Media and Communication
2024 • Volume 12 • Article 7836
https://doi.org/10.17645/mac.7836

ARTICLE

Open Access Journal

Standardized Sampling for Systematic Literature Reviews (STAMP Method): Ensuring Reproducibility and Replicability

Ayanda Rogge, Luise Anter, Deborah Kunze, Kristin Pomsel, and Gregor Willenbrock

Institute of Media and Communication, TU Dresden, Germany

Correspondence: Ayanda Rogge (ayanda.rogge@tu-dresden.de)

Submitted: 10 November 2023 **Accepted:** 5 February 2024 **Published:** 3 April 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Systematic literature reviews (SLRs) are an effective way of mapping a research field and synthesizing research evidence. However, especially in communication research, SLRs often include diverse theories and methods, which come with a considerable downside in terms of reproducibility and replicability. As a response to this problem, the present article introduces the method of standardized sampling for systematic literature reviews (STAMP). The method is a structured, four-stage approach that is centered around score-based screening decisions. Originating from principles of standardized content analysis, a method common in communication research, and supplementing established guidelines like Cochrane or PRISMA, the STAMP method contributes to more transparent, reproducible, and replicable SLR sampling processes. As we illustrate throughout the article, the method is adaptable to various SLR types. The article also discusses the method's limitations, such as potential coder effects and comparatively high resource intensity. To facilitate the application of STAMP, we provide a comprehensive guideline via the Open Science Framework that offers a succinct overview for quick reference and includes practical examples for different types of SLRs.

Keywords

content analysis; replicability; reproducibility; STAMP method; standardized sampling; systematic literature review

1. Introduction

Systematic literature reviews (SLRs) are an effective way of mapping a research field, synthesizing research evidence, and distinguishing "between real and assumed knowledge" (Petticrew & Roberts, 2006, p. 2). However, when applied in the social sciences, such as in communication research, scholars often face a

theoretically and methodologically diverse range of publications that do not permit quantitative summarization of results. For example, when summarizing evidence on the production or use of news, literature samples might well include both semi-structured interviews and ethnographies as well as standardized or automated content analyses and experiments (e.g., Anter, 2023; Melchior & Oliveira, 2022). Furthermore, the research interests of SLRs in communication research often extend beyond synthesizing knowledge on a specific question. Instead, these SLRs repeatedly aim to build theories, critically investigate a research field, and identify its weaknesses and research gaps (e.g., Engelke, 2019; Ratcliff et al., 2022; see also Paré et al., 2015). Against this background, traditional meta-analyses that aggregate quantitative data from similar studies to determine an overall effect estimate (Davis et al., 2014) are often unsuitable for SLRs in the field of communication. Instead, quantitative analysis is primarily employed for mapping the research field in terms of study characteristics and theories as well as applied methods (e.g., Kümpel et al., 2015). For synthesizing existing evidence, many SLRs also apply qualitative techniques such as textual analysis (e.g., Humayun & Ferrucci, 2022) or open coding of relevant text parts (e.g., Belair-Gagnon & Steinke, 2020). However, despite the necessity and utility of these mixed-methods or qualitative SLRs, these approaches come with a considerable downside in terms of reliability and reproducibility. In fact, it is hardly possible to calculate reliability scores for qualitative content analysis, nor is it feasible to reproduce the analysis precisely since qualitative techniques generally allow for more interpretive freedom. Against this backdrop, the screening and selection process of the publications becomes even more critical for the diverse range of SLR efforts in communication research, including (partly) qualitative, quantitative, computational, or mixed-methods SLRs. Accordingly, this article addresses the need for improved reproducibility and replicability in SLR samples given a heterogeneous research field, advocating for using standardized content analysis to systematically apply eligibility criteria to literature selection.

To date, to the best of our knowledge, this issue has not been thoroughly addressed in established practical guides for SLRs nor in existing SLRs. For instance, while Petticrew and Roberts (2006) offer detailed information about developing criteria for the inclusion or exclusion of publications and conducting the literature search, they appear to "skip" the critical step of applying the criteria to the publications resulting from the literature search. Similarly, the PRISMA protocol (Page et al., 2021, Item 8), which many SLRs in the field of communication already adhere to (e.g., Melchior & Oliveira, 2022; Pirkis et al., 2019), the SLR preregistration form recently provided by van den Akker et al. (2023), and the guidelines provided by the Methods of Synthesis and Integration Center (Pigott & Polanin, 2020) emphasize the importance of the screening stage and discuss different approaches (e.g., priority screening for only highly relevant records or single screening of all records). Still, they neither offer guidance on how to apply the eligibility criteria systematically to specific parts of the publications nor do they elaborate on efficiently documenting the screening and selection process—especially for the sake of reproducibility.

Correspondingly, many existing SLRs in the field of communication describe both their literature search procedure and the development of their criteria in a detailed manner (e.g., Belair-Gagnon & Steinke, 2020; Engelke, 2019). However, information about the inclusion/exclusion process often remains limited to whether and when researchers applied criteria to the abstract or the whole text. In addition, it is often unclear under what criteria researchers consulted the abstract or the full text for their inclusion decision (e.g., Hase et al., 2023), or full texts are not consulted at all (e.g., Joris et al., 2020). Therefore, even though the screening result (inclusion/exclusion) is sometimes provided, readers and researchers may find it challenging to replicate the sampling process or even reproduce the sampling decision for single

publications—especially regarding the content criteria, such as for assessing the thematic fit of a publication. This issue becomes even more significant in the case of SLRs that do not (e.g., Gambo & Özad, 2020) or only vaguely (e.g., Castells-Fos et al., 2023) provide their eligibility criteria.

