

Google Insights for Search: A Methodological Innovation in the Study of the Public Agenda?

Michael Scharkow*
Universität der Künste Berlin

Jens Vogelgesang**
Freie Universität Berlin

February 6, 2009

First Draft for Comment

Prepared for DGPuK FG Journalismus & Methoden,
Berlin, Feb 2009

*Institut für Theorie und Praxis der Kommunikation, Universität der Künste Berlin,
scharkow@udk-berlin.de

**Institut für Publizistik- und Kommunikationswissenschaft, Freie Universität Berlin,
gesang@zedat.fu-berlin.de

Abstract

The reliable and valid measurement of issue salience is a key problem of agenda setting research. While exposure, awareness and salience are usually investigated using surveys, the behavioral consequences of those processes, like follow-up communication or information seeking, lend themselves very well to observation. Many theoretical models of agenda setting incorporate these concepts, but only few empirical studies exist that actually measure salience-driven behavior. We propose a new method for measuring aggregate issue salience by analyzing data from search queries typed into Google. We illustrate this approach with a case study from the 2005 Bundestag elections in Germany, focussing on the fiscal expert Paul Kirchhof of Angela Merkel's campaign team. Using both survey and online observation data, we find substantial correlations between the two longitudinal measures of issue salience.

Keywords: agenda setting, issue salience, information seeking, search engine

If you type Google into Google,
you can break the Internet.

*(Jen Barber, Head of IT in the
British Comedy "The IT-Crowd")*

1 Introduction

The research note we present here explores the idea of measuring the public agenda without employing classical survey techniques. Drawing on data from the web service Google Insights for Search (GIFS), we examine the possibilities of using online observation to measure issue salience.

We claim that the traditional first-level approach of agenda setting comprising of cognitive processes like awareness and salience can be refined by broadening the theoretical perspective to include behavioral aspects. Although cognitive by nature, first and subsequent audience responses to media content (Becker et al., 1975) only occur when individual action has taken place beforehand (to select), in between (to think about), and afterwards (to select again). These behaviors form an essential part of the agenda setting process.

In contrast to researchers examining agenda setting effects such as candidate choice or turnout at the polls, we theoretically and empirically focus on more immediate audience responses. First, we present a blueprint for a broadened theoretical framework of first level agenda setting processes. Second, we recapitulate how to measure the public agenda by means of public opinion polls, and introduce GIFS as an alternative observational tool to measure the public agenda. Third, we present a case study and report cross check findings about the convergent validity of the GIFS results. Fourth, we conclude by summing up the potential merits and drawbacks of using GIFS for further agenda setting research.

2 Theoretical Model

The theoretical blueprint we present here (Fig. 1) refers to what is called first-level agenda setting and draws on the classic model by Becker et al. (1975). According to this model, agenda setting comprises of consecutive cognitive processes that involve several states which are entangled by threshold mechanisms. The initial state of a first-level agenda setting process is called selection (e.g. reading a newspaper). After that decision is made and the newspaper is flipped open, the reader may become aware of some issue with a certain

probability. Becoming aware of some newspaper issue may result from the individual's disposition (e.g. level of concentration) on the one hand, and the characteristics of the newspaper content (e.g. caption style) on the other hand. However, when the reader has become aware of a certain issue, state change has taken place. While the selection state is a prerequisite condition for the awareness state transition, the awareness state, in turn, is a prerequisite condition for the salience state transition. State change from awareness to issue salience may, in turn, depend on the individual's disposition (e.g. issue preference) on the one hand and the issue's characteristics (e.g. obtrusiveness) on the other hand. Put more general, a first-level agenda setting process resembles a cascade dynamics model. The dynamics of such a cascade model can be described as a stochastic process with a Markov property.

