gesis
Leibniz-Institut
für Sozialwissenschaften

Working *Papers* 2013|27

Reliability and Validity of RTR Measurement Device

Georgios Papastefanou

GESIS-Working Papers 2013|27

Reliability and Validity of RTR Measurement Device

Georgios Papastefanou

GESIS – Leibniz-Institut für Sozialwissenschaften 2013

GESIS-Working Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 553
Telefax: (0621) 1246 - 100
E-Mail: georgios.papastefanou@gesis.org

ISSN: 1869-0483 (Print)
ISSN: 1869-0491 (Online)
Herausgeber,
Druck und Vertrieb: GESIS - Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

1 Introduction

Real-time response measurement was first used by Paul F. Lazarsfeld and Frank N. Stanton in the late 1930s (Lazarsfeld/Stanton 1944). Lazarsfeld, who- already in the 1920s - tried to capture feelings while listening to music (Levy 1982, p. 31), was interested in measuring the emotional, i.e. automatic (spontaneous) and affective responses towards media stimuli. Together with Frank Stanton he developed and used a technical device to capture people's momentary evaluations, called the Lazarsfeld-Stanton Program Analyzer, which then - under CBS control – was used with no change until end of 60s as a rating system for governmental, educational and commercial/advertising purposes (Biocca et al. 1994, p. 35, see also Millard 1992, Levy 1982). With a boom in the early 80s this kind of a RTR device underwent technological improvements and differentiation as a feasible device to monitor and report cognitive and affective responses to messages (Biocca et al. 1994, p. 26).

While the RTR method in US is regularly used in presidential elections TV aired debates, a new interest in this method grew up in the last decade also in Germany on occasion of introducing TV debate in German election campaigns. Several studies were conducted in Germany on occasion of TV debates in two consecutive national election campaigns. From this background Maier et al. (2007) propagated real-time response (RTR) measurement as a valuable tool complementing survey based political attitudes research. They ask for RTR method to "become part of the standard toolkit to analyze televised debates as well as other media stimuli because it can provide unique insights into the micro-dynamics of media effects" (Maier et al. 2007, p. 70). In a recent collection of papers Maier et al. (2009) underline the aspiration on real-time response (RTR) measurement as a method for the social sciences.

According Biocca et al. (1994, p. 21) RTR measurement is a kind of digital version of a pencil-paper questionnaire. A TV debate then is a kind of a stream of stimuli, which people are asked to evaluate. The stimuli to respond to are defined ex ante in the context of a research question. Specific aspects like emotional, behavioral, rhetoric, general or specific features of attitudinal objects can be taken into focus as stimuli, which respondents are asked to rate in real-time.

Actually respondents introspect their reactions to message stimuli and then report perceived changes in their mental states by using a scale as is known in standard attitude measurement. In contrast to Lazarsfeld's approach of using a binary positive/negative rating scale, research tended to use even more high resolution with 5, 7 or 10 point interval scales. Biocca et al. (1994, p. 20) show, that scale granularity is limited only by the research question and by respondents' ability to make the requested discrimination in real time. In some studies devices even 100 points scale are used, but there seem to be a tradeoff between scale resolution and cognitive burden.

Usually, respondents are asked to use a technical device for response reporting. According Maurer/Reinemann (2009, p. 5) most RTR studies use dial devices, with metric scales (1-7) or sliders (0-10). Alternatively push button devices can be used for capturing positive or negative evaluation responses in real-time (Maier/ Faas 2003, p. 25). For a discussion of advantages and disadvantages of dials versus push button devices see Maier/Faas (2003). The authors see some doubts on the dial technique. They wonder about which pieces of information exactly were processed, as well as point to a possible bias by the tendency to not reset recurrently the dialing device. In sum they favour push button technique, which is said to provide researcher with better data (Maier/Faas 2003, p. 25).

Like for any data collection instrument, reliability and validity are crucial for RTR measurement.

Maurer/Reinemann (2009, p. 9) differentiate between test-retest reliability, split-half and parallel-test reliability. They see test-retest reliability as problematic when measuring spontaneous responses, because – as they say – in a retest the stimulus is not new any more. But this is true for any retesting procedure, even if you retest responses to a single item of a questionnaire. Nevertheless only few studies report test-retest reliabilities of real-time response measure. Fenwick/Rice (1991) found quite high test-retest correlations scores, but quantitative indicators are not reported. Based on one week delay of measurement repetition, Boyd/Hughes (1992) found a diverse pattern of test-retest correlations, depending on the type of real-time response (feeling vs. usefulness responses) and type of media stimulus streams. They report an average test-retest correlation of 0.53, which they see as "low score" (Boyd/Hughes 1992, p. 653). In their study participants watched the same media two times, with different response tasks. When viewing the video clips first time they were asked to report their affective responses (feelings), then they had to watch the video again and to report cognitive-evaluative responses (usefulness of the video).

Reliability can also be understood as a measure of internal consistency, basically calculated by split-half measures, either by splitting the sample of respondents or the sample of questionnaire items. A standard measure is Cronbach's alpha, which denotes the stability of responses over persons or over items. It seems meaningful to calculate this measure for real-time responses, as it provides information on how consistent people are using the device while watching the stimulus stream. In the very first of RTR studies high internal consistency of real time reporting response was reported. Hallonquist/Suchman (1944, p. 334) document split-half correlation coefficients between 0.95 and 0.99, Hallonquist/Peatman (1947) found correlations of 0.80 and 0.94, Schwerin (1940) reports coefficients of 0.89 and 0.93.

In a recent study, Maier et al. (2007) do not report internal consistency score, even if their data could be easily used for that. Instead Maier et al. (2007) examined *parallel-test-reliability*. In their analysis they computed for each point in time of the debate (using a one-second-interval time scale) for each study separately the z-scores of a binary response in favor of one candidate respectively against the other candidate. Finally they correlated the resulting z-score time series of the studies and found a significant overall correlation coefficient of 0.38. For different passages of the debate they found time series correlations ranging from 0.46 to 0.69 (Maier et al. 2007, p. 63).

