
Article:
Praetorius, A-K., McIntyre, N.A. and Klassen, R.M. (2017) Reactivity effects in video-based classroom research: an investigation using teacher and student questionnaires as well as teacher eye-tracking. Zeitschrift für Erziehungswissenschaft, 20 (Suppl 1). pp. 49-74. ISSN: 1434-663X

https://doi.org/10.1007/s11618-017-0729-3

Reactivity Effects in Video-Based Classroom Research: An Investigation Using Teacher and Student Questionnaires as Well as Teacher Eye-Tracking

Reaktivitätseffekte in der videobasierten Unterrichtsforschung: Eine Untersuchung mittels Lehrer- und Schülereinschätzungen sowie Lehrer-Eye-Tracking

Version 13th of August 2016, accepted for publication in Zeitschrift für Erziehungswissenschaft

Anna-Katharina Praetorius
German Institute for International Educational Research, Germany

Nora A. McIntyre
University of York, United Kingdom

Robert M. Klassen
University of York, United Kingdom

Author Note

Anna-Katharina Praetorius, German Institute for International Educational Research (DIPF), Germany; Nora A. McIntyre & Robert M. Klassen, Department of Education, University of York, United Kingdom.

Correspondence concerning this article should be addressed to Anna-Katharina Praetorius, German Institute for International Educational Research (DIPF), Schlossstrasse 29, 60686 Frankfurt, Germany. Tel: +49 69 24708229. E-mail: praetorius@dipf.de

Abstract

One prominent problem of conducting observational assessments of teaching quality is the possibility of reactivity effects. To date, the issue of reactivity has received limited empirical attention. The present study, therefore, investigated reactivity in 447 students from 24 classes as well as their 12 teachers. We compared reactivity during lessons that were video-recorded with those that were not: according to t-test analyses of teacher ratings and MIMIC analyses of student ratings, no significant differences emerged in teaching quality or teaching practices. Significant differences were found in teacher and student emotions, as well as in student cognition and behavior. Supplementary eye-tracking analyses indicated reactivity depleted after 1 minute 20 seconds. The results are discussed with respect to their relevance for future video studies on classroom instruction.

Keywords: eye tracking; observer ratings; reactivity; student ratings; teacher ratings; video-based classroom research

Kurzzusammenfassung

Ein zentraler Nachteil von Beobachtereinschätzungen zur Erfassung von Unterrichtsqualität sind potentielle Reaktivitätseffekte. In welchem Ausmaß solche Effekte auftreten, wurde bislang kaum empirisch untersucht. Im Rahmen einer Videoerhebung wurden Daten von 447 Schüler(inne)n aus 24 Klassen sowie deren 12 Lehrkräften erhoben. Der Vergleich der Video- sowie der Nicht-Video-Bedingung zeigte sowohl für die Lehrereinschätzungen (analysiert mittels t-Tests) als auch die Schülereinschätzungen (analysiert mittels MIMIC-Modellen) keine Unterschiede hinsichtlich Merkmalen des Unterrichts (Unterrichtsqualität und Unterrichtspraktiken); in Bezug auf Lehrer- und Schüleremotionen sowie Schülerkognition und –verhalten zeigten sich hingegen Unterschiede zwischen den beiden Bedingungen. Die ergänzenden Eye-Tracking-Analysen deuten darauf hin, dass reaktive Blickbewegungen von Lehrkräften nach maximal 1 Minute 20 Sekunden nicht mehr nachzuweisen sind. Die Ergebnisse werden im Hinblick auf ihre Bedeutung für zukünftige Videostudien diskutiert.

Schlagworte: Reaktivität; Beobachtereinschätzungen; Unterrichtsforschung; Lehrereinschätzungen; Schülereinschätzungen; Eye-tracking

Reactivity Effects in Video-Based Classroom Research: An Investigation Using Teacher and Student Questionnaires as Well as Teacher Eye-Tracking

Over the last decades, knowledge has advanced considerably with regard to the components of quality instruction, teaching practice and student behavior in classrooms. This knowledge is largely derived from video-based studies which are rated by trained classroom observers (Brophy, 2006). In such studies, it is conventional for researchers to interpret observed events and actions as if these occur naturally. However, what actually is investigated is behavior occurring in the presence of a video camera or an observer (see also Masling & Stern, 1969; Samph, 1976). The established assumption that both conditions are comparable is only true if either both video cameras and observers have no influence on what is happening in classrooms at all, or the observation-related influences disappear quickly (Masling & Stern, 1969). In spite of the potential impact of reactivity to video-based research, notably few empirical investigations explore reactivity. Accordingly, the aim of the current study was to investigate the extent to which reactivity occurs in video-based classroom observation studies.