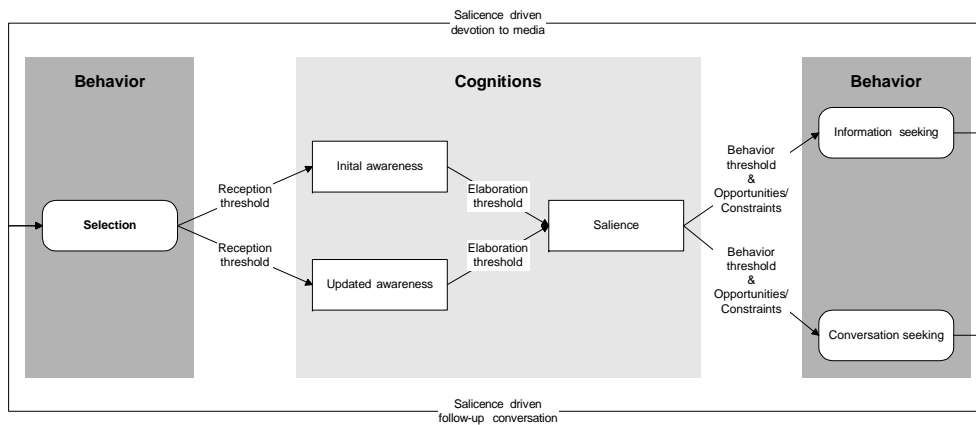


Figure 1: Cascade dynamics of agenda setting

Classical first-level agenda setting is typically confined to cognitive audience processes. Within the agenda setting tradition, salience has been described as “the degree to which an issue on the agenda is perceived as relatively important” (Dearing & Rogers, 1996, p. 8). Issue salience is the theoretical starting point for two additional facets of first-level agenda setting we want to shed light on subsequently. We follow the ideas of McLeod et al. (1974) that classical first-level agenda setting has to be both theoretically and empirically enriched by considering consecutive behavioral audience responses stemming from a prior salience state. The foregoing cognitive cascade may activate a newspaper's reader to learn more about a special issue (i.e. salience driven devotion to media content) or may motivate the reader to seek opinions about

that special issue (i.e. salience driven conversation with other people).

From a measurement point of view, awareness and salience refer to intrapersonal latent states for a given issue and person, whereas both salience driven devotion to media or conversation people refer to manifest states. The latter result from former latent audience states and, thus, represent behavioral consequences of issue salience. A valid measure of the public agenda, however, has to consider both facets of issue salience, i.e. the latent and the manifest audience state. While the measurement and modeling of the individual states and transitions is certainly a worthy challenge for any communication researcher, we pursue a more modest goal in this paper: the measurement of the manifest (behavioral) dimension of agenda setting.

3 The Measurement of the Public Agenda

3.1 Measuring the cognitive dimension using surveys

The reliable and valid measurement of issue salience is a key problem of agenda setting research. Issue salience is typically measured by means of special questions in public opinion polls (Dearing & Rogers, 1996, p. 45-49) focussing on some facets (i.e. perceived issue salience or interpersonal issue salience) of the "multi-faceted concept" (Gadziala & Becker, 1983, p. 122) termed salience. However, public opinion polls may only provide some vague and indirect measures of intrapersonal latent audience states, since researchers can only observe what people *tell* about the most important problem facing the country or the respondent today. Even with the most sophisticated statistical techniques which combine data from closed- and open-ended survey questions, agenda setting researchers do have to rely on the *answers* given by the respondents. This is equally true for studies in which behavioral aspects of agenda setting, such as interpersonal conversations, were measured using survey responses (an early example is Atwood et al., 1976).

Notwithstanding the general problems related to this kind of method bias, the problem of *how* to ask about issue salience (or one of its subdimensions) remains to this day (cf. McLeod et al., 1974; Niemi & Bartels, 1985; Edelman, 1993). The frequent use of the "most-important problem" question can often be attributed to convention rather than methodological thought (cf. Wlezien, 2005). We argue that observation of audience behavior can often provide a more direct and less problematic measure of (consequences of) issue salience, especially when using unobtrusive methods of data collection.

3.2 Measuring the behavioral dimension using (online) observation

The behavioral consequences of issue awareness and salience, information seeking and follow-up conversations are basically observable processes, there are very few agenda setting studies that use observation as a means of measurement. Indeed, unlike in journalism studies (Quandt, 2008) or media psychology (Ravaja, 2004; Unz & Schwab, 2005), observation is rarely used in mass communication research. This is most obviously due to the difficulty of observing more than a handful of selected people at the same time outside of a laboratory. Consequently, the one prominent study where naturalistic observation was used in agenda setting context, conducted by Kepplinger and Martin (1986), was quite limited in scope.