Admitting, that the overall correlation of the two different time series curves was lower than what was usually reported by split-half designs, they did not conclude, that RTR measurements yield reliable results (Maier et al. 2007, p. 63). It seems also remarkable, that Maurer/Reinemann see these correlation as "large correlations" (Maurer/Reinemann 2009, p. 10). This conclusion seems inadequate, given that in standard literature internal consistency coefficient alphas around 0.60 are evaluated as low to moderate (Gray/Watson 2007, p. 175).

In sum, knowledge on reliability of real-time response is scattered. While internal consistency measures show very high reliability of devices, the few studies, which provide test-retest results, show low reliabilities. Comparison between different instruments (being comprised of differences in device and instructions) are also of low to medium reliability. Overall, it seems that devices are used as response expression facilities in a consistent manner. Being a basic feature in applying these devices as means of response collection, internal consistency scores like Cronbach's alpha should be reported also in more recent RTR studies.

When judging **validity** of RTR data, one has to differentiate between external vs. internal validity (Shadish et al. 2002).

Internal validity of RTR data is examined by its correlation with other variables, either as a dependent (discriminative validity) or as independent (predictive or criterion validity) variable. Boyd/Hughes

(1992) report a of 0.69 correlation between RTR score and attitude measures. In a comparative study on TV debates in Germany (2005) and Sweden (2006) Maier/Strömbäck (2009) found similar discriminative and predictive validity for RTR responses in the Swedish and the German sample. In a study on TV debate between two opponent party leaders, Maier et al. (2007) report a significant effect of party identification on a balanced RTR score and a significant effect of the balanced RTR score on the post-debate overall evaluation of who won the debate. Hence, like in the study of Maier et al. (2007) one can expect that viewers of a debate evaluate the performances of the main contenders very differently depending on their party identification.

But, the genuine contribution of the real-time responses in predicting the perceived winner judgment is different for the two studies, which they analyzed separately. The effect of RTR balanced score on post-debate verdict is significant only in one study, whereas in the other study the effect of party identification on debate winner rating gets non-significant after RTR score is added to the regression equation (see Maier et al. (2007), p. 68, Table 4).

In sum, available studies show, that RTR data can be seen as valid measures of attitudinal dispositions towards political candidates or commercial advertisement objects. But internal validity imbalance between study 1 and study 2 (in one study a very high explained variance, while in the other study no predictive effect of RTR on winner candidate perception), might be a hint to strong effects of situational setting. Obviously, RTR studies utilized a specific setting for collecting individual responses to audiovisual media, namely a setting of public audience.

Asking for **external validity** means to ask whether the results can be generalized to natural settings (Brewer 2000, p. 12), settings which usually are normal or typical for responses to media in real-time in real-world. In fact it means to look on how people usually consume media like TV debates, when they are not researched. As Maurer/Reinemann (2009, p. 10) state, there is no empirical study on the external validity of RTR measurement. Two recent studies (Reinemann/Maurer 2009, Fahr/Fahr 2009) do explicitly try to examine external validity of RTR measurement, but they are focusing on reactivity effects of the RTR measurement approach. They do not discuss possible bias by a given public audience setting. From this background there seems to be some doubt on the ecological validity of traditional methods for collecting RTR data.

The audience setting issue was aware of in the very first RTR studies of Lazarsfeld. In his RTR studies respondents did not stay in the same room, but in different rooms.

This setting was not used in later studies as respondents were conceived as an audience in the meaning of a social category, not in the meaning of low interaction social group of co-acting people. In following decades RTR measurement research seemed to follow mainly the social group concept of audience. Merton, naming the real-time response measurement procedure as continuous responses measurement (Merton et al. 1990, p. 27), used it as a memory and attitudinal probe in focus groups. Millard used it as a measure of audience "attention (Millard 1992). Biocca et al. (1994) saw RTR measurement as the rating of stimuli presented to an audience, where the "audience members were allowed to continuously signal changes in some mental state using an interval scale." (Biocca et al. 1994, p. 19).

Asking respondents to report real-time ratings in an audience setting, seemingly has become a standard procedure of RTR studies. All studies, which we found for conducting RTR measurement, used the audience setting (see table 1).

Table 1: Location settings in studies with focus on RTR measurement

Study	Page	Setting location
Maier/Faas, 2003a	3f	University campus (presumably) Bamberg
Fahr/Fahr, 2009	49ff	Research lab
Tedesco/Ivory, 2009	181f	Big TV screen
Fahr, 2006	4	No information
Fahr/Früh, 2007	5	No information
Fahr/Hoffmann, 2007	6	No information
Hughes, 1992	1f	No information
Kaid, 2009	141ff	No information
Neuhoff, 2009		No information
Reinemann/Maurer, 2009	33f	Lab with big TV screen
Jackob, 2008	219ff	Lab, university campus
Maier/Faas, 2003b	4	PC pool, university campus
Schill/Kirk, 2009	158ff	Studio
Biocca/West, 1994	56	Theatre
Lazarsfeld/Stanton, 1944	274f	Recording studio
Lazarsfeld/Stanton, 1944	279f	Recording studio
Lazarsfeld/Stanton, 1944	282f	Recording studio
Lazarsfeld/Stanton, 1944	267ff	Recording studio, participants did not see each other
Maier et al, 2005	60f	Auditorium, university campus, big screen
Neuhoff, 2009		University campus, big screen
Maier/Strömbäck, 2009	99	Universität Jena, University of Lundsval
Maier et al, 2006	2ff	University campus, big screen
Mauer et all, 2007	21f	University campus, big screen
Maier/Maier, 2007	333f	University
Maier/Maier, 2009	72f	University campus, TV Studio, big TV screen
Maier et al, 2005	61f	University Mainz, Großleinwand
Meyer/Ségur, 2009	193ff	University campus

This is true also for the very latest RTR studies on occasion national parliament election in 2009 (Rattinger et al. 2011), in which participants were recruited to watch the TV debate in an audience room at university.

