with the same data and analyses. To validate findings at this level, accessibility of data and analysis code are sufficient. On the next level, we aim for replicability. When the same results and conclusions are obtained as those found in the original study using a new random sample and following the reported procedures, we can report a successful replication of results. Moreover, we distinguish between direct and conceptual replications (Zwaan et al., 2018). For direct replication, all critical facets of the study design in the original study have to be captured. A conceptual replication allows some differences in the study procedures. If findings are replicated independently of unmeasured factors in the original study, such as, for instance, sample characteristics or the country where the study took place, a finding can be considered as generalizable.

Replicability is already a problem at the level of individual studies. Reasons for failed replications can be found in every phase of the research cycle. If the findings are already known, researchers may propose hypotheses after the fact (a phenomenon known as HARKing – hypothesizing after the results are known; Kerr, 1998), leading to significant results deriving from their dataset. Another questionable research practice used to obtain significant results from the data is *p*-hacking (Simonsohn et al., 2014). Of the 64 studies they investigated, Banks et al. (2016) identified evidence of questionable research practices in 91%.

Apart from the questionable validity, individual studies often suffer from low statistical power (Fraley & Vazire, 2014), so that it is unlikely to find effects – especially small ones – even if they actually exist. Poor study quality and relevant differences in study design can also lead to different study results. Finally, published studies may not be representative of all studies, as significant results are published more often (Ioannidis, 2008). By chance, studies will provide meaningful results from time to time, even if the effects do not really exist. However, these results are often not replicable.

In meta-analyses, we expect a higher validity of the results due to the stronger evidence base and the heterogeneity in study designs and samples captured with a meta-analysis (Borenstein et al., 2009, p. 9). However, due to the frequency of questionable research practices in individual studies, the risk of bias of these should be accounted for in the first place (Hohn et al., 2019). Above this, when conducting a meta-analysis, there are a number of subjective decisions and sources of error threatening its validity (Cooper, 2017, p. 318).

Table 1 provides an overview of potential errors in each step of a meta-analysis. There might be several plausible alternatives for some decisions such as, for example, model specifications or treatment of missing data. However, the decisions must be justified and clearly reported to allow replication of the meta-analysis and also to investigate the impact of these decisions on the results by means of sensitivity analyses.

For meta-analyses, the four principles (open access, open methodology, open data, and source) of the open science movement advocated by Kraker et al. (2011) may be applied to overcome the challenge of making meta-analyses reproducible. Based on these principles, we derive the following requirements:

1. The transparent documentation of all steps and decisions along the meta-analytic process as presented in Table 1 enables the assessment of possible biases.
2. Common standards for interoperable and usable open data and scripts allow the verification of the results of a review. Subjective decisions may be modified, and new procedures may be applied with minimal effort to check the robustness of the results.

Updating and Cumulative Evidence

Static snapshot meta-analyses may quickly become outdated if they lack reusability to be expandable. Regular updates of meta-analyses are necessary. For example, Cochrane reviews should be updated every 2 years (Shojania et al., 2007) and Campbell reviews within 5 years (Lakens et al., 2016). Créquit et al. (2016) examined the proportion of available evidence on lung cancer not covered by systematic reviews between 2009 and 2015 with the finding that, in all cases, at least 40% of treatments were missing.

For systematic reviews, an update is defined as a new edition of a published review. It can include new data, new methods, or new analyses. An update is recommended if the topic is still relevant and new methods or new studies have emerged that could potentially change the findings of the original review (Garner et al., 2016).

Shojania et al. (2007) define signals of relevant evidence changes to warrant the update of reviews. These signals are changes in statistical significance, a relevant relative change in effect magnitude, new information on the clinical relevance of a review, or the emergence of new approaches not considered previously. For 100 reviews, the time between the publication and the occurrence of a signal for updating is measured and the median survival time of a meta-analysis in their analysis is 5.5 years. Within 2 years,

Table 1. Selected potential sources of error in the conduction of a meta-analytic process

Step of the meta-analysis according to Cooper (2017)	Sources of error	Threats to validity
Formulating the problem	Poorly defined constructs and relationships (e.g., Baer et al., 2019)	Questionable construct validity of measures
Searching the literature	Studies found in the literature search may not correctly represent the relevant population of studies (e.g., Pietschnig et al., 2010; Kepes & McDaniel, 2015)	Publication bias
Gathering information	Incorrect information retrieval resulting in misrepresentations in the data (e.g., London, 2016)	Unreliability in coding (inter- and intrapersonal), biased effect size sampling (favoring one direction of findings)
Quality appraisal	Evaluation approach not exhaustive concerning study design characteristics, nonexplicit weighting scheme (e.g., Hohn et al., 2019)	Biased exclusion of studies, biased weighting of studies
Synthesis methods and analysis	Nonweighting of effect sizes, unjustified model specifications, for example, in case of heterogeneous or dependent effect sizes (e.g., Voracek et al., 2019)	Biased estimates and inferences
Interpretation of results	Inaccurate treatment of missing data, no consideration of confounded moderator effects and heterogeneity in samples and study setting (e.g., Tran et al., 2014)	Biased estimates, overgeneralization of findings to contexts not represented in the meta-analysis

almost one-fourth of the reviews were already outdated (Shojania et al., 2007). As the number of publications is continuously growing (Bastian et al., 2010), we can expect the survival times of reviews to become even shorter.

Meta-analyses can also reveal research gaps by providing an overview of potential moderators or moderator combinations not yet sufficiently studied. In the case of Zhu et al. (2014), previous meta-analyses on the effect of thiazolidinedione treatment on the risk of fractures had mainly focused on postmenopausal women. New evidence provided the opportunity to study gender as a potential moderator of the effect, and it turned out that increased risk of fractures was only detected for women.

The ongoing accumulation of evidence informs researchers about the latest findings in a specific research area, for example, when the results are robust enough to no longer justify further research investment, at least without taking into account existing results and perhaps specific research gaps. A systematic review of cumulative meta-analyses (Clarke et al., 2014) reports many illustrating examples, speaking for the high relevance of cumulative research to enable more informed decisions and at the same time a more efficient distribution of research funds and efforts.

As requirements to overcome the challenge of updating meta-analyses, we can thus derive:

1. There is a need for infrastructures that are able to monitor the currentness and validity of meta-analytic evidence and to provide and apply decision rules for the necessity of updates.
2. Open access to data and metadata provides pre-existing research a usable and sustainable future. Extracted metadata and coding can be used to update a meta-analysis or even to conduct another meta-analysis on a similar subject with an overlap in the relevant literature.
3. Accumulating science and keeping evidence updated is a cooperative task and the participation in this task has to be supported and incentivized, for example, as proof of achievement instead of, or in addition to, the classical single publication.

Community-Augmented Meta-Analysis (CAMA) as a Publication Format

Openly available and regularly updated meta-analyses support the efficiency of science. Researchers can get a quick overview of a research field, can use the latest evidence for power analyses and study planning, and may make use of curated information and data to identify research gaps, as understudied moderator variables. As a solution for comprehensive, dynamic, and up-to-date evidence synthesis, Créquit et al. (2016) call for living systematic reviews, that is high-quality online summaries, that are continuously updated. Similarly, Haddaway (2018) proposes open synthesis.

Actually, a concept for a publication format for meta-analyses that meets these requirements already exists. There are slightly different forms that have been suggested for this meta-analytical concept including living (Elliott et al.,

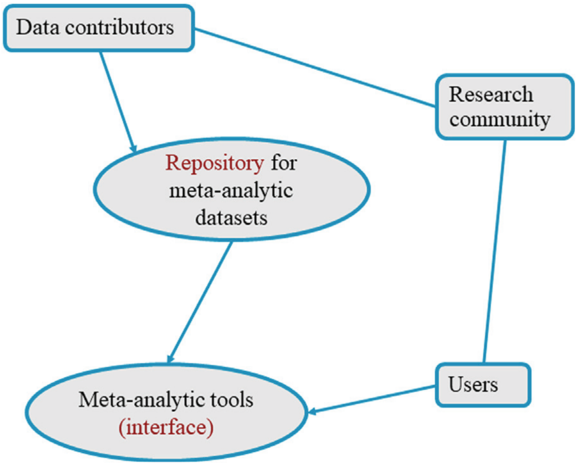

Data contributors
Research community
Repository for meta-analytic datasets
Meta-analytic tools (interface)
Users

Figure 1. The basic structure of a CAMA

2017), dynamic (Bergmann et al., 2018), or cloud-based meta-analysis (Bosco et al., 2015). Braver et al. (2014) describe an approach called continuously cumulating meta-analysis (CCMA) to incorporate and evaluate new replication attempts to existing meta-analyses. In our conception, we use the term community-augmented meta-analysis, CAMA for short (Tsuji, Bergmann, & Cristia, 2014). A CAMA is a combination of an open repository for meta-analytic data and an interface offering meta-analytic analysis tools.

