Jürgen Maier/Michaela Maier/
Marcus Maurer/Carsten Reinemann/
Vincent Meyer (eds.)

Real-Time Response Measurement in the Social Sciences

Empirische und methodologische Beiträge zur Sozialwissenschaft

Herausgegeben von Jürgen Falter, Jürgen Maier,
Katja Neller und Harald Schoen

26

PETER LANG Internationaler Verlag der Wissenschaften

Jürgen Maier/Michaela Maier/
Marcus Maurer/Carsten Reinemann/
Vincent Meyer (eds.)

Real-Time Response Measurement in the Social Sciences

Methodological Perspectives and Applications

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at <http://dnb.d-nb.de>.

ISSN 0172-1739
ISBN 978-3-631-57743-1

www.peterlang.de

Contents

RTR Measurement in the Social Sciences: Applications, Benefits, and some Open Questions 1
Marcus Maurer & Carsten Reinemann

Methodological Perspectives

Measuring Spontaneous Reactions to Media Messages the Traditional Way: Uncovering Political Information Processing with Push Button Devices 15
Jürgen Maier & Thorsten Faas

Is RTR Biased Towards Verbal Message Components? An Experimental Test of the External Validity of RTR Measurements 27
Carsten Reinemann & Marcus Maurer

Reactivity of Real-Time Response Measurement: The Influence of Employing RTR Techniques on Processing Media Content 45
Andreas Fahr & Annette Fahr

Measuring the Perception and the Impact of Verbal and Visual Content of Televised Political Ads: Results from a Study with Young German Voters in the Run-Up to the 2004 European Parliamentary Election 63
Michaela Maier & Jürgen Maier

The Explanatory Power of RTR Graphs: Measuring the Effects of Verbal and Nonverbal Presentation in Persuasive Communication 85
Thomas Roessing, Nikolaus Jackob & Thomas Petersen

Advantages and Limitations of Comparing Audience Responses to Televised Debates: A Comparative Study of Germany and Sweden 97
Michaela Maier & Jesper Strömbäck

Applications

Tracing Ad Experience: Real-Time Response Measurement in Advertising Research 117
Gary Bente, Lisa Aelker & Mirjam Fürtjes

Immediate Responses to Political Television Spots in U.S. Elections: Registering Responses to Advertising Content 137
Lynda Lee Kaid

Applied Dial Testing: Using Real-Time Response to Improve Media Coverage of Debates 155
Dan Schill & Rita Kirk

Health Message Primes and Sexual Health Campaign Messages: Evaluating Viewers' Real-Time Responses 175
John C. Tedesco & Adrienne Ivory

Immigration as Translated by the Media: (Re)Production, Representation and Response to Images 193
Vincent Meyer & Céline Ségur

Contributors 207

RTR Measurement in the Social Sciences: Applications, Benefits, and some Open Questions

Marcus Maurer & Carsten Reinemann

1. Why traditional survey research sometimes fails

Without any doubt, the formation of attitudes during the reception of communicative stimuli is a process rather than a single event. Whether people like a movie, a TV commercial, or the performance of a political candidate in a televised debate is the result of several verbal and nonverbal impressions during the reception adding up to a final opinion. Nevertheless, traditional study designs in communication research measure the effects of communicative stimuli *after* the audience has been exposed to the *entire stimulus*. Consequently, they treat communicative stimuli – intentionally or not – as one single unit. This holds true for laboratory experiments, measuring short term effects right after the treatment, and even more for field studies, measuring long term effects by comparing content analysis and survey data under real life conditions. In both cases, changes in attitudes, cognitions, or emotions can be traced back to changes in media content but they cannot be traced back to specific elements of media content. In other words: Using questionnaires and other traditional methods in media effects research might give a good impression of *whether* communicative stimuli changed recipients' attitudes – but they do not give any hint *why* recipients changed their attitudes. From traditional survey research we can learn that people like a movie, a TV commercial, or the performance of a candidate in a debate. But we cannot learn which elements of the film, which persuasive strategies in the commercial, and which arguments or gestures of the candidate made people like it or him.