While location varied, with some studies gathered people in a lab, while others chose an auditorium, respondents usually are gathered as low-interaction group in a room of a public location, getting an audiovisual presentation.

So, obviously audience setting seems to have got established as a standard method of RTR measurement. Only recently some research is trying to estimate bias, which is given by traditional settings of RTR measurement (Reinemann/Maurer 2009, Fahr/Fahr 2009). But also these studies use the standard audience group setting for measuring real-time responses to media like TV debate or advertisement.

Overall, there are several reasons, which cast some doubt on external validity of usual RTR measurement.

First, when people are recruited for a RTR study, this means a self-selection of those, which are highly interested in giving feedback on their ratings to a research public. In advance, participants get a self-definition as a special group of rating subjects many days before the TV debate is aired. So their attention to relevant information on the debate might be stronger than average people without the task orientation.

Second, people are gathered together in a special presentation location, usually an auditorium or a lab on the university campus is chose. By this they enter a special socio-spatial, academic context, which aligns their focus on their special rating task. This task identity might be fostered, when they - before the debate begins - go through an elaborate process of instructions, on how to use the device and how the stimuli are defined. It seems to be a truism, that when people watch a TV debate in real-life, they do this usually in the living room using their home TV set. Watching TV in modern society usually takes place alone or together with some friend or relative at home.

Third, as the rating task is achieved while are others being present, respondents' rating behaviour inevitably will be biased by social facilitation processes. Since Zajonc's study (1965), effects of mere presence of others on individual cognition and behaviour are well-researched: in the presence of others (e.g. as a transient audience) people tend to perform better with what they are good and to perform worse with what they are generally unsecure. These processes are effective, even if respondents are told not to communicate with each other and not to produce disturbance noise, as it is proposed for controlling social bias (Reinemann/Maurer 2009).

From the background of social facilitation research it seems reasonable to assume, that people will rate politicians' debate behaviour more pronounced, if their attitude towards the politician is well-established. People's' arousal level is increased involuntarily in presence of others (Zajonc 1980, 2000). As this contributes to increased salience of their general and candidate specific attitude, ratings in an audience setting might be more explicitly articulated than in a private situation. For people, who do not have well-established attitudes, resp. specific political party identification, increasing arousal in audience setting will lead them to report more frequent and inconsistent ratings.

In sum, it seems reasonable to assume that when RTR data are collected in audience settings, as most RTR studies do, social facilitation processes cast doubt on the external validity of RTR data. But RTR measurement technology, as Biocca et al. (1994) pointed out, is not necessarily restricted to the lab, or to auditorium. They asked for "systems that are portable, to meet subjects at any location, or to connect input devices to computer telephonically, enabling measurement of communication behaviour in more natural settings (Biocca et al. 1994, p. 19).

Meanwhile, dynamic measurement in natural settings has got available by utilizing wearable computer technology, as it is documented in a growing body of ambulatory assessment research e.g. on experience sampling in natural settings (Ebner-Priemer/Kubiak 2010) Originally most research was focused on psychiatric and clinical research, but ambulatory assessment methodology finds many applications in natural settings like family living, work life and urban experience. Recent studies show, that special designs of wearable computing can be a valuable supplement to survey based social research (Papastefanou 2008, 2009).

Being designed for unobtrusive, long-term capture of peripheral physiological parameters, wearable computing devices can be also equipped with an interface connecting a keypad for recording momentary subjective responses, e.g. by pushing tagged buttons. These kind of ubiquitous computing device makes possible to run RTR measurement in natural settings, e.g. watching TV debate at home. To see if RTR measurement is feasible and adequate in real-world settings, three studies were run using a prototypical ambulatory RTR measurement device. In following we report results on reliability and validity of these RTR data on occasion of TV debates in parliament elections campaigns in Germany in 2008 and 2009.

2 Three empirical studies with ambulatory RTR measurement

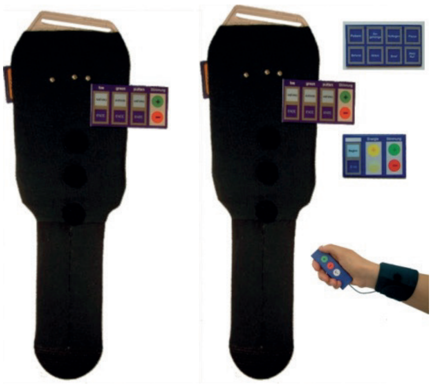
2.1 General considerations on measuring momentary responses

The studies, which are reported below, were conducted to test a method for measuring momentary responses to media in situ, at the location in which people usually view advertisement or political debate in TV, namely at home in their living room. Conducting a study with RTR means to reach several methodological decisions on: device, stimulus definition, response mode, response scale, study design. As these aspects are general and common for all of our three studies, we will discuss them in advance, before details of the studies are described.

Wearable computing response device: The guiding question of this research is aimed at testing a way of capturing momentary responses to media, when people watch TV situation at home. This everyday process of intermittently having a response to the TV debate should be as little as possible be affected by the measurement method. This would mean to put as low as possible effort to the respondents, and to keep this situation as much as possible in its common process watching the show at home in the living room.

Wearable computing seemed to be an adequate solution, as it works unobtrusively. Lab device would not be feasible as they require special computing systems infrastructure. So, a sensor wristband, which initially had been developed for capturing physiological data in comfortable and unobtrusive way, was extended by a small sized push button tablet, which fits neatly in the palm of the hand (see figure 1).

Figure 1: Sensor-wristband with multiple marker facility



Source: www.bodymonitor.de

Response mode: To keep the normal watch TV situation as natural as possible participants were asked to indicate their liking respectively disliking of the statements made in the course of the televised debate, whenever they wanted to. They were told that they could push the green or red button whenever they see as good or as bad what participants in the show said. This procedure does not restrict RTR to two candidates' debate, but is open for multiple person debates, as it was the case in post-election debate where the party leader joined the show.