A possible solution to the small sample problem lies in automated observation tools that work without human observers. The most commonly used methods of this kind are audience ratings (Webster et al., 2000), which enable researchers to observe the media use of many recipients in real time. However, these data are less well-suited for agenda setting research because (a) interest can only be measured at the granularity of programmes not issues, (b) the users' possibilities for salience-driven information seeking are limited in linear media like TV¹ and (c) information seeking may not be the only relevant motive for program choice.

A relatively new and, we argue, promising way of unobtrusive observation has been available since the advent of the internet, both for observing follow-up communication and information seeking. Any communication that happens via protocols like HTTP (World Wide Web), SMTP (Email) or NNTP (Usenet) can be logged, stored and analyzed with some technical but little human effort. The study of online communication blurs the boundaries of observation and content analysis as well as mass and interpersonal communication, and provides researchers with almost endless streams of data. Consequently, many issues in communication research can and often have been studied online.

Roberts et al. (2002) examined salience-driven follow-up communication in electronic bulletin boards (EBB). The authors analyzed the content of four media outlets as well as AOL's political message board. For three out of four selected issues (immigration, health care, taxes and abortion) significant positive cross-correlations between the media content and the online discussions

¹One could argue that at least non-interest in an issue can be measured by the subsequent switching of the programme.

were found, leading the authors to conclude that “[m]edia coverage apparently can provide individuals with information to use in their Internet discussions” (Roberts et al., 2002, p. 464).

While the analysis of user generated content, like message boards, comments, chats or Twitter messages, promises to be a great method for the measurement of follow-up communication, we are interested here in observing information seeking. Fortunately, the internet provides users with a tool for this task – search engines which are arguably the first, and possibly only, device for information seeking for millions of people. According to recent estimations by comscore, U.S. online users submitted 137 billion search queries in 2008, which means 1.7 queries per day and user (South Florida Business Journal, 2008). Following Cohen’s famous dictum, this is certainly a lot of information about what people think and like to know more about.

We argue that search queries are close to perfect as an indicator of issue salience: Compared to survey questions, there is no interviewer bias or social desirability involved, the measurement is completely unobtrusive and happens in the field. Moreover, for many users there is virtually no effort involved in using search engines, compared to buying a book or searching a paper encyclopedia, so that the threshold from issue salience to active information seeking is quite low. Of course, the demographic of internet users is still different from the general population, so that we cannot take the complete public agenda, as measured by search queries, at face value. But unless the mechanisms leading people to look for information about salient issues are fundamentally different for onliners than for the rest of the population, we see no problems for agenda setting research in this respect.

The central problem until recently has been the lack of availability of search query data to most researchers. Basically, only the providers of search engines like Google, Yahoo or Microsoft are able to collect such data which are heavily used by the companies themselves in order to optimize services for users and advertising partners.² Applications of scientific analysis of those queries were therefore quite rare and prominent, possibly because it could be demonstrated that searches for pornographic and illegal downloads accounted for the most frequent queries (Silverstein et al., 1999).

Fortunately, the world’s largest search engine, Google, has recently begun to make aggregate log file results available to the public. Google started their web service Insights for Search (available at <http://google.com/insights/search>) in

²Technically, internet service providers (ISP) that offer dial-in, DSL or cable connectivity could log search queries as well as any other data streams on the TCP or protocol level.

August 2008 as a follow-up to their earlier Google Trends site. GIFS provides public access to Google's logged search queries and allows for filtering by search term, time frame and region. Unlike older services, users cannot only see graphical presentations of these data, but download actual data tables of the search volume for a particular query. In the remainder of this paper, we will investigate the utility of this data for agenda setting research.

4 Case Study

4.1 Method

In order to check the validity of Google search queries for the measurement of the public agenda, we compare the aggregate search query data provided by GIFS with aggregate survey data from about 500 telephone interviews conducted daily by the FORSA institute. If our hypothesized model is true, we expect to see a strong correlation between the two time series.