The core of a CAMA, as shown in Figure 1, is the data repository, where meta-analytic data contributions from researchers in specific research areas are stored. It serves as a dynamic resource and can be used and augmented by the research community to keep the state of research updated and accumulate knowledge continuously. Tools to replicate and modify analyses with these data are accessible via an open web-based platform, usually encompassing a graphical user interface (GUI). For example, examining moderator effects beyond the analyses presented in the original meta-analysis may be conducted. The available evidence from the meta-analyses archived in a CAMA can also be used to improve study planning. Estimates of the expected size of an effect can serve as input for power analyses. The examination of possible relevant moderators can help to identify research gaps and guide the design of new studies (Tsuji et al., 2014).

Overview of Existing Systems Implementing CAMA in Psychology

There are already several systems and initiatives in psychology aiming at developing an infrastructure for the continuous curation and updating of meta-analytic evidence and, thereby, fulfilling the call to make meta-analyses reproducible and dynamic. In the following, five of these systems are reviewed and compared. These systems have been identified by conference meetings (metaBUS and MetaLab were presented at the Research Synthesis Conferences 2018 and 2019), and by successive searches for similar systems. However, the selection is not exhaustive. There are other CAMA systems outside psychology and the life sciences (e.g., MitiGate: https://mitigate.ibers.aber.ac.uk/), and systems aiming for open meta-analyses, but providing less information and guidance for users, thereby rendering them less adequate for comparison purposes (e.g., openMetaAnalysis: https://openmetaanalysis.ocpu.io/home/www/).

A project located in the domain of management and applied psychology is metaBUS. It is based on a hierarchical taxonomy of the field and provides a database consisting of correlations between clearly defined concepts within this taxonomy (Bosco et al, 2020). MetaBUS is a cloud-based platform and search engine providing access to more than 1.1 million curated findings from over 14,000 articles published in applied psychology journals since 1980 (https://www.metaBUS.org). It relies on the RStudio Shiny architecture for the GUI (Bosco et al., 2015) and the R statistics package metafor (Viechtbauer, 2010) for the meta-analytic calculations and visualizations.

To collaboratively collect and curate meta-analyses in the fields of early language acquisition and cognitive development, MetaLab offers a shiny webapp to reproduce meta-analyses and visualizations conducted with the statistical software R (Tsuji et al., 2017). Unlike the approach of metaBUS to retrieve single correlations, the data in MetaLab is organized in single meta-analyses, each focusing on the experimental evidence of one specific phenomenon (Bergmann et al., 2018). These meta-analyses are modified and improved collaboratively over time. At the moment (March 2020), MetaLab consists of 22 meta-analyses with information from 477 papers reporting on about 1,804 effect sizes (http://metalab.stanford.edu/).

Primarily located in the fields of cognitive and social psychology, the crowdsourced platform Curate Science (https://curatescience.org) allows the permanent curation of findings by the psychological research community. The design of the platform is guided by a unified curation framework enabling a systematic evaluation of empirical research along four dimensions: the transparency of methods and data, the reproducibility of the results by repeating the same procedures on the original data, the robustness of the results to different analytic decisions, and the replicability of effects in new samples under similar conditions (LeBel et al., 2018).

In the field of life sciences and health, Cochrane is piloting a project called Living Systematic Reviews (LSRs,

Table 2. Characteristics of selected CAMA systems in psychology and the life sciences

Domain	MetaBUS	MetaLab	Curate Science	Cochrane LSR	PsychOpen CAMA
Thematic scope	Management and applied psychology	Language acquisition and cognitive development	Psychology in general	Life sciences and health	Psychology in general
Data structure and effect sizes	Hierarchical taxonomy of concepts, over 1.1 million correlations	Single meta-analyses with over 1,800 effect sizes	Curated findings allowing evaluation (1,127 re-plications of 168 effects)	Published reports only	Single meta-analyses with effect sizes of any type
Analysis engine	metafor	metafor	metafor	None	metafor
Data collection and curation	Semi-automatic data collection and curation	Template sheets for standardization, responsible curators	Crowdsourcing and peer review to curate findings in evidence collections	Searches, screening, and updating by review author on a regular basis	Template sheets, synergy effects with related products, user accounts
User interface	R shiny webapp	R shiny webapp	Django/React (early beta, only open to small group of researchers)	None	PHP web application
Possible specifications	Multilevel, filter options, trim-and-fill parameters	Multilevel, moderator, and filter options	Filter replications by study characteristics	None	Multilevel, filter, moderator, effect size
Functionalities and tools	Flexible querying, funnel, violin plot, model output	Funnel, forest, violin, power plot, model output	Search and evaluate findings, forest plot	Follow updates on study website	Funnel, forest, power plot, model output, EGM
Export functionalities	Modification and download of query results	Dataset download as csv or excel file	None	None	Link to PsychNotebook for further analyses

Synnot et al, 2017), suggesting continuous updating for reviews with a high priority for health decision making, low certainty in the existing evidence, or a high likelihood of emerging evidence affecting the conclusions. An LSR is a review that is continually updated and incorporates new evidence immediately (Elliott et al., 2017). Cochrane LSRs and corresponding updates are published in the Cochrane database of systematic reviews (https://www.cochranelibrary.com).

At the Leibniz Institute for Psychology (ZPID), PsychOpen CAMA is currently under development with a first version becoming available in 2021. This service aims to serve the psychological research community as a whole by covering different psychological domains and meta-analyses on diverse effect sizes and study types. The approach for data storage and curation is similar to the one of MetaLab. Single meta-analyses can be published via the platform to become accessible to and expandable by the community. Instead of using an R shiny architecture for the GUI, PsychOpen CAMA relies on a PHP web application with an OpenCPU server for the R calculations. This improves the scalability of the web application according to the number of users, which is of special relevance for a service provided by a research infrastructure institute covering a broad scope of potential research domains, possibly reaching more users than rather narrowly specified applications targeted to a small research community.

Comparison of Data Collection, Augmentation, and Curation Approaches

The systems differ in terms of how data are collected and stored, augmented, and curated. As the data repository is the basis of a CAMA system, we will compare the systems previously introduced in terms of data administration. Table 2 sums up the central results of the comparison between the systems in terms of both data administration and, as discussed in the next section, data analyses.

The effect size of interest in metaBUS is correlations. On average, an empirical article contains 75 zero-order correlations, many of which would be overlooked during a traditional literature search. These correlations are collected by a semi-automated matrix extraction protocol. Trained coders supervise this process and additionally classify each variable according to the hierarchical taxonomy of variables and constructs in applied psychology. For each variable, further attributes, such as its reliability and response rate, are coded. The metaBUS database is constantly growing, but it relies exclusively on recruited, trained, and paid coders, as crowdsourcing efforts have not paid off yet due to the difficulty to motivate and train potential collaborators (Bosco et al., 2020).

As mentioned above, MetaLab is organized in single meta-analyses. The founders of the project have defined a general structure of potentially relevant meta-analyses.

Thus, the core parts of each meta-analysis are standardized. Templates and tutorials to explain how data have to be extracted and coded using this standardized structure are provided to guide external contributors when updating or adding meta-analyses to MetaLab (Tsuji et al., 2017). To guarantee the quality of data added to a meta-analysis, there is a responsible curator for each dataset. The standardized data in MetaLab allow the computation of common effect size measures as odds ratios or standardized mean differences. Meta-analyses in MetaLab are organized following a multilevel approach. Data usually originate from experimental studies sometimes reporting multiple effect sizes in one study and perhaps various studies in one paper. As effect sizes within a study and studies within a paper are usually more similar than effect sizes between studies or papers, the shared variance has to be considered to provide unbiased estimations (Bergmann et al., 2018).