If we want to know the causes of opinions or opinion changes, why not ask respondents directly? There are several reasons: Imagine respondents watching a movie or a televised debate lasting 90 minutes or more. On the one hand, they will be able to state whether they liked the movie or which candidate in their opinion won the debate. On the other hand, they will not be able to remember every single aspect of the plot or every single argument used in the debate. When asked about their reasons for liking a movie or a candidate's performance, viewers will mention some aspects they still remember. These will not necessarily be the ones which impressed them most. Rather, the possibility that, e. g., an argument in a debate is remembered increases when it has been

presented at the beginning (primacy effect) or at the end (recency effect) of the debate. Moreover, it increases when the argument fits to the respondent's former opinions (consistency effect). Arguments in the middle of the debate or inconsistent arguments are forgotten much more quickly. Consequently, respondents might give invalid answers because they have already forgotten arguments which had been important for their opinion formation while others seem to be important just because they come to mind first.

But even if respondents could remember every single argument during a debate they simply might not know which of them were most important for their opinion formation. Because opinion formation is a rather unconscious process, respondents are sometimes wrong about their criteria for forming attitudes. For example, they tend to mention arguments as being important just because this seems to make sense afterwards (rationalization) or because it seems to make them look like good citizens (social desirability). Consequently, in traditional surveys most people tend to state that they liked a movie because of its witty plot rather than of its good looking actresses and that they liked a candidate's performance because of his plans for economic growth rather than of his sympathetic smile.

Taken together, post stimulus questionnaires seldom lead to valid results about the causes of respondent's opinions and opinion changes. Traditional survey research fails when the process of opinion formation during the reception of communicative stimuli is under examination. In this case, a research tool is needed which continuously measures recipients' impressions during the reception of the stimulus. By matching those measures with elements of the stimulus, it is possible to trace back recipients' impressions to every single verbal or visual signal: issues, arguments, rhetorical strategies, gestures, music etc. This tool is called real-time response measurement (RTR) or continuous response measurement (CRM).

2. Measuring audiences' responses in real-time

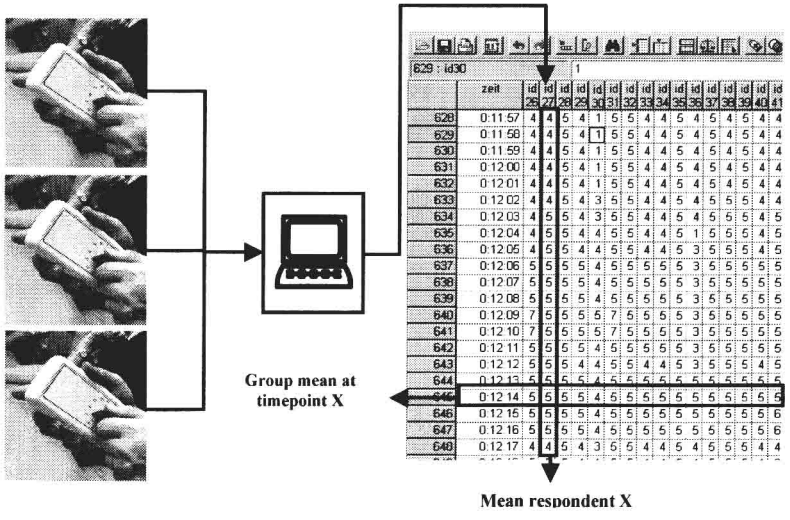
2.1. A short history

RTR measurement is not a new idea. Paul F. Lazarsfeld and Frank N. Stanton were the first to employ this technique with respect to media content in the 1930s (for a detailed overview of the history of RTR measurement see Levy 1982; Millard 1992). Their "Princeton Radio Research Project" was based on the so-called "program analyzer". Listeners of a radio program were asked to continuously indicate their impressions using two buttons – one green for positive, one red for negative impressions. Pushing no button indicated "indifference". The data could be analyzed on the individual level. In addition,

an aggregated line was generated representing the overall impression of the listeners. Based on these data, it was possible to identify those passages of the radio program that evoked the most pronounced reactions (see, e.g., Peterman 1940; Peatman & Hallonquist 1950). The measurements were often complemented by a questionnaire and focus group discussions afterwards (see Millard 1992). The early program analyzer was later used by CBS into the 1980s.