In response to the issues regarding the replicability and reproducibility of SLR sampling processes outlined above, we argue that the screening and selection of literature (i.e., sampling) significantly benefit from standardized content analysis, a method that is genuine to and prevalent within communication research (Haim et al., 2023, p. 280). The method is described as a valid and replicable measurement of texts' meaning by assigning categories to content (Krippendorff, 2004, p. 18). This describes what standardized SLR sampling should involve: predefined categories (i.e., eligibility criteria) that are systematically applied to specific text parts in an academic publication (i.e., title, keywords, abstract, or full text). This coding provides the basis for inferring a publication's fit with the SLR's research interest. Understanding SLR sampling as standardized content analysis also means making use of the research techniques that have been developed to ensure reproducibility and replicability—methodological requirements that do not only apply to content analysis but have been dealt with prominently in the respective literature (e.g., Haim et al., 2023; Lombard et al., 2002). Most basically, applying these techniques instructs researchers not only to report the eligibility criteria and the inclusion results of their SLR sampling (as is already common for SLRs in our field) but also to plan, conduct, and document the whole sampling process systematically and transparently.

To guide this process, we propose a four-stage approach for systematically sampling publications applicable to the variety of SLR efforts in communication research. At its core, the procedure employs scores that quantify a paper's eligibility for the respective SLR to standardize and protocol the sampling process, which is why we propose our so-called STAMP method for STandardized sAMPling. Specifically, the STAMP method includes:

- Stage 1: The development of eligibility criteria for including publications and reflecting the review's objective and scope;
- Stage 2: The identification of potentially relevant literature through databases;
- Stage 3: Narrowing the sample by assigning an abstract-based screening (ABS) score;
- Stage 4: Determining the final sample by assigning each publication a full-text reading (FTR) score.

We particularly emphasize the importance of thorough documentation of every step in the review protocol, which functions as a codebook for the content analyses conducted in Stages 3 and 4, when the eligibility criteria are applied to the publications in order to deduce the score-based screening decision. This continuous documentation of each screening stage allows researchers to constantly reflect on the SLR's progress to conform to its goals. The procedure also increases both the replicability and the reproducibility of the SLR's findings: Using STAMP and its rigorous documentation enables both replicating the SLR with newly compiled literature samples and reproducing the coding process within the same literature sample. Another benefit of the proposed method is its customizability to different research questions and techniques: Our prior SLRs show that the procedure is convenient for both evidence-summarizing SLRs that investigate a clearly defined research field (Anter, 2023) and concept-building SLRs that have to iteratively approach a literature sample to provide a valid theory synopsis (Rogge, 2023). Moreover, our approach involves the strength of measuring sample reliability as the ABS and FTR stages evaluate the content fit of publications by applying the principles of standardized content analysis to the SLR sampling process.

Importantly, we do not consider STAMP as a substitute for established SLR procedures, such as the Cochrane guidelines for systematic reviews (Higgins et al., 2023) or the PRISMA 2020 statement (Page et al., 2021). Rather, we see it as a *supplement* that refines the sampling process in order to increase the reproducibility and replicability of SLRs. Both the Cochrane and the PRISMA guidelines originate from the context of health care research, whereas STAMP is based in the field of communication. Thus, STAMP uses the opportunity to connect field-specific practices of communication research with established SLR guidelines. Inherently, the STAMP method aligns with guidelines established by Cochrane and PRISMA, employing their key practices such as using a review protocol, defining eligibility criteria, and documenting the search strategy. Additionally, STAMP provides more concrete guidance for inclusion and exclusion decisions than Cochrane and PRISMA, which merely stress the significance of this process in general and provide basic suggestions regarding the sampling process, such as pre-testing the eligibility criteria, training the (independent) reviewers (Cochrane), and transparently documenting the number of reviewers per publication (PRISMA). Moreover, while Cochrane and PRISMA primarily address possible bias within individual publications as well as bias within the SLR's synthesis, STAMP focuses on the *reduction* of bias within the sampling process.

To present our methodological procedure and facilitate its adoption by researchers planning to conduct an SLR, we describe the standardized sampling method for SLRs in detail. In addition, we provide a guideline including practical examples where the method has been successfully tested in concept-building (inductive) and evidence-summarizing (deductive) SLRs via OSF. These SLRs cover diverse communication research areas, such as health communication, human–machine communication, and journalism studies, highlighting the adaptability of the STAMP method.

2. The Four Stages of STAMP

In this section, we present the method designed to guide communication researchers through a systematic and reproducible sampling process for SLRs. We introduce the four stages, as summarized in Figure 1, as well as their goals, procedures, and practical recommendations. There are different ways to apply the approach to SLRs, so a guideline including practical examples of the four-stage procedure is available on OSF (https://bit.ly/4atGsvN).

2.1. Stage 1: Eligibility Criteria

The primary goal of this stage is to transparently establish the scope and objectives of the SLR.

The foundation of this stage lies in the specification of a review protocol, an established way to document SLRs (Nightingale, 2009; for an example, see also Page et al., 2021). First, a review protocol summarizes essential components such as the study's rationale, definitions of pivotal terms, and research questions/hypotheses. Second, the review protocol defines relevant keywords for constructing and validating the search string (Stage 2). To define the keywords for the SLR, researchers should choose keywords from pertinent literature as a starting point, such as background literature and related work (Mishra et al., 2009). The exploration of existing literature is crucial in order to identify synonyms for terms in other disciplines or earlier works. This initial selection of keywords should be enriched by using the snowball method or own keywords. Third, the review protocol contains the SLR's eligibility criteria, that is, criteria for deciding which publications remain in the literature review and which are excluded (Petticrew &