The approach of Curate Science mainly relies on crowdsourcing. It provides a decentralized platform for the research community to curate and evaluates each other's findings. To facilitate this, Curate Science offers various features. A labeling system allows researchers to indicate compliance with reporting standards for their studies to curate transparency. The curation of reproducibility and robustness is supported by uploading corresponding reanalyses. Finally, replicability is curated by allowing the addition of replications to preexisting collections of published effects and by enabling researchers to create new evidence collections. To ensure the quality of this crowdsourced data collection, new replications added to evidence collections are reviewed by other users or editors (LeBel et al., 2018).

Research syntheses published as Cochrane reviews can be suggested for continuous updating due to special relevance. In this case, they follow clearly defined update scenarios. Searches and screening for LSRs are conducted on a regular basis (e.g., monthly). If no new data is found, only the search date is reported. If new evidence is found, the decision must be made about whether it should be integrated immediately or at a later date. In the case of immediate updating, data is extracted, analyses are rerun, and the review is republished (Elliott et al., 2017). Because this task is time-intensive for the individuals responsible for the LSR (typically the authors), there are aspirations to crowdsource and automatize microtasks for LSRs in the future. Searches may be continuously monitored by LSR specific filters at bibliographic databases, registries, and repositories. Thus, notifications may automatically be pushed in case of new potentially relevant studies. Their eligibility is assessed either by machine learning classifiers alone or complemented by crowdsourced efforts. Automation technologies for data extraction, synthesis, and reporting are still rudimentary (Thomas et al., 2017), and curation systems enabling the research community to maintain up-to-date evidence might be a better solution so far.

The meta-analytic data for PsychOpen CAMA is stored in PsychArchives, ZPID's archive for digital research objects in psychology. To update meta-analyses or to add completely new meta-analyses to PsychOpen CAMA, ZPID will ideally rely on synergy effects with its own related services and products. Research data from primary studies in PsychArchives can be used to update corresponding meta-analyses in CAMA. Alternatively, the results of studies or even complete meta-analyses preregistered at ZPID, as well as data collected in PsychLab will be used to extend the database for PsychOpen CAMA. As MetaLab, the template for data extraction assumes a multilevel structure and aims at standardizing data from different meta-analyses. In the future, user accounts should also serve data augmentation by giving users the possibility to edit data, for example, by adding new moderators or new studies. The suggestions made by users within their own accounts, however, have to be peer-reviewed before meta-analyses are updated accordingly.

Comparison of Available Analysis Tools

At the side of the user, the presented CAMA systems differ largely in the meta-analytic functionalities and the flexibility of the tools offered via the GUI. As the GUI is crucial for the accessibility of the meta-analyses to the interested users without expertise in meta-analysis, we will focus on the provided tools of the CAMA systems in the following.

The core functionality of metaBUS is the flexible querying via exact letter strings or taxonomic classifiers. There are two report modes. For the targeted search, two search terms are specified. Moreover, dependence in effect sizes may be considered, parameters for the trim-and-fill analysis can be specified, as well as the ranges of sample size, publication year, and the correlations (Bosco et al., 2015). An instant meta-analysis over all relevant bivariate relations and the corresponding metadata are returned. Users may refine their query for example via filtering by reliability or by checking the exact operationalizations of the concepts and if necessary, exclude individual entries. The newly developed exploratory search only requires one taxonomic node and instantly reports all meta-analyses with all other taxonomic nodes via an interactive plot (Bosco et al., 2020).

MetaLab offers meta-analytic modeling options, as the use of multilevel grouping, empirical Bayes estimations, and the use of selected moderator variables. Basic visualization tools, such as violin, forest, and funnel plots are available. Furthermore, prospective power analyses informed by the meta-analytic effect size of a given meta-analysis may be conducted to improve study planning. A simulation tool allows observing potential outcomes depending on key parameters of studies. Next to these basic tools available through a point-and-click interface, advanced users may also download the complete meta-analytic datasets and conduct their own analyses (Tsuji et al., 2017).

Curate Science essentially enables users to search for studies and evaluate findings based on characteristics related to transparency, reproducibility, robustness, and replicability. It provides an overview of the evidence on published and perhaps controversial effects. It also allows the meta-analysis of replications selected on the basis of study characteristics such as methodological similarities or preregistration status. Forest plots and meta-analytic estimations are then reported for the effect of interest (LeBel et al., 2018).

In their report on pilot LSR, Millard et al. (2018) sum up the processes and publication outputs from eight LSRs maintained during the pilot period. Depending on the amount of evidence published during this period, searches were conducted in time frames ranging from daily to once every three months. Updates were communicated on a regular basis to the readers via the study websites. For all but one study, new evidence was found during the pilot period. Only one LSR was completely republished. An interactive GUI, such as those used by metaBUS and MetaLab, is not yet available for Cochrane LSRs.

PsychOpen CAMA provides a user interface with basic meta-analysis tools, such as forest plots, funnel plots, and meta-analytic estimation. For these analyses, different effect sizes are available, dependency in effect sizes can be considered using a multilevel approach, and potentially relevant moderator variables can be chosen to be included in the model. Tools designed to inform about study planning decisions, such as evidence gap maps and power analyses, will also be included. Moreover, CAMA will be linked to PsychNotebook, a cloud-based electronic lab notebook for statistical analyses. Advanced users interested in applications that go beyond those directly available in CAMA may use the meta-analytic datasets within PsychNotebook.

To conclude, metaBUS, MetaLab, and PsychOpen CAMA address the open science principles to a great extent by providing data download and open analyses. The risk of bias of meta-analyses is also minimized by giving users the opportunity to filter included study results, including unpublished studies, add unconsidered moderator variables, and modify model specifications. Curate Science and LSR have no data export functionalities. Thus, relevant dimensions of the risk of bias, such as unjustified model specifications and unconsidered moderator effects, remain an issue, as the opportunities for open data and open analysis are not given.

Future Challenges for CAMAs

With a growing number of publications, efficient accumulation and synthesis of knowledge become the key to making scientific results usable and valid thus enabling more informed decisions. The survival time of the synthesized evidence in static meta-analyses, in many cases, is short. To keep this information up-to-date, the publication of meta-analyses in a format allowing the reuse of data already collected and an easy avenue to verify, update and modify meta-analyses is beneficial for the research community and the public.

A solution to enable dynamic and reusable meta-analyses is CAMA (community-augmented meta-analysis), a new, specialized publication format for meta-analyses. The core of such a system is the data repository, where effect sizes, completed meta-analyses, and metadata are stored and continuously curated.

The maintenance of such a repository, however, is challenging. Depending on the specific domain, a taxonomy for the concepts that are typically assessed, their designations, and the standards for the structure of the collected data has to be defined to allow the combination of research results assessing the same concepts or relations, regardless of how these were originally designated. This crucial, complex task must be undertaken for every meta-analysis to ensure that all research results are retrievable and comparable. Standards and taxonomies to ensure this is an essential aspect of a CAMA platform.

Furthermore, the continuous maintenance of a CAMA repository is both time- and labor-intensive. There are two ways to reduce the necessary workload, and these are already being applied to varying degrees in the systems presented here. The first one is crowdsourcing. MetaLab and Curate Science rely largely on this form of data accumulation. The difficulties encountered when relying on crowdsourcing, however, including how to motivate the crowd, how to educate contributors sufficiently to fulfill their tasks (e.g., by means of well-documented templates and tutorials), and how to ensure the quality of the contributions. Therefore, curation systems require quality checks, such as peer-reviewing of the added data or, as in the case of MetaLab, a curator who is responsible for checking all contributions before updating.

The second possibility to reduce the workload for the curation of the repository is the automatization of processes such as those involved in literature search (e.g., push notifications, database aggregators, automatic retrieval of full texts) and selection (e.g., machine-learning classifiers), or extraction of information from published reports (e.g., Robot Reviewer for information extraction and risk of bias assessment, Graph2Data for automatic data extraction from graphics) (Thomas et al., 2017). Currently, the software used to carry out these tasks is far from perfect and requires manual supervision. An R package facilitating all the single tasks mentioned at once, from abstract screening to data extraction and reporting of the literature selection process, is "metagear" (Lajeunesse, 2016). However, the further development of software is a research field in its own right. Algorithms need training data to learn how to decide on the inclusion of studies and extract information from reports. These training data have to be produced by manual effort.