A direct and explicit link between individual statement of political debater and response button was not installed. Participants were briefed, to push the red respectively the green button for a negative or positive evaluation of the momentarily active debate participant. This could be the politicians or moderator.

Respondents were told, that assignment of their response to the respective actor in the TV debate, would be achieved later by means of time synchronization of responses and talk activity of debate participants. In TV debates as in many talk shows, there seem to be a certain director's life cut principle to show for some seconds on screen the participants of the debate, who are addressed by the talking candidate. This procedure makes responding very easy and open for momentary responses, but it might be a source of error dampening the validity of real-time responses. We conceive this as a question, which can be answered empirically by reporting the duration of cross-over flashings as well as a factor in predictive and discriminative validity.

Response scale: The push button tool had two buttons, a green and a red one. Respondents were told to push the green button to indicate their approval, or the red button to indicate disapproval. Further, they were advised of the possibility to show intensity of approval respectively disapproval, by pushing the buttons as long as strong they wanted to.

To keep the task of momentary evaluative responding to a continuously changing stream of stimuli mostly close to usual watching TV we decided for a binary response scale, as Lazarsfeld and Stanton did with the Program analyzer. Compared to ratio or interval response scale seemingly this leads to a loss of information. But in fact in relation to what is intended to measure, namely an attitudinal disposition toward a politician's behavior, very rich data are collected even with a binary response scale. Counting negative and positive responses over the duration of specific statements, e.g. to selected issues, provides a ratio-scaled score of respondents real-time expressed attitude towards this attitudinal object over time.

Nevertheless, a possibility to use a more refined scale for rating the momentary favorability of a political candidate's talking was included. In study 1 the response tablet allowed for inputting different push pressure, as indicating differential positive or negative rating (this would be a ratio-scale response scale). In study 2 and 3 respondents were told, that they could use the push buttons to scale their response by the consecutive frequency of pushing the red or the green button.

Stimulus: Watching a TV debate means that respondents are exposed not to a single stimulus, but to a stream of audio and visual stimuli, as they are expressed voluntarily or involuntarily by the participants with verbal, facial, gestures and posture expressions. Definition of the stimuli to which respondents asked to provide evaluative responses, depends on research question.

In real-world life political TV debates are framed by information programs, by which people can learn more on politicians' positions and arguments. As Jackob et al. (2008) have shown, speech content is the central for rhetoric presentations. It seems adequate and necessary to define response stimulus close to what is framed in real-world situation of watching a political debate. Asking respondents to rate their general impression of the candidates seemed to be inadequate, because it expects counter-normative behavior by the respondents. This would define the measurement setting as an extraordinary evaluation task, which would be in contrast to liking and disliking responses while watching and listening to political candidate statements in a TV talk show.

Study design: For validation reasons external information on the political attitudes are needed. Pre-post design was used, by which discriminative information were collected before viewing the media by a short questionnaire. Criterion data were gathered by a questionnaire on the overall rating of who was the winner of the debate respectively on the politicians' performance in the debate. The self-administered questionnaire was to be filled in right after the end of the TV debate.

2.2 Studies description

Study 1

Media (stimulus): A TV debate on occasion of the Bavarian country parliament election in September 2008. A 45 minutes debate between the candidate of social democratic party Franz Maget, (which was the leader of the main opposition party) and the candidate of CSU Günter Beckstein, who was the Bavarian prime minister in that legislative period, was organized in September 2008 by the Bavarian regional broadcasting network and moderated by its chief editor. The televised debate was aired live on the 18th of September (the election took place on September 28th) by the Bavarian network 'Bayerisches Fernsehen' at 8.15 pm.

Sample: 21 male and 14 female participants were recruited being entitled to vote in Bavaria. A second group consisted of non-elective people residing outside of Bavaria, in cities of Mannheim respectively Ludwigshafen[4] .The sample was drawn by using academic personal and internet networks.

Device (response mode and scale): The device consisted of an electronic board housed in a textile elastic wristband, which was designed for being self-administered. The device captured also accelerometric and physiological data like skin resistance and skin temperature. Additionally it was equipped by a palm pad with two soft buttons, whose pressing would produce a voltage signal according the force of pressure. The voltage signal was captured and digitized at 100 Hz. One button was red and the other one was green. In relation to the push buttons participants were notified to push a button whenever they wanted to mark a moment of the candidates contribution as positive (by pushing the green button) or as negative (by pushing the red button). They were also told, that they could indicate level of rating by strength of pushing a button (for briefing see Appendix 1 and 2).

Setting: Participants viewed TV debate live at home, usually in their living room.

Design and Procedure: Participants got study materials (instructions, questionnaires and device) 1-2 days before the TV was aired life, either personally or by postal service. By personal contact (either face-to-face or by telephone) they were notified to follow instructions carefully. Accordingly they were to affix the wristband at the left hand 15 min before begin of the debate, then filling in the first questionnaire. Right after having watched the live aired TV debate at home, they were to fill in the post-debate questionnaire (see Appendix). Further they were asked to send back questionnaires and device by postal services using a pre-paid addressed parcel envelope.

Study 2

Media (stimulus): The media providing the stream of stimuli was a TV debate between the party leader of the conservative party of CDU, and its leader, chancellor (Angela Merkel) and social democratic party candidate for chancellor (Frank-Walter Steinmeier, who was also foreign secretary and vice chancellor in preceding legislation period government). The TV debate was organized and aired live on September 13 2009 by the four largest national broadcasting companies. The TV debate was moderated by four TV journalists, each representing one of the broadcasting companies. The debate started at 20.30 o'clock and lasted for one hour and a half.

Sample: Pupils of two secondary education schools (Gymnasium) were recruited as participants, being aged above 18 and entitled to vote. Pupils of two classes of one school A, and a group of pupils in another school B were asked for participation in an informal project presentation at school. In school A all participants belong to one class. Pupils from school B were drawn from several classes. In both schools research project was introduced by an audience presentation. Pupils in both schools were in their last school year before graduating and participated special courses in political and social science.