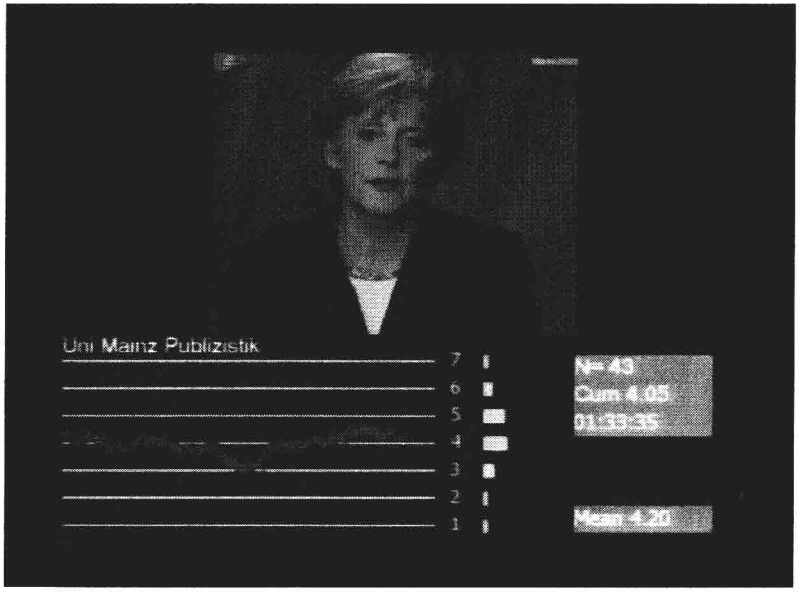
Since the 1950s, the input devices were steadily improved. The aim was to enable participants to give more specific, differentiated judgements. This seemed necessary as the early studies had shown that participants were characterised by rather high inhibition thresholds, which in turn had the consequence that they only indicated extreme sentiments – positive or negative (see Hallonquist & Suchman 1944: 326). As a result, in 1954 McCann-Erickson, Inc. developed the Televac system – a joystick-driven device measuring audience responses on a four-point scale (ranging from "very favorable" to "very unfavorable"). A five-point scale including a neutral mid-point was specifically rejected – mainly because the mid-point has been seen as a "don't know" response, causing problems in statistical analysis. Nevertheless, Televac was a first step into using dial devices, allowing participants to express their reactions on a metric dimension rather than in dichotomous ways (see Millard 1992; Biocca et al. 1994).

Figure 1: Measuring respondents' immediate impressions by RTR



Today, most RTR systems work with dial input devices (usually metric scales ranging from 1 to 7) or sliders (usually metric scales ranging from 0 to 10). Up to 250 respondents can take part in one session. The individual status of each respondent's input device is continuously recorded and sent to a central computer in intervalls which can be defined by the researcher (usually every second). The data is collected by the computer and transferred in a) a graph which can be seen during the session and matched with the stimulus on, e. g., the television screen by video overlay and b) a data set allowing for further analysis. The data can later be analyzed on the individual or aggregate level (see 2.3.). This procedure is shown in Figure 1. Figure 2 shows a graph generated by video overlay during the 2005 German general election televised debate. The line shows the immediate reactions of a group of voters to a statement by then challenger Angela Merkel. On the right side of the graph, time code and means are indicated.

Figure 2: RTR analysis by video overlay



Which kind of input devices are used is, of course, not irrelevant. Push buttons record respondents impressions only when they strike a key. If they do not, no signal is transmitted and the system automatically changes to the neutral position (reset mode). This is not the case when dial devices are used. Here, participants are asked to change the position of their input devices, whenever

their impression changes. If they do not, their most recent impression is recorded on and on (latched mode). Generally, the latched mode seems to generate much more data because respondents can give graded answers. On the other hand, it has been argued that dial devices measuring metric dimensions require more cognitive resources on behalf of the participants (see Baggaley 1987). While there is not much research on that issue, so far, this question will be intensively discussed in the contribution by Faas and Maier to this volume.

2.2. Fields of application

Generally, RTR is nothing else than a computerized version of a questionnaire. While in traditional surveys several questions are asked at one point in time, in RTR surveys one question is asked continously during a given time period. Therefore, RTR theoretically has a broad range of applications. In fact, RTR has been regularly used to measure recipients' perceptions of a communicative stimulus, recipients' evaluations of the stimulus, recipients' attention towards the stimulus, and recipients' emotions while being exposed to the stimulus (for a broader discussion see Biocca et al. 1994: 25ff.). These different measures can be best explained by having a further look at the most common fields of application:

For quite a long time, RTR has been more or less exclusively used in commercial research. Radio and TV stations employed it to test their programs, Hollywood studios to test their movies, and advertising agencies to test their commercials (see, e. g., Gitlin 1994: 32ff.; for an overview see also Biocca et al. 1994: 23). While these are still important fields of application, more recently, social sciences adapted RTR for their purposes. In academic research, RTR is often used to discover the effects of persuasive messages. Just like in applied research there are several RTR studies on the persuasive effects of commercial and political ads (see Hughes 1992; Thorson & Reeves 1985; see also the contributions of Bente et al. and Kaid to this volume). Other studies deal with the effects of educational television, mainly in the context of health communication (see, e. g., Baggaley 1986; see also the contribution of Tedesco and Ivory to this volume) and the effects of persuasive speeches like lawyer's opening and closing statements (see Biocca et al. 1994: 17). In each case the question is, which elements of the spot, film, or speech are most persuasive.