As a case study we use a single-issue study on the German General Election 2005. Specifically, we focus on Paul Kirchhof, a former judge at the Federal Constitutional Court, Professor at the University of Heidelberg, and fiscal expert in Angela Merkel's campaign team. Kirchhof's controversial ideas on a flat income tax as well as later comments on social issues and gender roles generated much media attention and debate, leading to his withdrawal from the shadow cabinet just before the election:

Thus, Kirchhof became yet another case in recent German electoral politics where an outsider was invited to join the political game, earned substantial initial appreciation, but ending up as a scapegoat for competing parties who had no problem finding ammunition in his person and his *vita* for their ruthless attacks. (Schmitt-Beck & Faas, 2006, p. 407)

Nonetheless, Paul Kirchhof and his flat tax proposal were dominant topics of the 2005 Bundestag campaign and proved consequential for the election results (Wüst & Roth, 2006).

We chose the *issue* Paul Kirchhof for this case study mainly for pragmatic reasons: Firstly, Kirchhof appears in public only for a limited period during the campaign, from his nomination in Angela Merkel's government team on August, 16th until some time after election day on September, 18th. This

enables us to restrict the coding of survey answers and data retrieval from Google to about 6 weeks. Secondly, a person's name can rather unambiguously be used in search queries, making it easier for the coders of the survey data to find references to this issue. Applying our approach to more complex issues will certainly require more thought and effort.

Search Query Data

The data on aggregate search queries was retrieved from the Google Insights for Search website with the following settings (see Figure 2): The selection of data was restricted to Germany, as defined by certain ranges of IP numbers from which the search query originated.³ The time range requested was August/September 2005, from which we removed the first two weeks afterwards.

Two important restrictions are currently implemented in GIFS: Daily search volume data can only be retrieved for up to 8 weeks from Google, for longer periods only weekly data is provided which is of less use for most agenda setting research. This could be remedied by sliding a two-month window over the period of interest, yielding 6 results sets for a year. However, the second restriction from GIFS makes merging these data difficult: The search volume is only available as normalized data, based on the relative frequencies of search queries, with the peak day within the period as a reference point. This is less problematic for our time series correlation, but estimates of absolute measures are currently impossible to obtain with GIFS.

The specification of the query string for GIFS is very basic. There is no support for complex queries connected by AND or NOT, no fuzzy matching or wildcards (cf. Hollanders & Vliegenthart, 2008). For many applied research cases, manual specification of different query variants and aggregation of the retrieved data will be necessary. For our case study, a simple search string is easily constructed: In order to account for different spellings of the name we use the following search terms, the plus sign marks a Boolean OR: 'paul kirchhof' + kirchhof + kirchhoff + 'paul kirchhoff'.

Using this query yields the result page depicted in Figure 2 which includes a graphical display of the time series, matching results from Google's news archive, and a downloadable CSV file with the numbers. The latter is only available to logged-in user while the rest of the interface is public.

³GIFS makes it possible to compare or restrict results to the level of Bundesländer. However, since we cannot verify the validity of the filter, we do not use this level of detail in our analysis.