We first present an introduction to observations as a method of investigating teaching quality. The potential problem of reactivity in these studies is then outlined. Next, video-based empirical studies on reactivity effects are presented as well as studies on change of reactivity over time. Finally, the research questions of this study will be presented.

1.1 Observations as a Measure of Teaching Quality

Observer ratings often are considered to be optimal measures of teaching quality (Clare, Valdés, Pascal, & Steinberg, 2001; Dalehefte, Rimmele, Prenzel, Seidel, Labudde, & Herweg, 2009; Helmke, 2009; Petko, Waldis, Pauli, & Reusser, 2003; Pianta & Hamre, 2009): some even consider the approach to be a definitive in instructional research (e.g., Helmke, 2009; Klieme, 2006). Observer ratings have many advantages over other measures such as student ratings or teacher ratings (see Clare, Valdés, Pascal, & Steinberg, 2001; Clausen, 2002; Helmke, 2010; Kunter, 2005; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Petko et al., 2003; Pianta & Hamre, 2009; Rakoczy, 2008; Waldis, Grob, Pauli, & Reusser, 2010). Some of these are: 1) Observers are trained how to observe and rate the aspects of interest and thus should rate them in a more valid way compared to teachers and students who are usually untrained. 2) Observers are not involved in teaching at the same time and can thus focus on observing and rating. 3) Observers usually observe a variety of different teachers and thus have a good amount of comparison possibilities.

At the same time, using observer ratings has several disadvantages. Over and above the mere fact that they are expensive, the main critical points are the short observation period and the potential for reactivity effects, both threatening the validity of inferences that can be drawn using observer ratings. Whereas some work exists on the problem of short observation periods (see e.g., Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012; Newton, 2010; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014), the problem of reactivity effects has gained only limited empirical attention so far.

1.2 The Problem of Reactivity Effects in Video-Based Classroom Research

The problem of potential reactivity effects is mentioned regularly in the literature on observer-based research in many fields (for an overview, see Kazdin, 1982). Taking a closer look at the argumentation regarding reactivity effects, it is obvious that researchers agree on the fact that something might be changed through observation, but what exactly is assumed to change and whether the direction of change is positive or negative, differs considerably between studies.

With respect to classroom research, some authors argue that there is no reactivity. Kerlinger (1973) for example stated that a "teacher cannot do what she cannot do" (p. 539, see Samph, 1976). Medley and Mitzel (1963) wrote about the problem of reactivity that "this argument has merit but should not be taken too seriously" (p. 306) and Stigler et al. (1999) mentioned that it "is highly unlikely that teaching could be improved significantly simply by placing a camera in the room" (p. 6; see also Helmke, 2009). At the same time, Stigler (1998) argued that reactivity might exist so that observed lessons could be a "somewhat idealized version of what the teacher normally does in the classroom" (p. 141; see also Helmke, 2009; Samph, 1976). More concretely, Stigler et al. (1999) mentioned the possibility that teachers prepare their lessons better and use different methods compared to regular lessons. Whereas most researchers mention the possibility that observing leads to an overestimation of regular teaching practice, Clausen (2002) pointed to the possibility that the opposite might be the case (e.g., students having problems to concentrate due to the presence of an observer). Similarly, Carter (2008) mentioned that teachers might feel more stressed and threatened when being observed.

1.3 Empirical Investigations Regarding the Occurrence of Reactivity Effects

Most of the existing studies that mentioned the problem of reactivity effects in classroom research are video studies. Discussing the findings revealed in the international TIMSS video study, Stigler et al. (1999) stated that "it may actually be easier to gauge the degree of bias in video studies than in questionnaire studies. Teachers who try to alter their behavior for the videotaping will likely show some evidence that this is the case. Students, for example, may look puzzled or may not be able to follow routines that are clearly new for them" (p. 6). Empirical evidence regarding the occurrence of reactivity effects is, however, scarce. Some authors have reported narrative evidence that they experienced no or only a small degree of reactivity when conducting their own studies (e.g., Blease, 1983; Helmke, 2009). Actually measuring the extent to which reactivity occurs is very hard for an observer who does not know very much about the teacher, the students, and the patterns of their regular instruction. Thus, an unfamiliar observer in all likelihood will not be able to disentangle whether students look puzzled because of the observer being present or because they would look puzzled anyway. The majority of existing evidence regarding reactivity effects is therefore not based on observation but on items the observed teachers answered themselves after being videotaped. Interestingly, the existing data regarding teacher reports on reactivity effects mainly comes from German video studies; to our knowledge, the US video studies that were conducted over the last years did not report on any such questions. In the following, results of two video studies are reported as examples.