Device (response mode and scale): The same device was used as was applied in study 1.

Setting: Participants viewed TV debate live at home TV.

Design: Pre-debate questionnaires were filled in by self-administered forms, after pupils have got instructions and device at the school meeting point. All participants were asked by written instructions and by oral communication to fill in the post debate questionnaire immediately after the TV debate ended.

Study 3

Media (stimulus): The TV show was also administrated in the context of the national election campaign 2009, organized by the national publicly funded broadcasting networks (ZDF and ARD). This TV debate was aired in the evening of the polling day, two hours after election were closed. Participants were leaders of the parties, which on the basis of the preliminary election results were represented in the national parliament. So, participants were Angela Merkel (designated chancellor), Frank-Walter Steinmeier (designated leader of the social democratic party), Guido Westerwelle (leader of the liberal party), Oskar Lafontaine (leader of the leftist party), Jürgen Trittin (candidate of the ecological party) and Peter Ramsauer (leader of the conservative regional Bavarian party, which acts nationally in union with the CDU). Peter Ramsauer was not physically present in the studio, but participated via video conferencing. The debate was moderated by two journalists, chief editors of the two organizing broadcasting networks.

Sample: pupils as participants school C participated, as well as some of those in school A, who participated in study 2.

Device (response mode and scale): The same device was used as was applied in study 1.

Setting: Participants viewed TV debate live at home TV.

Design and Procedure: Instructions, questionnaires and device were handed over personally at school either by the project leader, by teacher or by school secretary. Pre-debate questionnaires were filled in self-administered when pupils got the instructions and device at the school meeting point before they got study materials. All participants were asked by written instructions and by oral communication to fill in the post-debate questionnaire immediately after the TV debate ended. Filled-in questionnaires were returned via schools' secretary.

2.3 Data

The final samples show these descriptive measures.

Table 2: Descriptive measures of the samples

	Sample 1	Sample 2	Sample 3
Male (%)	58	35	46
Age (mean, sd)	40 (13.8)	18.7 (0.5)	18.8 (0.7)
N of positive markers of candidate A	16.1	44.6	27.7
N of negative markers of candidate A	10.9	21.9	48.8
N of positive markers of candidate B	19.5	63.5	61.9
N of negative markers of candidate B	7.7	15.6	17.8
Balance score candidate A	5.1	22.7	-21.1
Balance score candidate B	11.8	47.9	44.1
Candidate A preferred	24	33	33
Candidate B preferred	21	39	39
No candidate preferred	55	28	28
Candidate A's party preferred	9	22	11
Candidate B's party preferred	27	22	39
Any other party preferred	40	45	50
No party preference/undecided	24	11	0
N of observations	33	18	18

The resulting data sets consist of two types of data, namely of the cross-sectional survey data collected by the questionnaire before and after the TV debate and of the real-time captured physiological and real-time response data. The recording rate of the sensor-wristband was 100 Hz. These individual signal series data were collapsed afterwards into a second-by-second data format by averaging the parameter values of 100 observations per second. Individual time series data on having pushed a positive respectively negative button were differentiated according their time-based location relative to the statement turns of the debate actors. Positive and negative responses for each of the candidates were calculated, without taking response intensity into account. Timing of a response was defined as the beginning of response.

3 Results

Internal consistency

For calculating reliability of RTR data we followed the procedure of Hollonquist/Suchman (1944). With RTR data a longitudinal measurement design with regular measurement repetition (according the recording rate of the computing device) respectively with event-dependent measurement repetition (pushing the buttons depending on approved respectively disapproved events) is given. So this provides adequate data for calculating Cronbach's alpha to indicate test-retest reliability by. To keep computations modest, we did not use the second-by-second repetition, but collapsed the response data to 1 minute intervals over the 45 min of the debate show. So we've got 45 quasi-items, each of which comprised the number of positive resp. negative evaluation responses. It's this set of items, whose internal consistency – separately or positive and negative responses – is calculated by Cronbach's alpha. As we see in see table 2, Cronbach's alpha generally is above 0.9, which indicates a high internal consistency of the responses over the 45 minute intervals.

Table 3: Cronbach's alpha for one-minute response scores over 45 min of TV debate

Response scores	Beckstein - Maget	TV Duell	Berliner Runde
Positive ratings Merkel		0.9819	0.6916
Negative ratings Merkel		0.9616	0.8139
Positive ratings Steinmeier		0.9542	0.9315
Negative ratings Steinmeier		0.9136	0.6785
Overall positive ratings	0.92	0.9850	0.8393
Overall negative ratings	0.92	0.9601	0.8643

Validity

In relation to RTR scores and party identification we followed the interpretation, that the stronger their associations are the better the construct validity of RTR measures (Maier et al. 2007, p. 66).

Party identification as an individually stable affective tie to a political party does affect perception and evaluation of political actors and political issues. As we can see in table 4 and table 5, there is a clear differentiation of negative and positive response scores ascribed to the candidates by candidate preference and party identification.