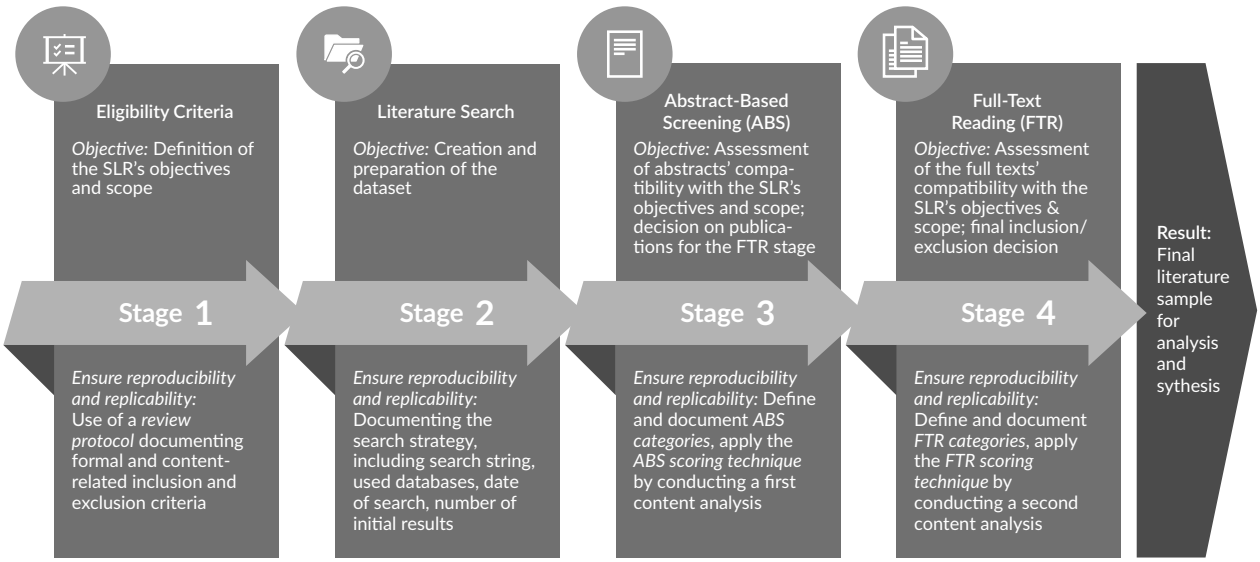

Figure 1. Illustration of the four stages of the STAMP method.

Roberts, 2006). These criteria comprise both formal (e.g., the language of the publications, publication years and types, countries of origin, and scientific disciplines) and content-related factors (e.g., topics or variables based on research questions). Defining and transparently documenting the inclusion and exclusion criteria is essential to enable a transparent research process. These eligibility criteria are also used for the ABS (Stage 3) and FTR (Stage 4). In a nutshell, the review protocol explicitly defines the observer-independent rules that govern the sampling process and are applied to all units of analysis in the abstract and full-text stages—a key requirement for replicability (Krippendorff, 2004, p. 19). Therefore, it is a central component of our method and it is conceived to take on the role of a codebook in Stages 3 and 4.

At its core, the first stage aims at setting up the review protocol and the eligibility criteria that ensure systematic alignment with the stipulated research objectives and scope. Thus, the first stage lies the foundation for an efficient and systematic review of the literature.

Is the SLR's scope feasible? Researchers need to consider available resources when planning and conducting an SLR. Primarily, these resources include time, personnel, and access (to databases, literature, etc.). The availability of these resources will impact the duration of the SLR project. We suggest that researchers prepare a project plan and consider the possibility of reducing the SLR scope if resources are limited. Also, regardless of the available resources, one should be open to adjusting the project plan and scope once researchers become more familiar with the nature of their analysis material.

2.2. Stage 2: Literature Search

The main objective of this stage is to conduct a literature search and collect the sampling units for the content analysis in Stages 3 and 4, i.e., a dataset including all possibly relevant publications.

For this purpose, researchers define the search strategy in the review protocol and document the search. This documentation includes the search string (Bartels, 2013; Mishra et al., 2009) as well as all employed bibliographic databases, search engines, and repositories. A well-curated search string equips the team to

discover a broad spectrum of related literature and is pivotal to discerning the core consistencies or discrepancies across the research landscape. The search strategy, and especially the search string, should be developed iteratively. Hence, exploring and documenting different search string combinations is essential. The quality of the search strategy can be assessed by the amount of possibly relevant literature that the search generates. For example, researchers can include only search strings that yield a defined number of results. Additionally, researchers should predefine pertinent publications, authors, or journals in the protocol whose appearance in the search results validates the search strategy. After defining and validating the search strategy, the actual literature search is executed and the results are transferred into a dataset, including the publications' abstracts (see Stage 3). For all databases, search engines and repositories, the same search strategy should be applied. Duplicates resulting from searching multiple databases must be cleared in the dataset and logged in the protocol to ensure reliability.

The second stage revolves around systematically sourcing and organizing relevant literature—again, an already established procedure (Page et al., 2021). As a result, the first dataset of publications for the SLR emerges from Stage 2.

How to decide on databases? The selection of databases (e.g., Web of Science), repositories (e.g., arXiv), and scholarly search engines (e.g., Google Scholar) depends on the SLR's underlying rationale. While field-specific databases (e.g., Communication & Mass Media Complete) are particularly relevant for SLRs focusing on a narrow and discipline-specific topic, others encompass a broader research landscape (e.g., Scopus; Falagas et al., 2008). If the SLR needs to include grey literature, researchers should also use "regular" search engines (Falagas et al., 2008). In any case, we recommend using multiple databases to combine the strengths of different providers.