Thus, neither crowdsourcing nor automatization completely solves the problem of the continuous curation of cumulative, meta-analytic evidence. All relevant processes in the selection, collection, and standardization of research results require human supervision. However, this is an effort providing benefits for the research community as a whole by improving the usability and currency of existing evidence. As continuously curated meta-analytic evidence also discloses and specifies research gaps, it enables efficient distribution of research funds for closing these gaps purposefully.

References

Baer, R., Gu, J., Cavanagh, K., & Strauss, C. (2019). Differential sensitivity of mindfulness questionnaires to change with treatment: A systematic review and meta-analysis. *Psychological Assessment, 31*(10), 1247–1263. https://doi.org/10.1037/pas0000744

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine, 7*(9), e1000326. https://doi.org/10.1371/journal.pmed.1000326

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bosco, F., Steel, P., Oswald, F., Uggerslev, K., & Field, J. (2015). Cloud-based meta-analysis to bridge science and practice: Welcome to metaBUS. *Personnel Assessment and Decisions, 1*(1), 3–17. https://doi.org/10.25035/pad.2015.002

Bosco, F. A., Field, J. G., Larsen, K. R., Chang, Y., & Uggerslev, K. L. (2020). Advancing meta-analysis with knowledge-management platforms: Using metaBUS in psychology. *Advances in Methods and Practices in Psychological Science, 3*(1), 124–137. https://doi.org/10.1177/2515245919882693

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Clarke, M., Brice, A., & Chalmers, I. (2014). Accumulating research: A systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. *PLoS One, 9*(7), e102670. https://doi.org/10.1371/journal.pone.0102670

Cooper, H. (2017). *Research synthesis and meta-analysis. A step-by-step approach*. Sage Publications.

Créquit, P., Trinquart, L., Yavchitz, A., & Ravaud, P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: The example of lung cancer. *BMC Medicine, 14*(8), 1–15. https://doi.org/10.1186/s12916-016-0555-0

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., & Thomas, J. (2017). Living systematic review: 1. Introduction – the why, what, when, and how. *Journal of Clinical Epidemiology, 91*, 23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010

Fraley, R. C., & Vazire, S. (2014). The *N*-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One, 9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Schünemann, H. J., Akl, E. A., Beyene, J., Chang, S., Churchill, R., Dearness, K., Guyatt, G., Lefebvre, C., Liles, B., Marshall, R., Martínez García, L., Mavergames, C., Nasser, M., Qaseem, A., Sampson, M., . . . Schünemann, H. J. (2016). When and how to update systematic reviews: Consensus and checklist. *British Medical Journal (Online), 354*, 1–10. https://doi.org/10.1136/bmj.i3507

Haddaway, N. R. (2018). Open synthesis: On the need for evidence synthesis to embrace open science. *Environmental Evidence, 7*(1), 4–8. https://doi.org/10.1186/s13750-018-0140-4

Hohn, R. E., Slaney, K. L., & Tafreshi, D. (2019). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology, 9*, Article 2667, 1–15. https://doi.org/10.3389/fpsyg.2018.02667

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kepes, S., & McDaniel, M. A. (2015). The validity of conscientiousness is overestimated in the prediction of job performance. *PLoS One, 10*(10), Article e0141468. https://doi.org/10.1371/journal.pone.0141468

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kraker, P., Leony, D., Reinhardt, W., & Beham, G. (2011). The case for an open science in technology enhanced learning. *International Journal of Technology Enhanced Learning, 3*(6), 643–654. https://doi.org/10.1504/IJTEL.2011.045454

Lajeunesse, M. J. (2016). Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution, 7*(3), 323–330. https://doi.org/10.1111/2041-210X.12472

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*(1), 1–10. https://doi.org/10.1186/s40359-016-0126-3

Lakens, D., van Assen, M., Anvari, F., Corker, K., Grange, J., Gerger, H., Hasselman, F., Koyama, J., Locher, C., Miller, I., Page-Gould, E., Schönbrodt, F. D., Sharples, A., Spellman, B. A., & Zhou, S. (2017 31). *Examining the reproducibility of meta-analysis in psychology: A preliminary report*. https://doi.org/10.31222/osf.io/xfbjf

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpeamel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389–402. https://doi.org/10.1177/251524591878748

London, J. E. (2016). *The effect of time period, field, and coding context on rigor, interrater agreement, and interrater reliability in meta-analysis*. Dissertation, North Carolina State University.

Millard, T., Synnot, A., Elliott, J., & Turner, T. (2018). *Results from the evaluation of the pilot living systematic reviews*. https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201905 LSR_pilot_evaluation_report.pdf

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence, 38*(3), 314–323. https://doi.org/10.1016/j.intell.2010.03.001

Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine, 147*(4), 224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Synnot, A., Turner, T., & Elliott, J. (2017). *Cochrane Living Systematic Reviews*. https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/LSRInterimguidance_v0.3_20170703.pdf

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., & Elliott, J. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology, 91*, 31–37. https://doi.org/10.1016/j.jclinepi.2017.08.011

Tran, U. S., Hofer, A. A., & Voracek, M. (2014). Sex differences in general knowledge: Meta-analysis and new data on the contribution of school-related moderators among high-school students. *PLoS One, 9*(10), Article e110391. https://doi.org/10.1371/journal.pone.0110391

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Tsuji, S., Bergmann, C., Lewis, M., Braginsky, M., Piccinini, P., Frank, M. C., & Cristia, A. (2017). MetaLab: A repository for meta-analyses on language development, and more. In International Speech Communication Association (Ed.), *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH, 2017)*. https://www.isca-speech.org/archive/Interspeech_2017/pdfs/2053.PDF

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift für Psychologie, 227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Zhu, Z. N., Jiang, Y. F., & Ding, T. (2014). Risk of fracture with thiazolidinediones: An updated meta-analysis of randomized clinical trials. *Bone, 68*, 115–123. https://doi.org/10.1016/j.bone.2014.08.010

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120. https://doi.org/10.1017/S0140525X17001972

History
Received April 24, 2020
Revision received June 18, 2020
Accepted August 17, 2020
Published online February 4, 2021

Conflict of Interest
We have no known conflict of interest to disclose.

ORCID
Tanja Burgard
https://orcid.org/0000-0001-9194-4821

Michael Bošnjak
https://orcid.org/0000-0002-1431-8461

Tanja Burgard
Leibniz Institute for Psychology (ZPID)
Universitätsring 15
54296 Trier
Germany
burgard@leibniz-psychology.org

Appendix F: Submitted manuscript of Study 4 (under review)

PsychOpen CAMA: Publication of Community-Augmented Meta-Analyses in Psychology

Evidence in psychology is crucial for decision-making in many fields for example based on the effectiveness of psychotherapies[1], the effects of interventions on health behavior[2], or the influence of workplace conditions on the mental health of employees.[3] To enable optimal decision-making in practice, research syntheses have to be accessible for decision-makers and the general public, and should be updated rapidly, if new study results or analytic methods become available. In order to fulfill these requirements, open meta-analytic data[4] and open synthesis[5] are called for. We present the architecture, user interface, and functionalities of a platform for community-augmented meta-analyses (CAMA), serving the psychological research community and neighboring fields.

1. Requirements for Open Synthesis for Evidence-Based Decision-Making

To serve the purpose of providing information for decision-making in practical contexts, the comprehensibility of results is of high relevance. A graphical user interface (GUI) providing visualizations including interpretation aids can enable users without proficient knowledge of meta-analytic methods to use meta-analytic data to get an overview on the evidence on a research question.[6] Plain Language Summaries giving a summary of the existing evidence, in the tradition of Cochrane reviews[7], can complement the GUI to make scientific knowledge accessible for decision-makers and the public.