Recently, most RTR studies on the persuasive power of communicative stimuli have been undertaken in the context of televised political debates. Studies like that have first been carried out by television broadcasters in the U.S. and other countries in order to determine the winner of the debates (see Biocca et al. 1994; Clark 2000; Ward & Walsh 2000). Most recently, TV stations provide RTR reaction of focus groups to the audience (see the contribution of

Schill and Kirk to this volume). Later, several academic studies have been carried out in order to identify highlights and defining moments of the debates (see Delli Carpini et al. 1997; McKinnon & Tedesco 1999), the influence of viewers' pre-debate opinions on their perceptions of the debate (see Jarman 2005), the effects of different rhetorical strategies (see Reinemann & Maurer 2005), the influence of viewers' perceptions during a debate on their post-debate opinions and knowledge (see Reinemann & Maurer 2005; Maurer & Reinemann 2006), and the concurring effects of verbal and nonverbal signals during a debate (see Faas & Maier 2004; see also the contributions of Maier and Maier as well as Roessing et al. to this volume). While, thus far, there is a lack in international comparative RTR studies on the persuasive effects of televised debates, this question will be adressed in the contribution of Strömbäck and Maier to this volume by comparing debates from Sweden and Germany.

As already pointed out, RTR actually has not been developed for analyzing the effects of persuasive messages but for analyzing recipients' reactions towards entertaining media content. Consequently, other academic applications are concerned with emotions caused by several kinds of entertaining media. In the tradion of the early studies on radio programs, research deals with listeners' preferences for certain kinds of music (see Brittin 1996) and viewers' feelings of being entertainend while watching talk shows (see Gunter 1995) or movies (see Sneed 1991). In this case, RTR measurement can be combinend with other methods like the retrospective think-aloud method or psycho-physiological data like heart rate or skin conductance level (see the contribution of Fahr and Fahr to this volume).

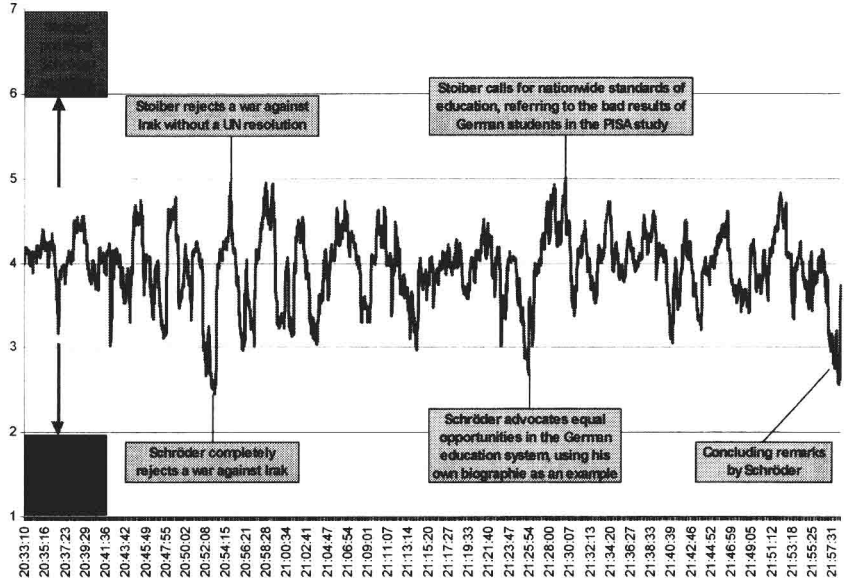
Sometimes RTR has also been used to measure recipients' attention to a stimulus. In this case, viewers are instructed to evaluate, e. g., television commercials by using RTR devices. The frequency and range of movement on the scale serves as an indicator for their attentiveness (see Thorson & Reeves 1985). Finally, RTR has been used as a tool for content or reception analysis (see, e. g., Rust 1985). In studies like that, trained coders or ordinary recipients indicate their perceptions of communicative stimuli by using dial devices. The contribution by Meyer and Ségur to this volume gives a good example of this research by analyzing viewers' perceptions of French television images on immigration.