How to decide on a narrow or broad search string? Depending on the SLR's objectives and scope, search strings can either be sensitive (i.e., include a broad range of keywords and their synonyms) with the aim of exhaustion or specific (i.e., include a small selection of pertinent keywords) with the aim of precision (Petticrew & Roberts, 2006). For example, concept-building SLRs could use a broader search string that employs multiple theoretical dimensions or incorporates interdisciplinary approaches. In comparison, SLRs aiming to summarize evidence on a specific question might use a narrower search string. However, depending on the characteristics of the research field, it might also be necessary to use a more sensitive search string. This means a search string does not only depend on the type of SLR but must also capture the scope of the research objective. Thus, an evidence-summarizing SLR can also use broad terms in the search string if the associated research question requires this search approach. Options to expand the search string include the generous use of synonyms (Anter, 2023; Rogge, 2023) or wildcards (Kunze, 2024). Importantly, a broad search string inherently results in a larger data set, which may require drawing a random sample in Stages 3 or 4 to render the SLR feasible.

How to create a search string? We recommend using advanced search techniques such as Boolean operators, truncations, and wildcards in search terms to amplify the search's reach (e.g., Bartels, 2013; Mishra et al., 2009). Here, different databases might have distinct requirements, such as word limits causing minor variations in the search string. To assure the search's reproducibility and replicability, the review protocol should contain every search string and, if applicable, additional filters and limitations used for the respective database or search engine.

What to do with additional literature? Researchers might identify additional relevant literature *after* the systematic search is carried out (e.g., through recommendations of colleagues or reviewers). If researchers consider ignoring this literature as detrimental, it is suitable to define an extended sample (Petticrew & Roberts, 2006), including all publications sampled because of the researchers' subjective decision. To ensure transparency, each publication's origin (systematic vs. extended sample) should be documented in the final sample.

2.3. Stage 3: ABS

The third stage focuses on an ABS of the publications collected in Stage 2. The primary purpose is to assess the compatibility of each abstract with the research objectives.

In Stage 3, researchers read all abstracts and assess their fit regarding the eligibility criteria. Therefore, inclusion and exclusion criteria (see Stage 1) are transferred into categories for a content analysis of the publications' abstracts (ABS categories), again defined in the review protocol, which now takes the function of a codebook. These categories are used to assess the fulfillment of the respective criteria through coding the abstracts. For each eligibility criterion, it is possible to use one or multiple ABS categories. While most formal criteria will only need one ABS category (e.g., language), using multiple categories for more complex content criteria (e.g., study findings) might be suitable. Moreover, as not every relevant section of a full article may be addressed in the abstract, it might be pragmatic to limit the ABS inclusion criteria to the presence of keywords in the abstract (e.g., a specific theory, topic, or method) and formal features of the publication (e.g., type). For instance, if an abstract mentions a relevant theory, concept, medium, method, or population, the respective category is coded with 1. As is apparent from this example, we recommend using binary coding, indicating whether an eligibility criterion (in the form of an ABS category) is *met* (1) or *not met* (0), but ordinal coding (e.g., 2 = *met*, 1 = *possibly met*, 0 = *not met*) is also reasonable for a more detailed evaluation. Additionally, researchers should adopt a "generous" coding attitude at this stage: In case of uncertainty about the fulfillment of a category, researchers can still include the respective publication and "postpone" the final decision to the following FTR stage. In order to ensure the reproducibility and replicability of coding, all coders should be provided with detailed coding instructions, precise definitions, and examples for each category (Krippendorff, 2004, p. 132). Additionally, as also highlighted in the Cochrane guidelines, we recommend coder training during which coders become familiar with the categories and code a small number of abstracts together with the head researcher, allowing the computation of intercoder reliability and quality assessments.

Based on the abstract coding, the ABS score is calculated as a sum index of all categories representing a publication's fit: the higher the ABS score, the better the fit. If multiple categories per eligibility criterion were used (e.g., several categories to assess the thematic fit), it is also possible to calculate separate scores per criterion. While an aggregated ABS score provides a holistic overview and simplifies the inclusion, scores per ABS category provide a more differentiated understanding of an abstract's alignment with various eligibility criteria. Usually, only publications that fulfill all ABS categories remain in the dataset. Alternatively, researchers can define a threshold value beforehand. The level of detail of the ABS assessment and the use of thresholds depends on the requirements of the literature review. For example, evidence-summarizing SLRs usually want to include only studies that deal with a specific question. Therefore, it is usually sufficient to directly transfer the eligibility criteria in an aggregate ABS score and exclude all publications that do not reach their maximum value (e.g., Anter, 2023). A concept-building SLR rather employs multiple ABS

categories reflecting different dimensions of the topic. In comparison to the evidence-summarizing SLR, the inclusion metric is less rigorous; abstracts must, for instance, address at least half of the ABS categories to remain for the FTR (e.g., Rogge, 2023).

Stage 3 comprises a truncated content analysis of the given abstracts. This screening process provides no analytical depth with regard to the SLR's research questions. Instead, its purpose is filtering, distinguishing relevant from less relevant literature. The benefit of the proposed scoring technique lies in a transparent, reasonable, and reproducible evaluation of the abstracts. As a result, the ABS score determines the publications that remain in the dataset for Stage 4.

How to deal with missing abstracts? This is particularly likely if an SLR also includes chapters in books or edited volumes, which often do not provide an abstract. In these cases, researchers can calculate the ABS score based on the publication's introduction and conclusion. Importantly, this "workaround" should be transparently documented in the review protocol.

Are the provided ABS decisions appropriate? Although it might be tempting, we do not recommend validating the ABS scores by delving deeper into the full texts. Not only could this cause self-referential loops, where researchers recurrently adjust the ABS categories and their coding decisions without moving to the next stage, but it also includes the risk of adjusting the ABS categories based on a potentially biased selection of full texts. To address uncertainty, we instead recommend using ordinal categories whose middle category indicates uncertainty. These publications can be re-assessed either during the next stage or at the end of the ABS phase. After reading a fair amount of both highly and hardly relevant abstracts, researchers usually gain sufficient experience to make clear coding decisions.