1.3.1 The amount of reactivity based on teacher reports

Many video studies have included rather broad questions regarding reactivity related to the *lesson in general*. In the German addition to the TIMSS video study, 96% of the 53 teachers included in the analyses reported that the two mathematics lessons that were videotaped were to a large degree typical or completely typical (Clausen, 2002). In the German IPN video study, 90% of the 50 videotaped physics teachers agreed their lesson was mostly or completely typical (Seidel, Prenzel, Duit, & Lehrke, 2003; Seidel, Prenzel, & Kobarg, 2005). The teachers, who agreed that the two lessons were different in the IPN study, explained in an open answer format that they *prepared their lessons* more intensively than was the norm (Seidel et al., 2003). With respect to *conducting the lesson*, 98% of the teachers in the German sample of the TIMSS video study reported that the methods they used in the videotaped lesson were similar or very similar to the lessons they conduct on other occasions (Clausen, 2002). In the IPN video study, teachers explained that their lesson was different, among others, with the fact that they used more or less experiments in their physics classes than usual (Seidel et al., 2005). Concerning the *quality of the lesson*, 72% of the German teachers in TIMSS video reported that the lesson was at least as good as usual (Clausen, 2002). Regarding *teacher behavior*, 48% of teachers reported that they were not or not overly nervous in the IPN video study (Seidel et al., 2005). Similarly, in a study involving American university teachers being observed by their peers, Kohut, Burnap, and Yon (2007) reported that there were nearly as many participants who did not feel stressed by being observed as participants who did feel stressed (3.38 on a 5-point scale; 5 = strongly disagree). Concerning *student behavior*, 79% of German teachers in the TIMSS video study reported that their classes behaved in a similar or very similar way (Clausen, 2002); the percentage was slightly lower (69%) in the IPN video study (Seidel et al., 2005). That the lesson differed to some extent from usual lessons was explained by teachers in this study as being due to differing levels of cooperation of students (higher or lower, depending on the teacher) as well as differing levels of silence and concentration (higher or lower, again depending on the teacher).

To sum up existing evidence, teachers themselves tend to judge reactivity effects regarding most aspects as not very problematic. However, it would be useful to not only ask about reactivity very broadly, but also about specific aspects of the phenomenon as some of them were mentioned by teachers to be influenced by reactivity. For example, the following aspects of teaching might be influenced by observation: (1) Preparation of lessons (e.g., investing more time to prepare), (2) conducting the lesson (e.g., implementing more student group work), (3) quality of the lesson (e.g., better classroom management), (4) teacher behavior (e.g., asking fewer questions than usual), and (5) student behavior (e.g., higher level of concentration). Distinguishing these aspects would allow better insight into whether reactivity occurs in video-based classroom research.

Although some teachers have reported differences between observed and unobserved lessons, it can be challenged whether teacher self-reports are an appropriate data source for assessing the extent of reactivity effects. The potential problems are many. First, teachers are involved in many interaction processes that require quick perception and action as well as the distribution of attention on several things at the same time (e.g., Doyle, 1986). Thus,

teachers may not be able to accurately perceive and process instruction in its full complexity. Second, teachers are involved in the instructional process as actors. They thus cannot separate their actions from their teacher role (see e.g., the research on actor-observer differences, Jones & Nisbett, 1972; Storms, 1973). Third, teachers can be influenced by self-enhancement biases, whereby they may present themselves in socially desirable ways at the expense of objectivity (e.g., Wubbels, Brekelsmans, & Hooymayers, 1992). Fourth, teachers usually do not have many possibilities for comparing their instruction with that of their peers, making it even more difficult to judge the quality of their own teaching (e.g., Clausen, 2002). More objective measures for reactivity are therefore needed.