2.3. Analyzing RTR data

RTR data can be analyzed on the individual or the aggregate level. On the aggregate level, a mean series curve for all respondents is generated. At first glance, it gives a good impression of how well viewers liked a movie, a television commercial, or the performance of a political candidate in a debate.

Additionally, the mean evaluation for all timepoints can be calculated – for example in order to compare the evaluations of different movies or commercials. The next step might be the analysis of the most significant moments in the time-series. While watching a movie, viewers will like some passages more than others. While watching a televised debate, viewers will be impressed by certain arguments more than by others. These "peaks" can be identified by analyzing the mean curve (for rules for identifying peaks see Biocca et al. 1994: 38). Figure 3 shows a peak analysis of the 2002 German general election debate. The most striking peaks are marked. Additionally, the figure shows what the discussion was about at those points in time.

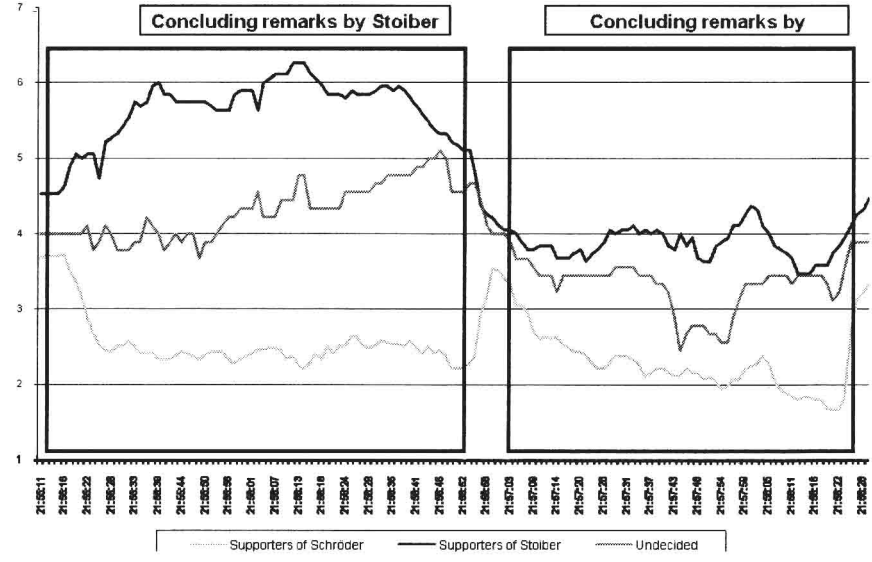
Figure 3: Peak analysis of an RTR graph (2002 German general election debate)



While descriptive analyses like that are relatively simple, it is not always easy to decide *why* viewers liked a certain part of a movie or a certain argument in a debate. In other words: Difficulties arise when peaks have to be explained. A solution to this problem is analyzing the commonalities of all peaks and looking for characteristics that separate them from the other parts of the movie or other arguments. For example, in the 2002 German general election debate almost all peaks were caused by candidates speaking so vaguely about their own future

plans that everbody could agree. In contrast, other arguments, e. g., candidates attacks were not that successful (see Reinemann & Maurer 2005). A different strategy of analysis was comparing respondents evaluations to a systematic content analysis of the stimulus, e. g., by time series analysis.

Figure 4: Analysis of group differences (2002 German general election debate)



On the individual level, the results of RTR measurement can be compared to external data like traditional surveys before and after the presentation of the stimulus as well as psycho-physiological measurement conducted during the presentation of the stimulus. Survey data gathered *before* an RTR study can be used to analyze group differences in the reception or evaluation of the stimulus. In the context of movie or advertisement research it might be important to know whether a movie or TV spot is evaluated differently by men and women or younger and older viewers. In the context of debate research it is especially important to know whether the perception and evaluation of candidates' arguments depends on viewers' pre-debate party affiliations. Figure 4 shows an analysis of differences in the perceptions of the concluding statements of both candidates in the 2002 German general election debate between supporters of the two different candidates and undecided voters. It shows that short-term perceptions during the debate are heavily influenced by pre-debate opinions (see Reinemann & Maurer 2005).