How to assure impartiality? An essential aspect of the ABS (and the FTR, Stage 4) is impartiality. For example, researchers might (unwillingly) be more generous with publications from scholars who are established and well-known in the field and oversee less prominent research that could be even more relevant. Therefore, researchers should ideally remain unaware of the authors behind each abstract/full text. A blind procedure is easily implemented in Stage 3 by anonymizing all bibliographic information in the data set (Stage 2) with an ID, which becomes the exclusive reference during Stages 3 and 4. Consequently, researchers can realize blind sampling at a low cost by using the export functions of today's databases with subsequent anonymization.

How to incorporate tools for automation? Automation tools can assist researchers in dealing with a large number of relevant publications, particularly when coding for relevant formal and content-related categories in ABS and FTR. When integrating automation tools into the STAMP method, it is essential to thoroughly document in the review protocol which tools (when and with which settings/outcomes) were used. One example of an SLR automation tool is the open-source software ASReview (van de Schoot et al., 2021). ASReview employs a machine-learning algorithm trained by the researchers to rank publications regarding their relevance to the SLR. This training process should be carefully documented, as there is no standardized guideline for when to end the training. While ASReview has the potential to reduce the amount of (irrelevant) publications that have to be screened during Stage 3, such automation tools limit the comprehensibility of the ABS coding and, in turn, the calculation of an ABS score. Consequently, while ASReview enhances efficiency, especially when a high number of publications have to be screened, it could hinder the detailed and transparent coding technique employed by the STAMP method.

2.4. Stage 4: FTR

In the final stage of the STAMP method, researchers perform an FTR of the publications remaining in the dataset and evaluate their fit with regard to the FTR categories by employing an FTR score. This stage aims to evaluate articles exclusively based on their contribution to the research interest and to create a final literature sample for the subsequent analysis and synthesis.

The FTR again utilizes content analytical principles to maintain consistency with the ABS stage. This includes creating an FTR category system with (binary/ordinal) categories, providing coding guidelines to ensure reliable assessment and comparability, as well as providing a coding scheme to document all coding decisions and the calculated FTR score(s). Therefore, similar to Stage 3, researchers create FTR categories based on the inclusion and exclusion criteria (eligibility criteria). As the FTR stage usually focuses on the publications' fit regarding *content*, the content criteria based on the SLR's research questions/hypotheses are transferred into concise categories—as before, documented in the review protocol serving as a codebook. For example, FTR categories could assess whether a publication investigates a specific variable relation or uses certain definitions and theories. Methodological aspects (e.g., study design) and results (such as effect sizes) might also be part of the FTR categories. Depending on the SLR's objective and scope, the FTR categories may be identical to the former ABS categories. In these cases, the researcher can confirm or revise a coding decision from Stage 3 based on the full text. This approach is convenient in evidence-summarizing SLRs with clearly defined research questions. On the other side, concept-building SLRs may require a more iterative procedure, which implies specifying the categories from the ABS to the FTR stage. Iteratively developing categories based on the material is necessary when the scope of a category system cannot be estimated initially. Although this procedure includes more effort, it brings the benefit of pretesting categories, reflecting on the purpose of a category, and inductively developing categories based on the given full texts. Moreover, especially for SLRs focusing on theoretical constructs, determining the fulfillment of inclusion criteria through standardized categories might be challenging. In these cases, it is important that the category system provides indicative phrases (e.g., "anchor examples" that include typical phrases of a definitional approach) for each category to facilitate the coding process.

After preparing the category system, researchers (roughly) read the full texts and code whether they contain relevant passages or arguments. As with the previous stage, an unbiased approach is crucial, so the evaluation should ideally be conducted blindly using the ID of an article. In Stage 4, blind sampling can be efficiently implemented by saving the full texts using the ID and blanking all bibliographic information. Again, there are different options to create an FTR score. Binary coding is appropriate if researchers only want to decide *if* a full text contains important information to answer a research question/hypothesis. For instance, the FTR category "topic" could indicate if a publication deals with a topic *within the scope of the SLR* (1) or *not within the scope* (0). Another option would be to evaluate the importance of a publication per eligibility criterion, e.g., through categories counting relevant arguments that a full text provides. For example, if the SLR aims to derive a definition, the associated categories capture how many definitional statements the full texts contain. As with the ABS, the FTR score can be calculated as a sum score per eligibility criterion or aggregated into a comprehensive FTR sum score. The FTR score then serves as an indicator of a publication's overall importance regarding the review's objectives.

Researchers transparently have to define the required value of the FTR score for a publication to remain in the final literature sample. Again, usually, only publications with the maximum score are included, as only these

fulfill all required eligibility criteria. However, defining a score threshold (cut-off criterion) might be suitable as well. For example, concept-building SLRs that aim at defining a concept might be confronted with the problem that almost all publications meet the categories because most publications include (short) sections that define pertinent concepts. For synthesis, however, it is more useful to include only publications that contribute substantially to the definition of a concept. In this case, a defined amount (e.g., the upper third) of publications with the highest FTR score (which indicated the number of, e.g., definitional statements) is retained for the analysis. This approach follows the premise that publications with a higher FTR score are particularly suitable to answer the research questions/hypotheses.

In the FTR stage, researchers conduct a second truncated content analysis based on the full texts, ensuring each included publication contributes to answering the SLR's research questions. Each publication is evaluated through FTR scoring based on clearly defined FTR categories. As a result, researchers end up with the final—transparent and reproducible—literature sample for the subsequent analysis and synthesis.

What if the dataset is too large to perform the FTR? Researchers' resources might be insufficient for reading all publications included during the ABS stage. In this case, we suggest drawing a random sample from the ABS publications to proceed with the FTR scoring technique. A randomized sample presents a cross-section of the research landscape under investigation. Since all relevant publications have an equal chance to be considered for the FTR sample (and therefore the later analysis), this does not violate the systematic approach of the SLR and, thus, presents a valid alternative to full suspension.