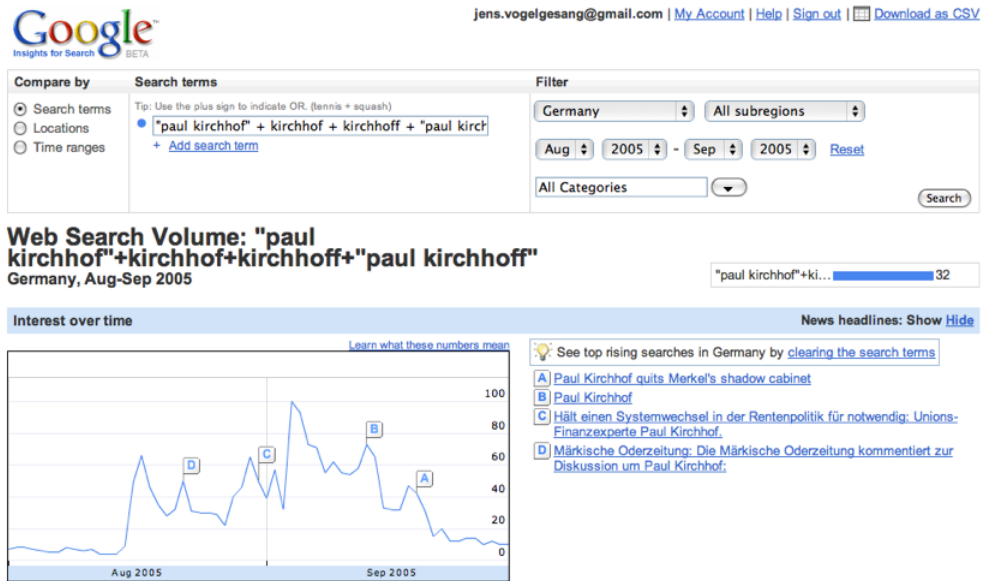


Figure 2: Result page of Google Insights for Search

To sum up: The data collection using GIFS is free of cost, easy to understand and very quick, even when some tuning of the search query is required. Another big advantage is the availability of daily search volume data without missings, which makes time series analysis a lot easier.

Survey Data

The survey-based time series in our analysis is the aggregate number of responses to an open ended question about important issues. The data was collected by FORSA using standardized computer-assisted telephone interviews. About 500 interviews were conducted every Monday to Friday during the period of interest in 2005. Unfortunately, this means that there are no survey data available for the weekends. In order to test the impact of the missing data we conducted separate analyses with and without imputed missings.

The question asked was: "Can you recall any important issues recently covered in the news media that interest you?" and the respondents were entered as free text by the interviewers.

The wording of the question is somewhat multi-dimensional and does not clearly capture one of the processes described in section 2. Given that it relates to issues recently covered by the media, we argue that the item is an indicator

for awareness or perceived issue salience rather than individual salience. We do not expect this to be an issue for correlational analyses but caution against inferring absolute levels of issue salience from this question.

Two student coders then analyzed all responses and coded every answer referring to Paul Kirchhof. As mentioned before, the effectiveness and reliability in the codings is due to the simple nature of the topic of interest. The coding was part of a larger research project, so the results were already available. For future analyses with equally simple search terms, a dictionary based automatic coding is certainly feasible (cf. Krippendorff, 2004).

4.2 Results

As can be expected for any minor topic, at least compared to issues like unemployment, only a small fraction of the respondents named Paul Kirchhof an important topic. Kirchhof does not appear on the public agenda until his nomination and then quickly peaks at about 3 per cent of all answers, then a rather volatile level of interest and again a small rise in the week before the election (see Fig. 3).

The time series for Google search queries related to Paul Kirchhof looks very similar to the aggregate survey responses. There is a sharp rise in interest on September 4th, the day of the first TV debate between Merkel and Schröder, in which Kirchhofs proposals for taxation were heavily debated. Another peak in search queries occurs some 10 days later, after Kirchhof suggested forming a team with prominent CDU fiscal expert Friedrich Merz.

After the visual inspection of the data, we computed a simple correlation between the two time series. The coefficient of $r = .49$ is quite large and highly significant ($p < .01$). Note that correlations of this strength are not unusual when using aggregate data. But there are still some caveats to consider: We did not specify ARIMA cross-correlations and transfer functions (Krause & Fretwurst, 2007) because of too few data points and too much missing survey data, but there is an AR(1) process in the Google search data which is hardly surprising. We expect that some of the common variance between the two time series is due to both following this autoregressive pattern.