For researchers with further interest in a published meta-analysis, data access and a thorough documentation of the underlying methodology is crucial to be able to replicate or re-use the data. Published results can thus be replicated within the GUI. Above this, reuse of the data for subgroup analyses, or the modification of methodological decisions, as estimation method and modeling choices provides these users the opportunity to use existing data resources for novel purposes.[5]

Due to a large and growing number of published findings[8], the results of meta-analyses outdate fast[9] and updating existing meta-analyses is often time-consuming, as relevant resources cannot be accessed for an efficient use.[10] To prevent waste in research, infrastructures are needed to facilitate the accumulation of evidence by fostering FAIR data sharing for optimal re-usability and exploitation of data.[11]

2. Community-Augmented Meta-Analysis

Actually, a concept for a platform that meets these requirements already exists. There are slightly differing forms that have been suggested including living[12], dynamic[13], or cloud-based meta-analysis.[6] Braver et al.[14] describe an approach called continuously cumulating meta-analysis (CCMA) to incorporate and evaluate new replication attempts to existing meta-analyses. All of these approaches have in common, that they aim at accumulating scientific results to keep evidence up-to-date.

We use the term community-augmented meta-analysis, CAMA for short[15], to describe a platform providing an open repository for meta-analytic data and a GUI for meta-analytic tools. A CAMA is supposed to facilitate and foster the accumulation of evidence by providing a user-friendly infrastructure to the research community. Existing systems in psychology have been reviewed and presented in a previous article[16].

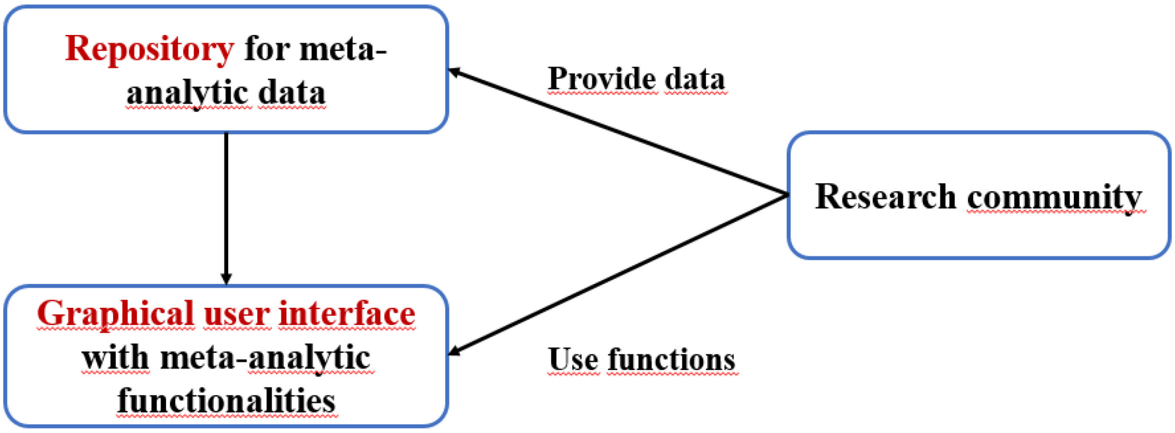

Repository for meta-analytic data
Provide data
Research community
Graphical user interface with meta-analytic functionalities
Use functions

Figure 1: Basic Form of a CAMA

Figure 1 illustrates the essence of a CAMA. The research community at the one hand feeds the CAMA with new data and on the other hand benefits from it by having easy access to data and meta-analyses. The role of the infrastructure provider is to set standards for the data submitted to the repository and to store the data according to these standards. The functionalities on the GUI are interoperable with these data and thus, as meta-analytic datasets are augmented or new meta-analyses are implemented, analysis outputs can be requested by users and are automatically available on the GUI. The platform thus serves as a dynamic resource enabling the research community to keep the state of research updated and accumulate knowledge continuously by providing a common language for the data.

To sum up, a CAMA adheres to the requirements of FAIR data by making research results findable, complete datasets accessible, ensuring interoperability of data and analysis scripts, and thus, making data reusable.[17]

3. From meta-analytic data to dynamic synthesis in PsychOpen CAMA

At the Leibniz Institute for Psychology (ZPID), PsychOpen CAMA is currently under development with a first version becoming available in 2021. This service aims to serve the psychological research community as a whole by covering a broad scope of potential research domains. Meta-analyses can be published via the platform to become accessible to and expandable by the community. As Figure 2 shows, PsychOpen CAMA relies on a PHP web application with an OpenCPU server for the R calculations. This improves the scalability of the web application according to the number of users compared to commonly used R shiny architectures.

Original meta-analytic data from users is standardized according to a spreadsheet defining the structure and naming of CAMA data. Standardized data becomes part of a self-maintained R package, that also contains all functions needed for analysis requests offered on the GUI. The user can choose a dataset and request meta-analytic outputs for this data on the

GUI. The request is then sent to the server, where the computations are executed. The resulting outputs of the analyses are given back to the user via the GUI.

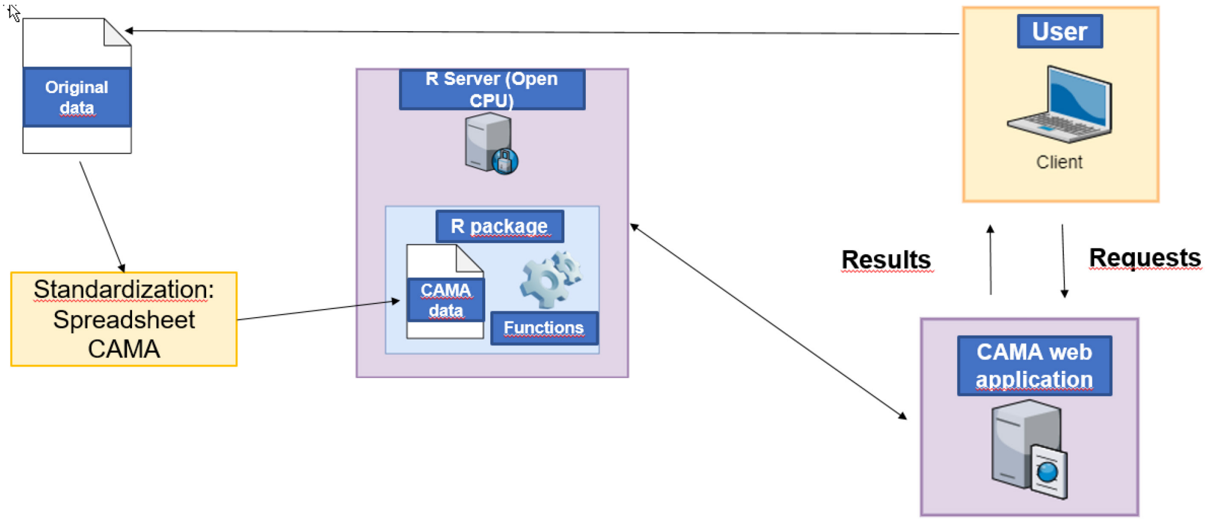


Figure 2: Architecture of PsychOpen CAMA

3.1. Interoperability: Data standardization and metadata

Interoperability enables operational processes and information exchange between different systems. Optimally, standardized identifiers and metadata for all data and digital objects allow for an automated access and use of data by humans and machines.[18] To achieve interoperability of different datasets with the analysis functions used in PsychOpen CAMA, a template for meta-analytic data and machine-readable metadata are used.

As Figure 3 illustrates, the template of PsychOpen CAMA the collection of data on different levels. There may be dependencies in the outcome measures of meta-analyses. This might occur, if the effect of an intervention is measured using multiple outcomes, for example competences in different domains. If multiple outcomes are measured using the same study sample, results within a sample might be more similar than between samples. Not accounting for this potential covariance of the outcome measures from the same sample can bias statistical inferences.[19] Data do not have to be nested necessarily. In some meta-analyses, there might only be one outcome measure per sample and report. However, the structure and variable naming enable to distinguish the information levels of the variables and – in case of dependencies – automatically trigger the use of a multilevel model in the analysis scripts.