How can the subsequent analysis be prepared during the FTR stage? We recommend copying the text passages that were decisive for the coding into a data extraction form (e.g., a spreadsheet or an online survey; Petticrew & Roberts, 2006). Using additional software, such as survey tools, supports researchers to comfortably collect and synthesize data during the FTR stage, which reduces human labor costs and prepares the literature sample to be further processed during the phase of result synthesis by using automated content analysis. In addition, such software could not only be used to document the coding but also to automate the screening decisions (e.g., through IF/THEN conditions), which facilitates reproduction by future researchers.

3. Discussion

We introduced the STAMP method, a score-based sampling technique designed to enhance the reproducibility and replicability of SLRs by standardizing the literature selection procedure. Based on the principles of standardized content analysis, our proposed approach draws on the strengths of a genuine communication research method to offer advancements over preceding SLR methodologies, emphasizing aspects of scientific quality that are "seminal" for content analysis (Haim et al., 2023, p. 281) and add value to a diverse range of systematic research efforts.

3.1. Transparency and Reliability

Applying the STAMP method to SLRs improves *transparency* since each screening decision is systematically documented. In accordance with established approaches, our procedure also focuses on the review protocol (Nightingale, 2009), but deviates by integrating principles from standardized content analyses: The review

protocol not only documents all decisions made while designing the SLR. It also serves as a codebook that transfers eligibility criteria into a comprehensive category system with transparent coding guidelines. Consequently, the abstracts and full texts yielded by the systematic literature search are not merely used for heuristically deciding on a publication's inclusion—They are units of analysis, systematically coded with intersubjective categories. These categories, in turn, lead to scores that quantify the publications' relevance for the SLR. Through translating the selection process into two truncated content analyses during Stages 3 and 4, our approach also enforces standardized documentation, which enables assessing the reliability of the sampling. To be able to quantify the agreement of ratings from multiple coders is a fundamental prerequisite for replicability and reproducibility of content analysis (Krippendorff, 2004, p. 211). Hence, as for standardized content analysis, both intra-coder and inter-coder reliability scores can be computed for the ABS and the FTR stage, facilitating a reproducible and replicable sampling process.

To do so, each inter-coder *reliability* test should start with training, where all raters discuss the codebook. For example, an inter-coder reliability test could be conducted with two or three raters, which code around 10% of the dataset's publications (Petticrew & Roberts, 2006). After separate coding procedures, reliability coefficients can be calculated (for an example of inter-coder reliability assessment, see Joris et al., 2020). Both inter- and intra-coder reliability coefficients can be computed separately for each ABS/FTR category as well as for the overall ABS/FTR scores. We recommend calculating and reporting a variety of coefficients like Krippendorff's alpha, Cohen's kappa, and Holsti's coefficient (e.g., Kunze, 2024). For Krippendorff's alpha and Cohen's kappa coefficients, it should be considered that homogeneity (i.e., little variance) and dichotomy (binary coding decisions like *a criterion is met* [1]/*not met* [0]) of the data could negatively influence those coefficients. Besides these coefficient-specific annotations, reliability tests still identify potential coder effects and increase the clarity and intersubjectivity of the ABS/FTR categories. Employing coefficients brings the benefit of a comprehensible decision basis for including/excluding categories after pretesting for the complete coding procedure, for example, defining 0.8 as the acceptance threshold for content-related and 1.0 for formal categories (Lombard et al., 2002, p. 593).

Reproducibility and replicability require open reporting (Freiling et al., 2021). Accordingly, the method section should include the research question(s), eligibility criteria, sample sizes during Stages 2 to 4, and information on the coding process (how many coders were involved in the sampling, when and how coder training took place, and coefficients for intra- and inter-coder reliability). As suggested within the STAMP guideline template (https://bit.ly/4atGsvN), the SLR's appendix should additionally include the review protocol and the codebooks for Stages 3 and 4 (including definitions and examples for each category as well as the construction of the ABS and FTR scores). Ideally, ABS and FTR coding sheets with basic bibliographic information are also made accessible (for instance, via OSF).

3.2. Adaptability and Validity

In addition to the aforementioned advantages, one major methodological contribution of the STAMP method is its *adaptability*. STAMP is not only adaptable to different publication types—for the content analyses during Stages 3 and 4, it does not matter whether the studies are qualitative, quantitative, computational, or mixed-methods—but also to diverse types of SLRs (Paré et al., 2015) that are common for the interdisciplinary field of communication research. In our article, we repeatedly compared concept-building SLRs and evidence-summarizing SLRs. Our prior SLRs fit in that spectrum, demonstrating

the STAMP method's practicability. For instance, Rogge (2023), as a concept-building SLR, sought to unify a multifaceted term within an interdisciplinary research domain. This necessitated an inductive and iterative STAMP application: Every stage used standardized categories, but they were refined between Stages 3 and 4. In comparison, Anter (2023), within a well-defined research field and, therefore, at the evidence-summarizing end of the spectrum, utilized a rigidly deductive STAMP method with predefined criteria and fixed categories. Finally, our method remains fitting for intermediary SLRs such as the one by Kunze (2024), which combines both efforts. In this study, one research question aims to define a term (inductive, concept-building), while another investigates factors influencing the dependent variable (deductive, evidence-summarizing). Accordingly, the STAMP method was realized both with more fixed categories and categories based on an iterative reflection in relation to the available material.

In this sense, our approach also contributes to the *validity* of the literature sample: A thought-through and comprehensive codebook is the first step towards ensuring the validity of content analysis (Potter & Levine-Donnerstein, 1999, p. 266). Iteratively refining the eligibility criteria allows tailoring the respective categories to the unique material of analysis. At the same time, researchers are continuously forced to check the alignment of the sampled material with the SLR's objectives. Nevertheless, another necessary step to ensure validity is to assess the coding decisions against external standards (Potter & Levine-Donnerstein, 1999, p. 266). Examples of these standards include using relevant literature and expert consultations when designing and defining the SLR, as well as using pertinent literature for validating the search string.