1.3.2 Measuring reactivity through approaches other than teacher reports

Some attempts have been made to receive more objective data on reactivity effects using experimental approaches. In a study by Samph (1976), for example, verbal interactions of teachers were recorded using microphones that were installed weeks prior to the observations. Two experimental conditions (observer present vs. absent) were compared regarding a number of characteristics. T-test analyses suggested that the amount of time used for praising students, asking questions, and accepting student ideas was higher in the observer present condition. In another experimental study, Codding, Livanis, Pace, and Vaca (2008) found no differences between the observer present versus not present conditions regarding the implementation of a classroom behavior management plan for students with disorders. The evidence based on experimental studies therefore does not point clearly in one direction. In general, the existing experimental studies can be criticized regarding how well they actually tested reactivity effects. In the complex situation of teaching it is not easy to succeed in actually achieving the two intended experimental conditions. In the study of Codding et al. (2008), for example, the observer absent condition was simulated by positioning the observer behind a one-way window. As the teachers saw the window, it can be challenged whether they really felt that the observer was absent.

Considering other, non-experimental measures would therefore be useful. One real-world option is student ratings, as the students know their teachers well and can compare the observed lessons to regular lessons. Student ratings on reactivity were used in single video studies, for example in the "German-English International" (DESI) study (Helmke, T., Helmke, A., Schrader, Wagner, Nold, & Schröder, 2008). However, the results of these ratings to our knowledge have not been published; the only hint can be found in Helmke (2009) where it is mentioned that students were asked whether the videotaped lesson was different from other lessons; according to the students, there was less noise in the videotaped lessons and the teacher used more material. Another real-world option is using other measures that can be conceived as comparably objective by using behavior-based measures (e.g., electroencephalogram, galvanic skin response, heart rate monitors). Given their intensive behavioral sampling capability, participants are less likely to manipulate their own behavior for social desirability when their movements are sampled a number of times within each second. Mobile eye-tracking is another such methodology. Other than placing minimal additional demand on the eye-tracked participant (Glaholt & Reingold, 2011), gaze is automatic by nature in that it is very much guided by the task at hand (Land & Hayhoe,

2001) and therefore reflects internal priorities that are driving the viewer (Yarbus, 1967; De Angelus & Pelz, 2009). The potential of gaze to indicate cognitive priorities is the basis of scanpath research (e.g., Foulsham & Underwood, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007) and wider vision research aiming to highlight the top-down guidance of gaze direction (Hristova, Georgieva, & Grinberg, 2010; Schyns & Oliva, 1994; Tatler, Gilchrist, & Land, 2005). Furthermore, since attentional selectivity increases with task complexity (Glaholt, Wu, & Reingold, 2010)—and teaching is widely recognized to be a high-complexity profession (e.g., Berliner, 2001; Feldon, 2007)—we can confidently regard attentional direction (or gaze direction) to be a valid indication of what teachers are focused on during teaching. Indeed, teachers have already been eye-tracked as part of instructional quality research (Cortina, Miller, McKenzie, & Epstein, 2015). We therefore selected teacher gaze as it is a suitable means to tracing teachers' focus of attention during video-based classroom research.

1.4 Does Reactivity Decrease Over Time?

Heyns and Zander (1953) recommended "to keep an observer in the observational setting for long enough to be perceived as a 'piece of the furniture'" (cited after Samph, 1976). Other researchers (e.g., Helmke, 2009; Masling & Stern, 1969) state as well that reactivity effects exist but disappear after a short period of observation time, i.e., suggesting that teachers quickly get used to being observed and resort to typical behavior. Again, empirical evidence for such a change over time is scarce. In the IPN video study, some teachers reported in an open-answering format that they were mainly nervous at the beginning of the lesson (Seidel et al., 2003). Masling and Stern (1969) used a more indirect measure and investigated the diminishing effects of reactivity over time through correlating early and late observer ratings of the same teachers over time. They concluded that there is not clear evidence for diminishing effects, as some correlations increased, some decreased, and for some there was no change over time at all. Drawing conclusions based on this finding is, however, only possible to a limited degree as the assumption that observer effects should be visible in correlations of ratings is questionable. Additionally, actual changes of behavior over time have not been taken into account.

1.5 Research Questions

Very few empirical investigations regarding reactivity effects in classroom research exist. We therefore aimed to investigate the amount of reactivity effects when using observer ratings in video-based classroom research (Research Question 1). The existing evidence is based on global judgments that do not differentiate the different dimensions reactivity could make a difference on. In the present study, reactivity is therefore investigated in multiple domains with respect to lesson preparation, teaching practices, teaching quality, teacher behavior, and student behavior.