Table 4: Different RTR measurement indicators by chancellor preference

	Candidate A preferred as chancellor	t-value (Markers)**	Candidate B preferred as chancellor	t-value (Markers)**	t-value (Group)***	df	Etasquare
N of positive markers for candidate A	30.1 (45.5)*	2.3	35.5 (42.3)	-0.8	-0.4	39	0.004
N of negative markers for candidate A	9.8 (15.3)		51.4 (94.4)		-2.0	39	0.09
N of positive markers for candidate B	24.9 (34.7)	1.8	80.1 (97.3)	2.9	-2.4	39	0.13
N of negative markers for candidate B	12.4 (16.7)		19.8 (32.8)		-0.9	39	0.02
N of cases	20		21				

Notes: *figures in parentheses are standard deviations

** t-value (Markers) refers to the comparison of the mean number of markers for the candidates

*** t-value (Group) refers to the comparison of mean preference for candidate A resp. candidate B

Table 5: Differences in RTR measurement indicators by party preference

	Preferred candidate A's party	t-value (Markers)**	Preferred candidate B's party	t-value (Markers)**	t-value (Group)***	df	Etasquare
N of positive markers for candidate A	70.2 (71.7)*	2.4	23.9 (27.1)	-1.3	2.6	27	0.19
N of negative markers for candidate A	20.3 (23.7)		50.5 (97.0)		-0.9	27	0.03
N of positive markers for candidate B	57.2 (61.3)	1.8	64.8 (98.5)	2.3	0.2	27	0.002
N of negative markers for candidate B	27.7 (26.9)		14.9 (31.0)		-1.1	27	0.04
N of cases	9		20				

Notes: *figures in parentheses are standard deviations

** t-value (Markers) refers to the comparison of the mean number of markers for the candidates

*** t-value (Group) refers to the comparison of mean preference for candidate A resp. candidate B

Additionally tested criterion validity RTR measure was tested, which means to examine of the subjective response score would predict the occurrence of an external criterion. In the present context, RTR measurements would be valid if participants, who rated governmental candidate's statements better than that of oppositional candidate's turns, evaluated governmental candidate as the winner of the debate in the post-debate questionnaire. In table 6 one can see, that those who perceived governmental candidate as the debate winner, had given him more positive scores over the debate, while those, for whom governmental candidate had accumulated more negative response score, the oppositional candidate was the perceived winner of the debate. It is important to underline, that this predictive significance of real-time responses on final perception of being winner of the debate is effective independent of the antecedent preference for the candidate as chancellor.

Table 6: Odds-ratios of candidate A as winner of the TV debate, compared to candidate B as winner (logistic regression)

Reference group	Predictor	Model M	Model 1a
	No of positive markers for candidate A	1.07*	1.08*
	N of negative markers for candidate A	0.98	1.02
	No of positive markers for candidate B	0.93*	0.94*
	N of negative markers for candidate B	1.01	0.96
Candidate B preferred as chancellor/prime minister	Candidate A preferred as chancellor/prime minister		38.9*
	N of observations	69	69
	LR chi-square(6)	18.86	31.29
	Pseudo R-square	0.22	0.37

Notes: * p<0.05

4 Summary and conclusions

Starting point of this study were the findings of Maier et al. (2007), who showed that specific devices of RTR measurement provide reliable and valid data on people's judging of political candidates talking in a TV debate. Nevertheless, these results are reached under specific settings, which hardly can be accepted as representing TV debate watching in daily life setting. In the studies reported by Maier et al. (2007), RTR measurement was conducted in special, extra-daily settings, either as a kind of "public viewing", where participants are gathered in some large rooms like a university auditorium or cinema auditorium. This limits the ecological validity of measuring of political position via RTR method.

To overcome this crucial restriction, an alternative device was developed for RTR measurement in natural settings, for example at home in the living room. Targeted at being applicable in real life settings as part of standardized mailing survey procedure, a wearable computing sensor wristband was developed extended by credit card-sized keypad. The keypad had two response buttons, similar to the approach of study 1 of Maier et al. (2007) in which also two buttons allowed respondents to indicate positive and negative ratings separately. In sake of simplicity and thereby to sustain spontaneous evaluation responses free of desirability and cognitive deliberations we opted for two binary buttons. Further, participants were freed from aligning specific response buttons to specific candidate, which would tend to distract their attention. Under the assumption, that in TV debate the focus usually is directed to the speaking person, it seemed sufficient to tell people to spontaneously rate a person's statement while he making his statement.

Our results show a high reliability of measuring spontaneous evaluative responses in real-time by this keypad device in stimulus settings under normal life conditions, namely watching a TV debate at home. Based on calculations of over-time consistency, we found high Cronbach's alpha scores for subjective ratings reliability. This finding supports the perspective, that a wearable computing key pad device together with parsimonious instructions provides reliable subjective real-time revaluations.

Further, our results support Maier et al.'s (2007) validity claim. We found significant construct and criterion related validity of real-time responses, measured by our wearable key pad instrument. Global measures of response inclination as sum scores of negative and positive responses shown over the TV debate, like Maier et al. (2007) did, proved to correlate significantly with party identification as well as with subjective evaluation of who was the winner after the debate.

In sum, the general results of our study are consistent with that of study 2 of Maier et al. (2007), which proves our method of RTR measurement to be equivalent to those used by Maier et al. (2007). So, we can claim, that RTR measurement can be accomplished in a reliable and valid way by a method, which has better ecological validity as it can be applied under natural settings of viewing TV debates, namely at home in the living room.

In some detail, our findings deviate from the results, which are reported by Maier et al. (2007) for their study 2. Actually R-square of our criterion regression of the balanced response score was lower compared to the exceptional high R-square of .71 the authors reported. Also, we found the effect of the positive or negative RTR scores on the perception of the debate winner to be smaller than the effect of party identification. In the Maier et al. (2007) study a RTR balance score had quite a larger effect. These differences might be due to differences between the settings of our studies: while in our study respondents watched TV as they generally do, at home maybe with relatives and friends. In the Maier studies respondents were put in a special defined situation of rating TV together with others in the same room. This socio-cognitively focused situation might lead to more deliberately response consistency, being reflected in a higher correlation between RTR responses and political attitude before and after the TV debate. Because of the multi dimensional design differences between our and Maier's study, explanation of the different effect size of RTR scores remains indecisive. One could hypothesize, that the predictive power of RTR responses is lower under real life conditions of watching TV debate at home than under socio-cognitive focused setting of public viewing. Of course, finding differences might be due to different samples, which in both cases are prone to self-selection by recruiting procedures. So as reported in this paper, most participants had higher educational status and more pronounced party identification. This could be reflected in lower correlations between RTR and self-reported political attitude measures.