3.3. Impartiality and Intersubjectivity

Especially for coding latent variables such as the content of a text, objectivity is "not a realistic expectation" (Potter & Levine-Donnerstein, 1999, p. 265). In these cases, it is all the more important to ensure the *intersubjectivity* and *impartiality* of coding decisions. Building on the standards of content analysis, STAMP includes multiple efforts to tackle potential biases in the sampling process. Among others, this is the persistent focus on eligibility criteria and their application by trained coders and on the basis of an established coding scheme (Stages 1–4), combining strengths of different databases in the search strategy (Stage 2), and employing scores instead of heuristic assessments for screening decisions (Stages 3 and 4). During conducting an SLR, biases can also arise through priming effects when prominent authors or popular journals are prioritized over lesser-known publications that could, however, equally or even more strongly contribute to answering the research questions. This is particularly problematic when such biases lead to systematic exclusion (and therefore, discrimination) of certain scientific publications—for example, by scholars from the Global South (Chakravartty et al., 2018). The STAMP method's hallmark is its blind assessment of the abstracts and full texts. We recommend disguising author names and—if compatible with the SLR's formal eligibility criteria—journal names in the reading phases (Stages 3 and 4). This enables blind sampling and prevents confirmatory tendencies or biased screening decisions. A paper is therefore judged based on its contribution to the research question, which means that every publication has the same chance of remaining in the sample as long as it fulfills the predefined and intersubjective assessment criteria. Accordingly, blind sampling increases impartiality and further reduces the risk of citation circles (continuously referring to the same authors within a particular topic), which would significantly limit the validity and intersubjectivity of the SLR's results.

4. Conclusion

SLRs are essential in synthesizing the abundance of available knowledge in the field of communication and related disciplines, building and further developing theoretical concepts, and consolidating empirical insights. Yet, the shortcomings of SLRs in our field often lie in their transparency, reproducibility, and replicability, challenging the foundation upon which syntheses are built. Recognizing gaps in existing methodologies, we proposed the STAMP method, a score-based sampling technique rooted in the principles of a method genuine to communication science: standardized content analysis. STAMP's methodological contribution in extension to existing SLR procedures is constituted in quantifying screening decisions and standardizing the sampling process. Further, several strengths characterize the method: transparency and reliability through full documentation and score-based screening decisions, adaptability to different SLR types, sustained validity evaluation, and measures towards increasing impartiality and intersubjectivity. The novelty of STAMP lies not in a sophisticated protocol or complex search strategies, but in a four-step model that guides and standardizes the sampling process for SLRs in a comprehensive and practical way. Moreover, since STAMP is rooted in a method genuine to communication science, it contains nothing alien to communication researchers, but brings together common methodological competencies with regard to systematic reviews and is, therefore, easy to adopt by communication scholars.

Many of the components of the STAMP method are already established. This applies not only to the definition of inclusion and exclusion criteria, as described above. Review protocols are already used in various SLRs in communication research as well (e.g., Melchior & Oliveira, 2022). Moreover, various SLRs in our discipline also include a precise description of the search string and the search results (e.g., Belair-Gagnon & Steinke, 2020) or distinguish between abstract and full-text screening (e.g., Ratcliff et al., 2022). However, due to a lack of publicly accessible documentation, replicability and reproducibility are often limited to single parts of the sampling process. Replicating and reproducing the whole sampling procedure is only guaranteed by a systematic and transparent approach as proposed by STAMP. Thus, applying STAMP for SLRs in our field would introduce an additional instance of control for these studies and their particular validity claim.

Of course, the STAMP method also comes with some limitations. Most obviously, while it ensures a reliable and valid sample construction, it only partially influences the reliability and validity of the synthesis itself: A proper sample is a necessary prerequisite for reliable and valid analysis but is not sufficient on its own. Additionally, as with every content analysis, the method is prone to coder effects. Consider, for example, learning effects: As coders become more experienced, their knowledge of the respective research field increases. Consequently, their coding decisions might become more elaborate, which would be detrimental to the validity as well as inter- and intra-coder reliability. Finally, the STAMP method is costly, as it includes detailed documentation, developing and pretesting codebooks and—if several coders are involved—training and supervising coders.

Notwithstanding these shortcomings, the STAMP method helps researchers to further increase the transparency and, ultimately, the reproducibility and replicability of SLRs. Via OSF, we provide a detailed yet concise guideline that encapsulates the four stages of STAMP and includes practical examples for inductive, deductive, and mixed SLR approaches.

Acknowledgments

The authors wish to thank the three anonymous reviewers and the academic editors of this thematic issue for their valuable and constructive comments.

Funding

The article processing charges were funded by the joint publication funds of the TU Dresden, including Carl Gustav Carus Faculty of Medicine, and the SLUB Dresden as well as the Open Access Publication Funding of the DFG. No further funding was received for conducting the research presented in this article.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online at https://bit.ly/4atGsvN

References

Anter, L. (2023). How news organizations coordinate, select, and edit content for social media platforms: A systematic literature review. *Journalism Studies*. Advance online publication. https://doi.org/10.1080/1461670X.2023.2235428

Bartels, E. M. (2013). How to perform a systematic search. *Best Practice & Research Clinical Rheumatology*, *27*(2), 295–306. https://doi.org/10.1016/j.berh.2013.02.001

Belair-Gagnon, V., & Steinke, A. J. (2020). Capturing digital news innovation research in organizations, 1990–2018. *Journalism Studies*, *21*(12), 1724–1743. https://doi.org/10.1080/1461670X.2020.1789496

Castells-Fos, L., Pont-Sorribes, C., & Codina, L. (2023). Decoding news media relevance and engagement through reputation, visibility and audience loyalty: A scoping review. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2023.2239201