Furthermore, we chose the time frame of the analysis so that, contrary to many long-time agenda setting studies, our data are not zero-inflated. Consequently, the correlation is not due to very frequent non-occurrences of events that dominate both time series (cf. Krause & Gehrau, 2007, p. 200). We conclude that the substantial correlation between two time series collected

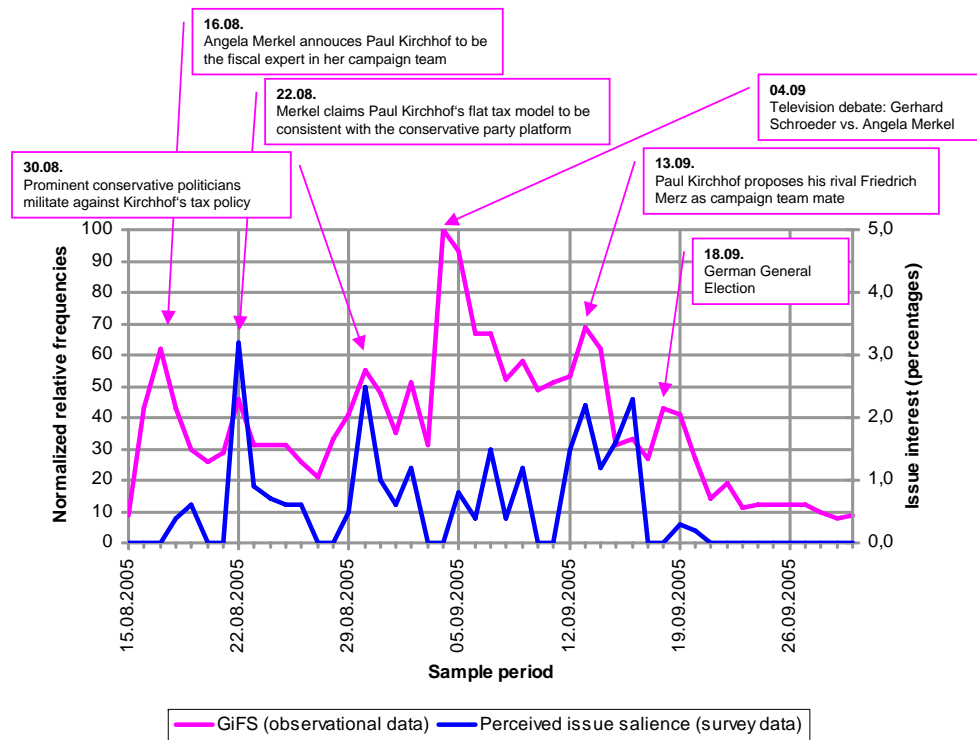


Figure 3: Survey responses and search queries compared for Paul Kirchof

with different means is a strong indicator that using search queries is a valid method for the measurement of the public agenda.

5 Discussion

In this research note we introduced a novel way of measuring the salience of issues by using aggregate data from queries to the Google search engine. Although our results are based on a rather simple case study, we are confident that using Google Insights for Search provides a powerful tool for the study of agenda setting processes. Moreover, since these data are available for virtually all imaginable topics, for many countries and even regions, and go back for several years, a plethora of interesting empirical research projects may profit from this data source – analyses of media resonance to public relations and advertising campaigns, cross-country and over-time comparisons of public

interest in political and social issues, even forecasting tomorrow's news from today's recipients' information seeking.

There are, however, some challenges and problems that need to be addressed, both methodologically and empirically:

1. We don't know yet how many people turn to Google for a given issue, and how this relates to the number of respondents in telephone surveys. The absolute amount of Google search queries related to, for example, Paul Kirchof is not (yet) publicly available, although it may be possible to infer these numbers using various other sources. This is less of a problem for agenda setting research using time-series correlations.
2. On a related issue, the normalization of GIFS data makes comparisons between countries, regions or time spans difficult. Further refinement in dealing with data provided by Google will be necessary.
3. The processes that lead to behavioral consequences of issue awareness and salience are not understood well enough. We don't know under which circumstances and when people turn to Google or their friends and colleagues in order to learn more about a salient issue.
4. Since the survey respondents and Google users are certainly from different populations, it is yet unclear whether the observed correlation of issue awareness and information seeking can easily be generalized. After all, internet users are at least younger and better educated than the general population. This becomes even more important when actual agenda setting processes relating media content with the public agenda are investigated online (cf. Roberts et al., 2002).

It is our hope that both the theoretical model and the measurement issues outlined in this paper will stimulate further agenda setting research online. We are anxious to see follow-up studies using more complex issues and covering a longer time span, and confident that the method proposed here will turn out to be useful. The incredible amount of communication taking place on the internet does not only necessitate much more theoretical, methodological and empirical scholarship but is also a great chance for exciting and rewarding research projects.