Survey data gathered directly or some days *after* a RTR study can be used to examine the relationship of short-term perceptions during media reception and opinion formation after media reception. As already pointed out, most theories on opinion formation explicitly or implicitly assume that post communication opinions are the result of several impressions during being exposed to communication. For example, viewers' perceptions of the candidates during a debate (individual means of second-by-second evaluations measured by RTR) should sum up to a final opinion after the debate (individual opinions measured by a traditional questionnaire). This model of opinion formation has been proven to be right by several studies on the 2002 German general election debate (see Reinemann & Maurer 2005; Maurer & Reinemann 2006; Maier et al. 2007).

Moreover, this research allows for analysis of primacy and recency effects in opinion formation, e. g., by comparing the correlations between individual RTR means and post-debate opinions when only the beginning, only the end or the whole debate is taken into account. Applying the same logic, it can also be analyzed whether crucial moments of a debate ("peaks") have a special impact on viewers' post-debate opinions. The only study on that problem, so far, suggests that this is not the case. Rather, viewers' post-debate opinions can be best explained by summing up every single impression during a debate (see Reinemann & Maurer 2005). Finally, using RTR and psycho-psychological data like heart rate or skin conductance level simultaneously *during* a debate, gives a better insight in viewers' emotional responses to entertaining media content (see the contribution of Fahr to this volume). Generally, the combination of RTR and external data can also be used in order to prove the validity of RTR measurement. This will be further explained in the following chapter (see 3.2.).

3. Methodological issues

3.1. Is RTR reliable?

Reliability concerns the question whether repeated measures of the same construct lead to the same results. Generally, the reliability of measurements can be tested in three ways: test-retest designs, split-half designs, and parallel-test designs. However, measuring test-retest reliability may be problematic when RTR is concerned. As the goal of RTR is to measure spontaneous reactions, in some cases particpants will react differently to a second presentation of the same stimulus. For example, watching a movie for the second time might not evoke the same reactions as far as the feeling of suspense or the attention towards certain characters is concerned. Besides that, a study by Fenwick and Rice (1991) on advertising evaluations found quite high levels of test-retest reliability for RTR measurements. More often, split-half designs are used to assess RTR

reliability. Participants are randomly assigned to two different groups and their perceptions are then compared afterwards. Doing so, early studies (Schwerin 1940; Hallonquist & Suchman 1944; Hallonquist & Peatman 1947) obtained reliablity scores between .80 and .99.

Finally, reliability can be measured by parallel-test designs. In the case of RTR, this has been done in a recent study by Maier et al. (2007). By comparing the results of two RTR studies on the same televised debate, using different sets of measurement devices (bush buttons vs. dial devices) and different instructions authors still found high correlations between the perceptions of the two groups of participants. This was especially the case during the crucial passages of the debate, when RTR curves in both studies moved exactly parallel. The coefficients of reliability are, of course, lower than coefficients that stem from split-half-designs which employ the *same* instrument. Taken together, the reliability of RTR measurement has been examined only in a few studies but has been proven to be quite high in all of them. Nevertheless, more research on the reliability of RTR measurements is needed.

3.2. Is RTR valid?

Concerning the validity of measurements, two different aspects have to be distinguished: external/ecological and internal/content valitity. *External validity* concerns the question whether the results of studies using experimental designs can be generalized to natural settings. In the case of RTR, this might be especially problematic because participants have to run control units during the reception process which might distract their attention from the media stimulus. Furthermore, due to economical or practical reasons the stimulus is often presented to a group of test subjects. This might lead to participants influencing each other, e.g., if subjects comment loudly on the stimulus or if switching the control units happen to be noisy. These problems have been discussed since the early studies using RTR (see Hallonquist & Suchman 1944; Hallonquist & Peatman 1947). These discussion revealed at least some suggestions for technical solutions: training test persons how to use the control units until they are so familiar with the use of the dials that they are able to concentrate on the stimulus; telling test subjects not to comment on the media stimulus or to talk to each other during the experiment; adjusting loudness of the stimulus in order to drown out the noise of switching the control units etc. Nevertheless, there is no empirical study on the external validity of RTR measurement, so far. This gap in research will be closed by the contribution of Reinemann and Maurer to this volume.