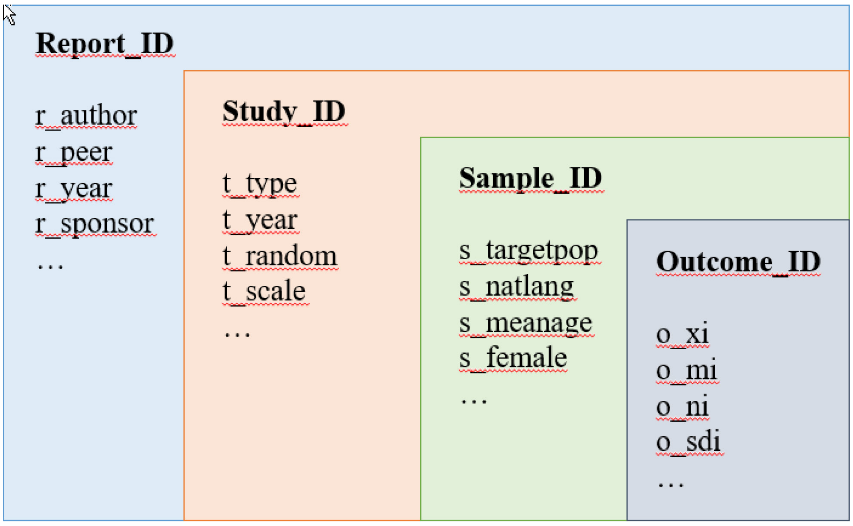


Figure 3: Schematic illustration of the data template for PsychOpen CAMA

As a first orientation for a template for basic meta-analytic information on report, study, and sample, the spreadsheet of metalab[15] served as a starting point. As the meta-analyses for which this spreadsheet serves as a template are all located in the domain of language acquisition and cognitive development, adaptations for other fields of research are necessary. As it is not possible to include each moderator variable that might be relevant in any field, the template is kept rather generic and leaves space for specific adaptations in the form of adding relevant moderators that are not included in the basic template.

The variable names of the outcome data follow the naming of potential measures serving as inputs to compute effect sizes with the escalc() function in metafor.[20] As PsychOpen CAMA operates mainly with this package and effect sizes are computed using the escalc() function, it is convenient to use the naming and description of the various outcome information potentially given in a report coded for a meta-analysis.

The naming of metafor is also used in the metadata, where the kind of effect size measure has to be given for each meta-analysis. The options for the ‘measure’ argument are used in various functions in metafor and therefore, following the standards of metafor in this case is also reasonable. Next to the kind of effect size measure, the metadata of each dataset contain the inclusion criteria, relevant moderator variables, research question, nesting of the data, and bibliographic information. The metadata thus serves the purpose of documentation of the methodological conduction of each meta-analysis. Moreover, the metadata are crucial for the automated analyses of the various datasets.

If a user for example selects a certain dataset, the GUI instantly reacts and offers the user the moderators that are available for this dataset. If the user asks for a meta-regression in the next step, the self-maintained R package takes the information on the type of effect size measure and potential dependencies in the data from the metadata to choose the right function and arguments.

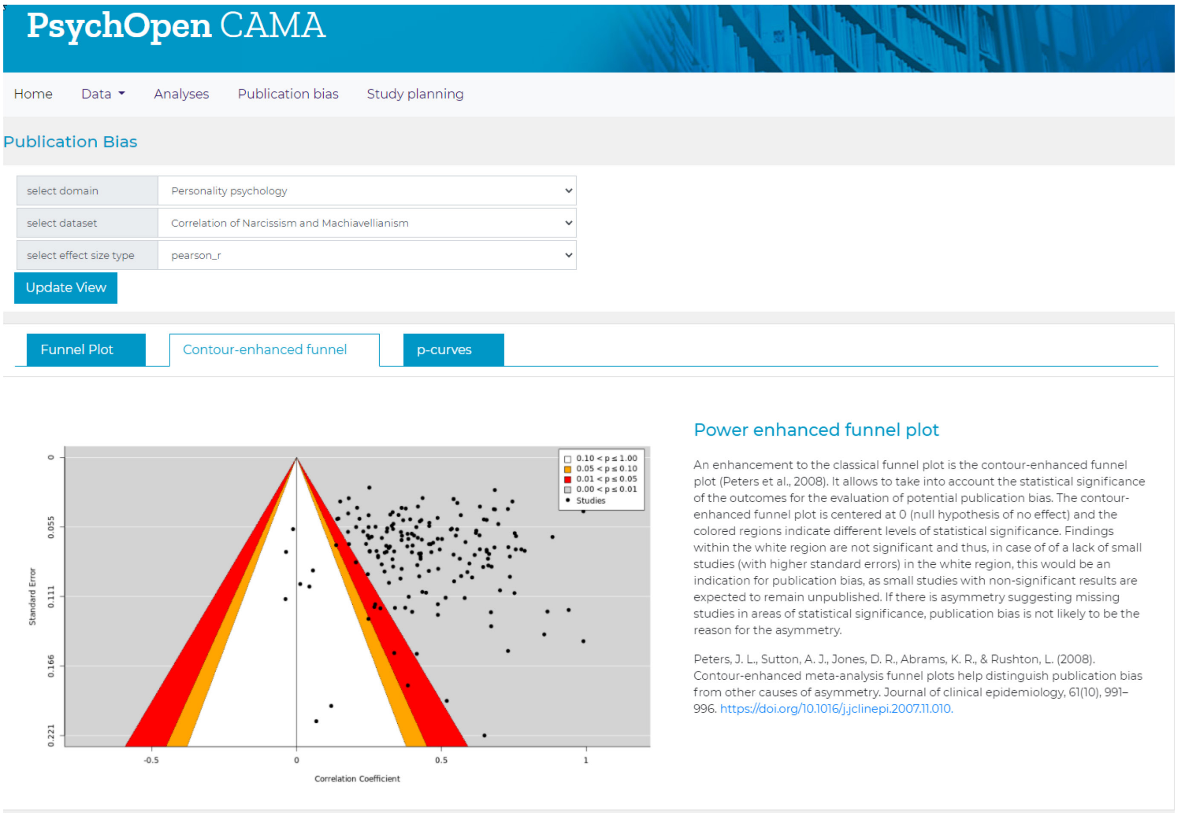
3.2. Graphical user interface: Use cases and functionalities

In the first version released, PsychOpen CAMA provides a GUI, offering the user easy access to the results of 14 meta-analytic datasets (February 2021). An intuitive and responsive point-and-click tool makes it easy to explore the data. Interpretation aids to each output make the results comprehensible, even for scientific laypersons. Moreover, these aids are also suited to serve educational purposes.

The menu item "Data" contains a thorough documentation, including bibliographic and methodological information, as well as links to primary studies included in the meta-analyses, and a data table for each dataset. Moreover, a data exploration tool provides a quick overview on the univariate distributions of effect sizes and potentially relevant moderator variables and the corresponding bivariate and trivariate distributions between these variables.

Basic meta-analytic outputs, such as forest plots and meta-analytic estimation, can be found under the item "Analyses". A dataset and an available effect size type, as well as moderators for inclusion in the meta-regression, can be chosen. If the data are nested, a multilevel model is automatically used to consider dependency in effect sizes. A detailed description of the statistical coefficients is given next to the output to give the user the opportunity to understand the statistics behind and to draw conclusions from the results.

In Figure 4, one of the outputs to assess potential publication bias is displayed, the contour-enhanced funnel plot. A classical funnel plot, the results of an Egger's test[21], as well as p-curve analysis[22] are also available in this context to give the user the opportunity to assess the evidential value and potential bias of a meta-analysis using different tools.



Publication Bias

select domain	Personality psychology
select dataset	Correlation of Narcissism and Machiavellianism
select effect size type	pearson_r

Power enhanced funnel plot

An enhancement to the classical funnel plot is the contour-enhanced funnel plot (Peters et al., 2008). It allows to take into account the statistical significance of the outcomes for the evaluation of potential publication bias. The contour-enhanced funnel plot is centered at 0 (null hypothesis of no effect) and the colored regions indicate different levels of statistical significance. Findings within the white region are not significant and thus, in case of of a lack of small studies (with higher standard errors) in the white region, this would be an indication for publication bias, as small studies with non-significant results are expected to remain unpublished. If there is asymmetry suggesting missing studies in areas of statistical significance, publication bias is not likely to be the reason for the asymmetry.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. Journal of clinical epidemiology, 61(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010.

Figure 4: Screenshot of the Test Version for PsychOpen CAMA

Finally, a study planning tool allows to conduct a prospective power analysis for a potential new study on one of the research questions of the included meta-analyses. Therefore, the meta-analytic estimate of the corresponding meta-analysis is assumed as the true underlying effect size. The sample size and desired significance level are chosen by the user. The tool calculates the expected power of the prospective study, as well as a necessary sample size to achieve a power of 80 %. This provides a quick indication of how large a study in a certain domain needs to be to achieve sufficient statistical power and may thus guide researchers in planning new studies.