Chakravartty, P., Kuo, R., Grubbs, V., & McIlwain, C. (2018). #CommunicationSoWhite. *Journal of Communication*, *68*(2), 254–266. https://doi.org/10.1093/joc/jqy003

Davis, J., Mengersen, K., Bennett, S., & Mazerolle, L. (2014). Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus*, *3*(1), Article 511. https://doi.org/10.1186/2193-1801-3-511

Engelke, K. M. (2019). Online participatory journalism: A systematic literature review. *Media and Communication*, *7*(4), 31–44. https://doi.org/10.17645/mac.v7i4.2250

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Freiling, I., Krause, N. M., Scheufele, D. A., & Chen, K. (2021). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*, *71*(5), 686–714. https://doi.org/10.1093/joc/jqab032

Gambo, S., & Özad, B. O. (2020). The demographics of computer-mediated communication: A review of social media demographic trends among social networking site giants. *Computers in Human Behavior Reports*, *2*, Article 100016. https://doi.org/10.1016/j.chbr.2020.100016

Haim, M., Hase, V., Schindler, J., Bachl, M., & Domahidi, E. (2023). (Re)establishing quality criteria for content analysis: A critical perspective on the field's core method. *SCM Studies in Communication and Media*, *12*(4), 277–288. https://doi.org/10.5771/2192-4007-2023-4-277

Hase, V., Mahl, D., & Schäfer, M. S. (2023). The "computational turn": An "interdisciplinary turn"? A systematic review of text as data approaches in journalism studies. *Online Media and Global Communication*, *2*(1), 122–143. https://doi.org/10.1515/omgc-2023-0003

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2023). *Cochrane handbook for systematic reviews of interventions: Version 6.4*. Cochrane Training. https://training.cochrane.org/handbook

Humayun, M. F., & Ferrucci, P. (2022). Understanding social media in journalism practice: A typology. *Digital Journalism*, *10*(9), 1502–1525. https://doi.org/10.1080/21670811.2022.2086594

Joris, G., De Grove, F., Van Damme, K., & De Marez, L. (2020). News diversity reconsidered: A systematic literature review unraveling the diversity in conceptualizations. *Journalism Studies*, *21*(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. SAGE.

Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115610141

Kunze, D. (2024). *Systematizing destigmatization in the context of media and communication: A systematic literature review*. Manuscript in preparation.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Melchior, C., & Oliveira, M. (2022). Health-related fake news on social media platforms: A systematic literature review. *New Media & Society*, *24*(6), 1500–1522. https://doi.org/10.1177/14614448211038762

Mishra, S., Satapathy, S. K., & Mishra, D. (2009). Improved search technique using wildcards or truncation. In R. Raghavan (Ed.), *2009 International Conference on Intelligent Agent & Multi-Agent Systems*. IEEE. https://doi.org/10.1109/IAMA.2009.5228080

Nightingale, A. (2009). A guide to systematic literature reviews. *Surgery*, *27*(9), 381–384. https://doi.org/10.1016/j.mpsur.2009.07.005

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Wiley. https://doi.org/10.1002/9780470754887

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Pirkis, J., Rossetto, A., Nicholas, A., Ftanou, M., Robinson, J., & Reavley, N. (2019). Suicide prevention media campaigns: A systematic literature review. *Health Communication*, *34*(4), 402–414. https://doi.org/10.1080/10410236.2017.1405484

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. https://doi.org/10.1080/00909889909365539

Ratcliff, C. L., Wicke, R., & Harvill, B. (2022). Communicating uncertainty to the public during the Covid-19 pandemic: A scoping review of the literature. *Annals of the International Communication Association*, *46*(4), 260–289. https://doi.org/10.1080/23808985.2022.2085136

Rogge, A. (2023). Defining, designing and distinguishing artificial companions: A systematic literature review. *International Journal of Social Robotics*, *15*, 1557–1579. https://doi.org/10.1007/s12369-023-01031-y

van den Akker, O. R., Peters, G.-J. Y., Bakker, C. J., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Moreau, D., Nordström, T., Pickering, J. S., Riegelman, A., Topor, M. K., van Veggel, N., Yeung, S. K., Call, M., Mellor, D. T., & Pfeiffer, N. (2023). Increasing the transparency of systematic reviews: Presenting a generalized registration form. *Systematic Reviews*, *12*, Article 170. https://doi.org/10.1186/s13643-023-02281-7

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*, 125–133. https://doi.org/10.1038/s42256-020-00287-7

About the Authors

Ayanda Rogge (MA, TU Berlin) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her dissertation focuses on human–machine communication and social robotics. In her project, she explores the social design of artificial companions, with a particular interest in their communication and interaction characteristics allowing users to perceive a technical agent as an artificial companion.

Luise Anter (MA, LMU Munich) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. In her dissertation, she explores how the characteristics of social media platforms shape journalistic production processes. Her other research interests include the perception and use of news and information in social media environments, news bias, and the interaction between journalists and their sources.

Deborah Kunze (MA, Leipzig University) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her research interests focus on the fields of health communication (e.g., the promotion of cancer prevention) and media psychology (e.g., effects on stigmatizing attitudes). In her dissertation project, she explores how destigmatization can be fostered, with a focus on the potential of media and communication in this context.

Kristin Pomsel (MA, TU Dresden) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her research interest focuses on political communication processes, media literacy, and media trust. Her dissertation project deals with recipients' expectations and perceptions of news quality as well as the actual quality of alternative and established German news media.

Gregor Willenbrock (MA, HMTM Hanover) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. His research interests include exploring the dynamics of online communities and their collaborative endeavors in peer production, specifically within the context of the networked public sphere. He is also dedicated to investigating the use of computational methods in communication science research, extending their application beyond the scope of his current project.