Acknowledgements

We thank Prof. Manfred Güllner (forsa Gesellschaft für Sozialforschung und statistische Analyse mbH) who provided us with the raw survey data. We are especially grateful to Ana Ivanova and Christiane Waas who coded the open ended survey responses.

References

- Atwood, L., Sohn, A., & Sohn, H. (1976). Community Discussion and Newspaper Content. In *Annual Convention of the Association for Education in Journalism*. University of Maryland.
- Becker, L., McCombs, M., & McLeod, J. (1975). The Development of Political Cognitions. In S. Chaffee (Ed.), *Political Communication* (pp. 21–63). Sage.
- Dearing, J. & Rogers, E. (1996). *Agenda-Setting*. Sage.
- Edelstein, A. (1993). Thinking About the Criterion Variable in Agenda-Setting Research. *The Journal of Communication*, 43(2), 85–99.
- Gadziala, S. & Becker, L. (1983). A new look at agenda-setting in the 1976 election debates. *Journalism Quarterly*, 60(1), 122–126.
- Hollanders, D. & Vliegthart, R. (2008). Telling what yesterday's news might be tomorrow: Modeling media dynamics. *Communications*, 33(1), 47–68.
- Kepplinger, H. & Martin, V. (1986). Die Funktion der Massenmedien in der Alltagskommunikation. *Publizistik*, 31(1-2), 118–128.
- Krause, B. & Fretwurst, B. (2007). Kurzfristige Agenda-Setting-Effekte von Fernsehnachrichten Eine Zeitreihenanalyse am Beispiel Ausländerfeindlichkeit und Rechtsradikalismus. In B. Krause, B. Fretwurst, & J. Vogelgesang (Eds.), *Fortschritte der politischen Kommunikationsforschung: Festschrift für Lutz Erbring*. VS Verlag für Sozialwissenschaften.
- Krause, B. & Gehrau, V. (2007). Das Paradox der Medienwirkung auf Nichtnutzer. *Publizistik*, 52(2), 191–209.
- Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage.

- McLeod, J., Becker, L., & Byrnes, J. (1974). Another Look At the Agenda-Setting Function of the Press. *Communication Research*, 1(2), 131.
- Niemi, R. & Bartels, L. (1985). New Measures of Issue Salience: An Evaluation. *Journal of Politics*, 47(4).
- Quandt, T. (2008). Methods of Journalism Research: Observation. In M. Löffelholz, D. Weaver, & A. Schwarz (Eds.), *Global journalism research: Theories, methods, findings, future*. (pp. 131–142). Malden, Oxford, Carlton: Blackwell.
- Ravaja, N. (2004). Contributions of Psychophysiology to Media Research: Review and Recommendations. *Media Psychology*, 6(2), 193–235.
- Roberts, M., Wanta, W., & Dzwo, T. (2002). Agenda Setting and Issue Salience Online. *Communication Research*, 29(4), 452.
- Schmitt-Beck, R. & Faas, T. (2006). The Campaign and its Dynamics at the 2005 German General Election. *German Politics*, 15(4), 393–413.
- Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33 (pp. 6–12): ACM New York, NY, USA.
- South Florida Business Journal (2008). Comscore: Online traffic soars, sales growth slows. Jan 30, 2009.
- Unz, D. & Schwab, F. (2005). Viewers viewed: Facial expression patterns while watching TV news. In L. Anolli, S. Duncan, M. Magnusson, & G. Riva (Eds.), *The hidden structure of social interaction. From Genomics to Cultural Patterns*.
- Webster, J., Phalen, P., & Lichty, L. (2000). *Ratings Analysis: The Theory and Practice of Audience Research*. Lawrence Erlbaum Associates.
- Wlezien, C. (2005). On the salience of political issues: The problem with ‘most important problem’. *Electoral Studies*, 24(4), 555–579.
- Wüst, A. & Roth, D. (2006). Schröder’s Last Campaign: An Analysis of the 2005 Bundestag Election in Context. *German Politics*, 15(4), 439–459.