The results of our studies open up extending the scope of RTR measurement to real life conditions of watching TV debates. They do support the claim of Maier et al. (2007), that "RTR should become part of the standard toolkit to analyze televised debates as well as other media stimuli" (p. 70).

References

Biocca, F., Prabu, D., & West, M. (1994): Continuous Response Measurement (CRM). A Computerized tool for research on the cognitive processing of communication messages. In: A. Lang (Ed.): Measuring psychological responses to media, 15-64. Hillsdale, NJ: Erlbaum.

Bohner, G., & Wänke, M. (2003): Attitudes and Attitude change. Hove, East Sussex: Psychology Press.

Boyd, Th. C., & Hughes, G. D. (1992): Validating Realtime Response Measurues. *Advances in Consumer Research,* 19, 649-655.

Brewer, J. D. (2000): Ethnography. Understanding Social Research. Buckingham: Open University Press.

Ebner-Priemer, U. W., & Kubiak, T. (2010): The Decade of Behavior Re-Visited: Future Prospects for Ambulatory Assessment. *European Journal of Psychological Assessment,* 26, 151-153.

Fahr, A., & Fahr, A. (2009): Reactivity of Real-Time Response Measurement: The Influence of Employing RTR Techniques on Processing Media Content. In: Maier, J., Maier, M., Maurer, M., & Reinemann, C. (Eds.): Real-Time Response Measurement in the Social Sciences. Methodological Perspectives and Applications, 45-61. Frankfurt am Main et al.: Peter Lang.

Faas, T., & Maier, J. (2003): *Wortlaut und Wahrnehmung des zweiten Fernsehduells im Bundestagswahlkampf 2002 – eine Dokumentation* [Wording and perception of the second televised debate in the federal election campaign 2002 – a documentation]. Bamberg: Universität Bamberg (Bamberger Beiträge zur Politikwissenschaft, Nr. II-17/2003).

Fenwick, I., & Rice, M. D. (1991): Reliability of Continuous Measurement Copytesting Methods. *Journal of Advertising Research*, 31, 23-29.

Gray, E. K., & Watson, D. (2007): Assessing Positive and Negative Affect via Self-Report. In: Coan, J. A., & Allen, J. J. B. (Eds.): Handbook of Emotion Elicitation and Assessment, 171-184. Oxford: Oxford University Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998): Measuring individual differences in implicit social cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Hallonquist, T., & Suchman, E. E. (1944): Listening to the listener. Experiences with the Lazarsfeld-Stanton Program Analyzer. In: Lazarsfeld, P. F., & Stanton, F. (Eds.) (1944): Radio Research 1942-1943, 265-334. New York: Duell, Sloan and Pearce.

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

Hughes, G. D. (1990): Diagnosing Communications Problems with Continuous Measures of Subjects' Responses: Applications, Potential Applications, Limitations, and Future Research. *Current Issues & Research in Advertising*, 13, 175-197.

Hughes, G. D. (1992): Real-time response measures redefine advertising wearout, *Journal of advertising research*, May/June, 61-77.

Jackob, N., Petersen, Th., & Roessing, Th. (2008): Strukturen der Wirkung von Rhetorik. Ein Experiment zum Wirkungsverhältnis von Text, Betonung und Körpersprache. In: *Publizistik*, 53(2), 215-230.

Lazarsfeld, P. F., & Stanton, F. (Eds.) (1944): Radio Research 1942-1943, 265-334. New York: Duell, Sloan and Pearce.

Levy, M. R. (1982): The Lazarsfeld-Stanton Program Analyzer: An historical note. *Journal of Communication*, 32 (4), 30-38.

Maier, J., & Faas, T. (2003): *Wortlaut und Wahrnehmung des ersten Fernsehduells im Bundestagswahlkampf 2002 – eine Dokumentation* [Wording and perception of the first televised debate in the federal election campaign 2002 – a documentation]. Bamberg: Universität Bamberg (Bamberger Beiträge zur Politikwissenschaft, Nr. II-16/2003).

Maier, J., Maurer, M., Reinemann, T., & Faas, T. (2007): Reliability and validity of real-time response measurement: A comparison of two studies of a televised debate in Germany. *International Journal of Public Opinion Research*, 19 (1), 54-73.

Maier, J., Maier, M., Maurer, M., & Reinemann, C. (Eds.) (2009): Real-Time Response Measurement in the Social Sciences. Methodological Perspectives and Applications. Frankfurt am Main et al.: Peter Lang.

Maier, M., & Strömbäck, J. (2009): Advantages and Limitations of Comparing Audience Responses to Televised Debates: Comparative Study of Germany and Sweden. In: Maier, J., Maier, M., Maurer, M., & Reinemann, C. (Eds.) (2009): Real-Time Response Measurement in the Social Sciences. Methodological Perspectives and Applications, 97-116. Frankfurt am Main et al.: Peter Lang.

Maurer, M., & Reinemann, C. (2009): RTR measurement in the social sciences: Applications, benefits, and some open questions. In: Maier, J., Maier, M., Maurer, M., & Reinemann, C. (Eds.) (2009): Real-Time Response Measurement in the Social Sciences. Methodological Perspectives and Applications, 1-13. Frankfurt am Main et al.: Peter Lang.

Merton, R. K., Fiske, M., & Kendall, P. L. (1990): *The Focused Interview: A Manual of Problems and Procedures*. Glencoe, IL: The Free Press.

Millard, W. J. (1992): A history of handsets for direct measurement of audience response. *International Journal of Public Opinion Research*, 4 (1), 1-17.

Papastefanou, G. (2008): Ambulatorisches Assessment und Empirische Sozialforschung [Ambulatory Assessment and empirical social research]. *soFid, Methoden und Instrumente der Sozialwissenschaften*, 2008 (2), 11-20.

Papastefanou, G. (2009): Ambulatorisches Assessment: eine Methode (auch) für die Empirische Sozialforschung [Ambulatory assessment: a method for empirical social research (as well)], *Österreichischen Zeitschrift für Soziologie*, Sonderheft 9, 443-469.