Internal validity concerns the question whether RTR really measures what it is supposed to measure. In the case of RTR, this might be especially problematic

because due to technical reasons and the limited capacity of human information processing, RTR studies can only measure one single dimension. Consequently, the question is whether this is sufficient when complex processes like viewers' information processing are under examination. The crucial decisions that have to be made, then, are the selection and definition of the dimension that is to be measured and, furthermore, the verbalisation of the instructions that are given to participants. Definitions and instructions need to be precise enough so that the participants know what is expected from them. At the same time instructions must leave some degree of freedom in order to measure recipients individual reactions to the stimulus. Therefore, the question whether RTR measurement yields valid results can only be answered study-by-study.

In their recent comparison of two RTR studies on a televised debate in the 2002 German national election, Maier et al. (2007) also analyzed different aspects of internal validity. *Construct validity* is given when the results of the measurement correlate with other variables in ways that one would expect. In this case, authors found that supporters of a candidate perceived his performance much more positive than supporters of the other candidate did. *Criterion or prognostic validity* is given when the results of the measurement correlate with a manifest external criterion in a way that one would expect. Here, authors found that respondents individual impressions during the debate strongly predicted their post-debate opinions (see also 2.3.). This held true for both studies despite the fact that they used different intructions. These results at least show that RTR measurement *can* be valid – in case proper instructions are used. Again, more research is needed in order to find proper instructions for at least the most common fields of application of RTR measurement.

4. A final look at the importance of RTR measurement

Despite not being a new research tool, RTR measurement has not often been used in social science research, so far. This might be the case for several reasons: Working with RTR is quite expensive, gathering and interpreting data is not always easy, there still are some open questions concerning reliability and validity. On the other hand, RTR is a unique research technique for analyzing respondents' immediate impressions during the reception of communicative stimuli. By combining RTR and content analysis data respondents' impressions can be traced back to single aspects of the stimulus, e. g., arguments, gestures, or certain dramatic elements. Moreover, it enables social scientists to prove some of their central models and theories, e. g., models of information processing and opinion formation.

Because RTR has rarely been used in social sciences, there are still open questions. This concerns fields of application as well as methodological issues.

To answer some of those questions is the aim of this volume. In its first part, some of the most common and some new fields of application are introduced. In its second part, methodological issues are discussed. While this volume might not provide answers to all open questions, it might be a starting point for further research. We believe that RTR measurement is an extremely helpful and important research tool for the social sciences. Therefore, we hope that this volume will encourage researchers to work with RTR in order to further improve the tool and the quality of the gathered data.

References

Baggaley, J. (1986). Formative evaluation of educational television. *Canadian Journal of Educational Communication* 15, 29-43.

Baggaley, J. (1987). Continual response measurement. Design and validation. *Canadian Journal of Educational Communication* 16, 217-238.

Biocca, F., David, P. & West, M. (1994). Continuous response measurement (CRM). A computerized tool for research on the cognitive processing of communication messages. In Lang, A. (Ed.), *Measuring psychological responses to media*. Hillsdale, 15-64.

Brittin, R.V. (1996). Listener's preference for music of other cultures. Comparing response modes. *Journal of Research in Music Education* 44, 328-340.

Clark, H. (2000). The worm that turned. New Zealand's 1996 general election and the televised "Worm" debates. In Coleman, S. (Ed.), *Televised election debates. International perspectives*. New York, 122-129.

Delli, C., Michael, X., Keeter, S. & Webb, S. (1997). The impact of presidential debates. In Norris, P. (Ed.), *Politics and the press.* Boulder, 145-164.

Faas, T. & Maier, J. (2004). Schröders Stimme, Stoibers Lächeln. Wahrnehmungen von Gerhard Schröder und Edmund Stoiber bei Sehern und Hörern der Fernsehdebatten im Vorfeld der Bundestagswahl 2002. In Knieper, T. & Müller, M. G. (Eds.), *Visuelle Wahlkampfkommunikation.* Köln, 186-209.

Fenwick, I. & Rice, M.D. (1991). Reliability of continuous measurement copy-testing methods. *Journal of Advertising Research* 31, 23-29.

Gitlin, T. (1994). *Inside prime time*. New York.

Gunter, B. (1995) Understanding the appeal of TV game shows. *Medienpsychologie* 7, 87-106.

Hallonquist, T. & Peatman, J.G. (1947). Diagnosting your radio program or the program analyzer at work. In Institute for Education by Radio (Ed.), *Education on the air. Yearbook of the Institute for Education by Radio, 1947*. Columbus, 463-474.