4. Future Challenges for PsychOpen CAMA

As a central research infrastructure institute for psychology, ZPID has resources and tools to provide users assistance with data submission and updating of data. Furthermore, the benefit of PsychOpen CAMA can be increased by giving users far-reaching and flexible analysis options. By using already existing resources within ZPID, we can reduce manual effort and simultaneously increase potential applications for the user. In the following, we will describe the planned synergies of PsychOpen CAMA with other ZPID services, concerning data acquisition, data analysis, and the methodological and topical scope of the service.

4.1. Data acquisition and updating: Crowdsourcing and automation

The continuous maintenance of a CAMA repository is both time- and labor-intensive. The workload can be reduced via crowdsourcing[23], if the research community is willing and able to provide relevant data, at best in the desired format. To support users in the submission of data, we plan to use the submission assistant of our archive for digital research objects in psychology, PsychArchives.[24] To ensure interoperability of the data with PsychOpen CAMA for the implementation on the platform, manual effort for validity checks will still be needed. We will automatize repetitive processes as far as possible, for example by using notifications in case of new data entries, and scripts for validity checks. But at least for the monitoring of these processes, additional plausibility checks, and necessary corrections in case of erroneous entries, manual effort cannot fully be replaced.

To strategically acquire new data for PsychOpen CAMA, there are more resources to be used. Research data from primary studies shared in PsychArchives can be used to update corresponding meta-analyses in CAMA. Alternatively, the results of studies or even complete meta-analyses preregistered at ZPID (https://prereg-psych.org/), as well as data collected in ZPID's online or offline laboratory will be used to extend the database for PsychOpen CAMA. For meta-analyses published in one of the journals of PsychOpen (https://www.psychopen.eu/), authors could be asked to share the meta-analytic data of the meta-analysis. The long-term goal of these strategies is to automate these linkages as far as possible to accumulate evidence in PsychOpen CAMA and keep pace with the mass of scientific results produced and published in various domains in psychology.

4.2. Analytical and methodological scope

To make data use for further analyses easier, PsychOpen CAMA will be connected to PsychNotebook, a cloud-based jupyter-lab notebook for statistical analyses in psychology, that will also be released soon. Advanced users interested in applications that go beyond those directly available on the GUI of PsychOpen CAMA may use the meta-analytic datasets within the free R environment in PsychNotebook. Furthermore, code snippets for advanced meta-analytic functionalities will be provided to facilitate the analyses in PsychNotebook, and to serve educational purposes. Users can create their own projects within PsychNotebook, where they can collect and save their own ideas, analysis scripts and outputs and share these with others.

There are various approaches in meta-analyses. For comparing multiple treatments in clinical psychology[25] or the effects of interventions on behavior[26], network meta-analysis is of particular importance.[27] For the estimation of overall estimates and interactions, the combination of available individual participant data (IPD) with aggregate data (AD) is superior[28], suggesting to use available raw data from studies included in meta-analyses whenever available in data archives. In behavioral and social sciences, relationships are often represented in the form of complex models including relations between several variables simultaneously. The results of structural equation models used to depict those relationships can be meta-analyzed with the help of the MASEM approach.[29]

All of the techniques mentioned require data standards and analysis outputs differing from those already available in PsychOpen CAMA. Data templates and the implementation of special analysis functionalities for these kinds of meta-analyses are therefore needed to broaden the scope of the platform.

5. Discussion

With a growing number of publications, the survival time of synthesized evidence is short. Efficient accumulation and synthesis of knowledge becomes the key to making scientific results usable and valid. To keep meta-analyses up-to-date, they have to be published in a format allowing the reuse of data and an easy avenue to verify, modify, and update meta-analyses.

PsychOpen CAMA is presented as a solution to enable dynamic and reusable meta-analyses. It provides a repository for interoperable meta-analytic data and a GUI for easy access to the results from the available analyses to the research community and the public. However, challenges regarding automation of workflows, flexibility of analysis options, and the scope of the platform remain. There is great potential to address these challenges by using further resources and tools at ZPID.

Highlights

- **What is already known?**
  Due to a growing number of research findings, accessibility of data and technical solutions to enable timely and easy updates of meta-analyses are needed to provide the best available evidence for practical decision-making. A concept for dynamic and reusable meta-analyses is CAMA (Community-Augmented Meta-Analysis).

- **What is new?**
  PsychOpen CAMA is a web application to serve the psychological research community as a whole by providing a repository with standardized meta-analytic data and a GUI to use data from this repository for meta-analytic calculations and visualizations.

- **Potential impact?**
  In PsychOpen CAMA, meta-analytic data can be reused by the research community and data curators to verify, modify, and update meta-analyses in psychology and neighboring fields.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study. All data that will be published in PsychOpen CAMA will be made available in PsychArchives before the release of the platform.

References

1. Cook, S.C., Schwartz, A.C., & Kaslow, N.J. (2017): Evidence-based psychotherapy: Advantages and challenges. *Neurotherapeutics*, 14, 537-545. https://doi.org/10.1007/s13311-017-0549-4

2. Ygram Peters, G.-J., de Bruin, M., & Crutzen, R. (2015). Everything should be as simple as possible, but no simpler: towards a protocol for accumulating evidence regarding the active content of health behaviour change interventions. *Health Psychology Review*, 9(1), 1-14. https://doi.org/10.1080/17437199.2013.848409

3. Harvey, S. B., Modini, M., Joyce, S., Milligan-Saville, J. S., Tan, L., Mykletun, A., Bryant, R. A., Christensen, H., & Mitchell, P. B. (2017). Can work make you mentally ill? A systematic meta-review of work-related risk factors for common mental health problems. *Occupational and environmental medicine*, 74(4), 301–310. https://doi.org/10.1136/oemed-2016-104015

4. Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4(1),* 1–10. https://doi.org/10.1186/s40359-016-0126-3

5. Haddaway, N. R. (2018). Open synthesis: On the need for evidence synthesis to embrace open science. *Environmental Evidence*, *7*(1), 4–8. https://doi.org/10.1186/s13750-018-0140-4

6. Bosco, F., Steel, P., Oswald, F., Uggerslev, K., & Field, J. (2015). Cloud-based Meta-analysis to Bridge Science and Practice: Welcome to metaBUS. *Personnel Assessment and Decisions*, 1(1). https://doi.org/10.25035/pad.2015.002

7. Langendam, M.W., Akl, E.A., Dahm, P., Glasziou, P., Guyatt, G., & Schünemann, H.J. (2013). Assessing and presenting summaries of evidence in Cochrane Reviews. Systematic Reviews, 2, 81. https://doi.org/10.1186/2046-4053-2-81

8. Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine*, *7*(9), e1000326. https://doi.org/10.1371/journal.pmed.1000326

9. Shojania, K. G., Sampson, M., Ansari, M. T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine*, *147*(4), 224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179

10. Créquit, P., Trinquart, L., Yavchitz, A., & Ravaud, P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: The example of lung cancer. BMC Medicine, 14(8). https://doi.org/10.1186/s12916-016-0555-0

11. Schultes E., & Wittenburg P. (2019). FAIR Principles and digital objects: Accelerating convergence on a data infrastructure. In: Manolopoulos Y., Stupnikov S. (eds). *Data analytics and management in data intensive domains.* DAMDID/RCDL 2018. Communications in Computer and Information Science, 1003. Springer, Cham. https://doi.org/10.1007/978-3-030-23584-0_1

12. Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. Journal of Clinical Epidemiology, 91, 23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010

13. Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

14. Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

15. Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*(6), 661–665. https://doi.org/10.1177/1745691614552498

16. Burgard, T., Bosnjak, M., & Studtrucker, R. (2021). Community-Augmented Meta-Analyses (CAMAs) in Psychology. Potentials and Current Systems. *Zeitschrift für Psychologie*, *229*, 15-23. https://doi.org/10.1027/2151-2604/a000431

17. Wu, M., Psomopoulos, F., Khalsa, S.J., & de Waard, A. (2019). Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, 18(3), 1-13. https://doi.org/10.5334/dsj-2019-003

18. Sansone, S.-A. & Rocca-Serra, P. (2016). Interoperability Standards - Digital Objects in Their Own Right. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4055496

19. Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6.

20. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package, *Journal of Statistical Software, 36* (3), 1-48. http://dx.doi.org/10.18637/jss.v036.i03

21. Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical research ed.)*, 315(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

22. Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspectives on Psychological Science, 9(6), 666–681. https://doi.org/10.1177/1745691614553988

23. McCarthy, R.J., Chartier, C.R. (2017): Collections^2: Using "Crowdsourcing" within Psychological Research. *Collabra: Psychology*, 3 (1): 26. https://doi.org/10.1525/collabra.107

24. Weiland, P., Baier, C., & Ramthun, R. (2019). *PsychArchives – Unterstützung des Forschungszyklus in der Psychologie mit DSpace.* (Vortrag, 11.04.2019). Bamberg: DSpace Anwendertreffen. http://dx.doi.org/10.23668/psycharchives.2416

25. Linde, K., Rücker, G., Sigterman, K., Jamil, S., Meissner, K., Schneider, A., & Kriston, L. (2015). Comparative effectiveness of psychological treatmens for depressive disorders in primary care: a network meta-analysis. *Family Practice*, 16, 103. https://doi.org/10.1186/s12875-015-0314-x

26. Molloy, G.J., Noone, C., Caldwell, D., Welton, N.J., & Newell, J. (2018): Network meta-analysis in health psychology and behavioural medicine: A primer. *Health Psychology Review*, 12(3), 254-270. https://doi.org/10.1080/17437199.2018.1457449

27. Tonin, F.S., Rotta, I., Mendes, A.M., & Pontarolo, R. (2017). Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. Pharmacy Practice, 15(1), 943. https://doi.org/10.18549/PharmPract.2017.01.943

28. Pigott, T., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods*, 3, 257-268. https://doi.org/10.1002/jrsm.1051

29. Cheung, M. W., & Hong, R. Y. (2017). Applications of meta-analytic structural equation modelling in health psychology: examples, issues, and recommendations. *Health psychology review*, *11*(3), 265–279. https://doi.org/10.1080/17437199.2017.1343678

Electronic Supplements

ESM 1: Literature Search Strategy and Results for Study 1

DOI: http://dx.doi.org/10.23668/psycharchives.4697

1. Literature Search Strategies

2. Results from Database Searches

3. DOIs of the Primary Studies Included

ESM 2: Databases Searched and Primary Studies Included in Study 2

DOI: http://dx.doi.org/10.23668/psycharchives.4698

1. List of Databases in CLICSearch

2. DOIs of the Primary Studies Included

ESM 3: R-Code for all outputs

DOI: http://dx.doi.org/10.23668/psycharchives.4699

ESM 4: Data and Codebooks

DOI: http://dx.doi.org/10.23668/psycharchives.4700

1. Data for Evidence Map (chapter 3)

2. Data for Meta-Analysis on Response Rates (chapter 4)

3. Codebook for Meta-Analysis on Response Rates

4. Data for Meta-Analysis on Panel Conditioning (chapter 5)

5. Codebook for Meta-Analysis on Panel Conditioning

Wissenschaftlicher Werdegang

Akademische Ausbildung

Seit 07/2018	**Promotion in empirischer Sozialforschung**, Universität Trier Thema: Kumulative Meta-Analyse. Robustheit der Evidenz in der Umfrageforschung
04/2011—07/2016	**Master of Arts Wirtschaftssoziologie** (1,2), Universität Trier Schwerpunkt: Industrielle Beziehungen, Sozialstaat und Umverteilung
04/2011—03/2016	**Master of Science Survey Statistics** (1,2), Universität Trier Schwerpunkt: Stichprobendesigns, Panels, Fehlende Daten
10/2007—03/2011	**Bachelor of Science Social Sciences** (1,5), Universität Trier

Auszeichnungen und Stipendien

11/2016	Beste Masterprüfung in Survey Statistics
11/2016	Beste Masterprüfung in Wirtschaftssoziologie
05/2012—03/2014	Stipendiatin der **Studienstiftung des deutschen Volkes**
10/2011—03/2012	Stipendiatin des **Deutschlandstipendium**

Publikationen

Burgard, T., Bosnjak, M., & Studtrucker, R. (under review). PsychOpen CAMA: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods*, xx.

Burgard, T., Bosnjak, M., & Studtrucker, R. (2021). Community-Augmented Meta-Analyses (CAMAs) in Psychology. Potentials and Current Systems. *Zeitschrift für Psychologie*, 229(1), 15-23. https://doi.org/10.1027/2151-2604 /a000431

Wedderhoff, N., Gnambs, T., Wedderhoff, O., **Burgard, T.**, & Bosnjak, M. (2021). On the structure of affect: A meta-analytic investigation of the 24 dimensionality and the cross-national applicability of the positive and negative affect schedule (PANAS). *Zeitschrift für Psychologie, 229*(1), 24-37. https://doi.org/10.1027/2151-2604/a000434

Burgard, T., Bosnjak, M. & Wedderhoff, N. (2020). Response rates in online surveys with affective disorder participants. A meta-analysis of study design and time effects between 2008 and 2019. *Zeitschrift für Psychologie, 228*(1), 14-24. https://doi.org/10.1027/2151-604/a000394

Burgard, T., Wedderhoff, N., & Bosnjak, M. (2020). Konditionierungseffekte in Panel-Untersuchungen: Systematische Übersichtsarbeit und Meta-Analyse am Beispiel sensitiver Fragen. *Psychologische Rundschau, 71*, 89-95. https://doi.org/10.1026/0033-3042/a000479.

Zillien, Nicole; Haake, Gianna; Fröhlich, Gerrit; **Bense, Tanja**; Souren, Dominique (2011): Internet Use of Fertility Patients: A Systematic Review of the Literature. In: *Journal of Reproductive Medicine and Endocrinology* 8, 281-287.

Konferenzbeiträge

Eingeladene Vorträge

Burgard, T., Kasten, N., Bosnjak, M. (2019). *Konditionierungseffekte in Panel-Untersuchungen. Meta-Analyse am Beispiel sensitiver Fragen.* Eingeladener Vortrag beim Panel-Workshop am Robert-Koch-Institut, Berlin.

Präsentationen auf wissenschaftlichen Konferenzen

Burgard, T. (2021). *PsychOpen CAMA – A platform for open and cumulative meta-analyses in psychology*. Talk given at the ESMARConf, January 22, 2021, online.

Bucher, L.; Tran, U.S., Prinz, G.M., **Burgard, T.**, Bosnjak, M., Voracek, M. (2020). *Expanding open and transparent meta-analytic data with PsychOpenCAMA: The implementation of a community-augmented meta-analysis on the Dark Triad of personality*. Paper presented at CSPD 2020, December 7, 2020, online.

Burgard, T., Wedderhoff, N., & Bosnjak, M. (2020). *Moderators of response rates in psychological online surveys over time. A meta-analysis*. Presentation given at GOR 20, September 10, 2020, online.

Burgard, T., Kasten, N., & Bosnjak, M. (2019). *Response rates in psychological online surveys. A meta-analysis on the effects of study design and time*. Presentation given at the ESRA 2019 conference, July 16, 2019, Zagreb, Croatia.

Burgard, T., Kasten, N., & Bosnjak, M. (2019). *Moderators of Panel Conditioning in sensitive questions. A meta-analysis.* Presentation given at the ESRA 2019 conference, July 18, 2019, Zagreb, Croatia.

Burgard, T., Bosnjak, M., & Kasten, N. (2019). *Participation in online surveys in psychology. A meta-analysis.* Presentation given at the Research Synthesis 2019 conference, CAAS, May 30, 2019, Dubrovnik, Croatia.

Kasten, N., **Burgard, T.**, Wedderhoff, O., Bosnjak, M., & Gnambs, T. (2019). *A meta-analytic investigation of the factor structure of the PANAS.* Presentation given at the Research Synthesis 2019 conference, CAAS, May 30, 2019, Dubrovnik, Croatia.

Burgard, T., Bosnjak, M., & Kasten, N. (2019). *Moderators of panel conditioning effects. A meta-analysis.* Presentation given at the 21st General Online Research Conference, TH Köln, March 8, 2019, Köln.

Posterpräsentationen

Burgard, T., Studtrucker, R., & Bosnjak, M. (2020). *Reproducible and dynamic meta-analyses with PsychOpen CAMA.* Poster presented at GOR 20, September 10, 2020, online.