Petty, R., & Cacioppo, J. T. (1986): The elaboration likelihood model of persuasion. In: L. Berkowitz (Ed.): *Advances in experimental social psychology*, 19, 123-205. New York: Academic Press.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Brettschneider, F., Faas, Th., Maier, J., & Maier, M. (2011): TV-Duel-Analyse, Real-Time-Response-Data (GLES 2009). GESIS Datenarchiv, Köln. ZA5310 Datenfile Version 1.1.0, doi:10.4232/1.10370.

Reinemann, C., & Maurer, M. (2009): Is RTR biased towards verbal message components? An experimental test of the external validity of RTR-measurements. In: Maier, J., Maier, M., Maurer, M., & Reinemann, C. (Eds.) (2009): Real-Time Response Measurement in the Social Sciences. Methodological perspectives and applications, 27-44. Frankfurt am Main et al.: Peter Lang.

Schwerin, H. (1940): An exploratory study of the reliability of the 'Program Analyzer'. *Applied Psychology*, 24, 742-745.

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

Spezio, M. L., Rangel, A., Alvarez, R. M., O'Doherty, J. P., Mattes, K., Todorov, A., Kim, H., & Adolphs, R. (2008): A neural basis for the effect of candidate appearance on election outcomes, *Social Cognitive and Affecitive Neuroscienc*e, 3, 344–352.

TIME Magazine (1942), *What Do They Like?*, Monday, June 29.

Thorson, E., & Reeves, B. (1986): Effects of over-time measures of viewer liking and activity during programs and commercials on memory for commercials. *Advances in consumer research*, 1986, 13, 5409-553.

Witte, E. H. (1987): Behavior in group situations: An integrative model. *European Journal of Social Psychology*, 4, 403-429.

Zajonc, R. B. (1965): Social Facilitation. *Scienc*e, 149 (3681), 269-274.

Zajonc, R. B. (1980): Feeling and thinking: preferences need no inferences. *American Psychologist*, 35 (2), 151-175.

Zajonc, R. B. (2000): Feeling and thinking: Closing the debate over the independence of affect. In: Forgas, J. P. (Ed.) (2000): *Feeling and Thinking: The Role of Affect in Social Cognition*, 31-58.

Footnotes

[1] I am grateful to Thomas Plischke (GESIS) for his political science expertise in discussing preliminary results of the study and Matthias Fleck and Steffen Weber for their extraordinal contribution to accomplishing this study. Further, my thanks to Prof. Hans Rattinger's (University of Mannheim and GESIS) decisive support in recruiting participants at the University of Bamberg. My thanks also to Prof. Peter Mohler's support of the GESIS pilot project on "Ambulatory Assessment in Social Research", which is the background of the present study.

[2] "The Lazarsfeld-Stanton program analyzer is a simple device. Subjects sit in comfortable chairs, hold a pair of push buttons in their hands, and listen to a program. When they like what they hear, they push the right-hand button. When they don't like it, they push the left button. Each button is electrically connected with a pen which draws a continuous line on a moving paper tape pulled under it at a constant speed of approximately one inch every five seconds. When a button is pressed, an electric magnet jogs the pen a quarter of an inch, keeps it off the apathy line until the button is released. Working from a timed script, researchers interview the subjects after the program, ask the cause for their likes & dislikes" (TIME, 1942).

[3] Further development was to enable participants to give more specific, differentiated judgments assessments and evaluations can be measured by using semantic differentials (see Biocca et al. 1994). Dial devices were developed, allowing participants to express their reactions on a metric dimension rather than only in a dichotomous way (see Millard 1992, Biocca et al. 1994). Obviously in this development the emotional aspect tended to be neglected.

[4] These cities are located in different state countries of Germany, about 260 km as the crow flies distant to Munich, the capital city of Bavaria respectively about 100-150 km to the Bavarian state border.

Appendices

APPENDIX 1 Instructions for self-administered questionnaire and ambulatory device

Please put on the sensor-wristband as described, watch the TV-debate and mark everything you like or don't like. Additionally we ask you to fill in the two questionnaires (questionnaire A and B) before and after TV-debate.

Please follow these succeeding steps:

1. put on the sensor-wristband about 5 Min. before 8 p.m. (see Instruction 1)
2. switch-on the sensor-wristband exactly at **8 p.m.** (to the split second as indicated by the Tagesschau-clock)
3. Fill in questionnaire A
4. Watch the TV-debate from 8.15 p.m. and give your marks during the broadcast (see Instruction 2)
5. Fill in questionnaire B after the TV-debate
6. Switch-off the sensor-wristband and detach it
7. Put all papers into the big envelope and send them back with the attached package.

On the next page you find the description how to put on the sensor-wristband.

APPENDIX 2 Instructions for affixing the sensor-wristband

Instruction 1: How to put on the sensor-wristband

Please wear the sensor-wristband on you **left** wrist.

To put it on follow the next steps:

APPENDIX 3 Defining the response task

Instruction 2: What you should do while watching the TV-debate

During the broadcast you should mark everything you like or don't like what the politicians (or the moderator) are saying.

Please use the push-button facility on the sensor-wristband to give your marks:

Whenever you **like** something you push the **green** button.

Whenever you **don't like** something push the **red** button.

You can use the buttons during the TV-debate as often as you want.

Whenever you especially like something or especially don't like something, you can indicate the intensity by pushing the green respectively red button correspondingly hard.

APPENDIX 4 Wording of used items

a) Who according your opinion has done best in today TV debate? - *name of governmental candidate* clearly did better, - *name of governmental candidate* did somewhat better, - *name of oppositional candidate* clearly did better, - *name of governmental candidate* did somewhat better, - none of both, - do not know

b) Overall, are you inclined towards a political party, and if yes, which one? – CDU, - CSU, - SPD, - FDP, - Bündnis 90 / Die Grünen, - Die Linke, - NPD, -another party, - no party