Additionally, most evidence regarding reactivity effects is based on teacher reports, although this data source may not be sufficiently trustworthy regarding all aspects of classroom instruction. In the present study, mobile (i.e., glasses) eye-tracking is used in addition to student and teacher reports in order to investigate the existence of reactivity effects.

Lastly, changes of reactivity over time might exist. Whether this is true is investigated in the present study using teacher eye-tracking data (Research Question 2).

2 Method

2.1 Sample and Procedure

Teacher participants consisted of 12 volunteers from one comprehensive (i.e., state) secondary school in the north of England. The teacher sample consisted of five males and seven females. Years of teaching experience was $M = 14.08$ years, ranging from 2 to 33 years. Teachers were $M = 41.17$ years old, the youngest being 27 and the oldest being 57 years old. Ten teachers participated in both the video (first condition) and control conditions (second condition); two teachers participated only in the video condition.

Student participants were recruited from the classroom groups of the teachers who participated. After each video-recorded (i.e., video condition) and non-video-recorded (i.e., control condition) lesson, teachers were requested to complete a reactivity questionnaire with their students. Thus, the aim was for each teacher to be associated with two sets of student data: one set of data for students who experienced a video-recorded lesson (i.e., the video condition) and data from a second group of students who experienced a technology-free lesson (i.e., control condition). Twelve student groups participated in the video condition: students were aged $M = 13.58$, ranging from 10 to 16 years; 114 (43.70%) were male and 136 (52.10%) were female. Ten student groups participated in the control condition: students in this condition were aged $M = 14.41$, ranging from 11 to 17 years; 95 (51.90%) were male and 88 (47.30%) were female.

Throughout, the sampling method for both teachers and students was opportunity sampling. Teachers were approached by the study's in-school contact-person for the video condition; teachers were then asked by researchers to participate in the control condition. Students in both conditions were chosen according to the lesson that teachers opted for.

Before the video-recorded lesson started, the eye-tracking glasses were fitted and calibrated to the teacher, gaze recording was initiated, and the data recorder was transferred to a waist bag worn by the teacher. As students approached the classroom, the researcher sat down at the back of the classroom. Once the teacher had settled the students, the researcher briefed the students, mainly to emphasize the importance of behaving 'as normal'—namely, to resist deliberately drawing attention to themselves in the video cameras, but rather to engage with the lesson as they normally would. At the end of the lesson, the questionnaires were distributed to both teacher and students, who completed the questionnaires immediately and independently.

2.2 Materials and Apparatus

We developed scales with the specific goal of investigating reactivity from both the teacher and student perspective. The scales were based on a screening of items used in previous studies (e.g., TIMSS video and IPN video) and on all additional areas mentioned in the literature where reactivity might occur (see also section 1.3.1). In developing the scales, we aimed to systematically differentiate between major areas of human perception and behavior likely to be relevant to reactivity. The following areas were distinguished for the student as well as the teacher questionnaire (for an overview of the single items used, see

Table 1): (a) student negative emotions (3 items), (b) student motivation (2 items), (c) student cognition (3 items), (d) student behavior (2 items), (e) teacher negative emotions (2 items), (f) teacher motivation (3 items), (g) teaching practices (6 items), and (h), teaching quality (3 dimensions with 2 items each: student support, classroom management, and cognitive activation). For teachers, additional items on lesson preparation (i) were developed (4 items); these were not added to the student questionnaire as it cannot be assumed that students have enough information to answer such items validly. All items used the same item stem ("Compare this lesson to normal lessons. Compared to other lessons, …" and a 5-point scale, ranging from 1 (= much less) to 5 (= much more). For students, the internal consistencies were sufficient to a large degree (see Table 2). For teachers, however, this was not the case for some scales. We therefore decided to analyze the teacher data item-by-item instead of aggregating items to scales.

SMI Natural Gaze eye-tracking glasses were used to record teacher gaze. This technology includes high-definition scene video recording as well as audio recording. The eye-tracker looks like normal glasses and thus can be assumed to be non-intrusive to the students (regarding the intrusiveness for teachers, see section 1.3.2). Additionally, two video cameras (Panasonic TM700) were used. In keeping with conventional video-based classroom research (e.g., Seidel et al., 2005), one camera was set up in the front of the classroom to record student behavior and a second camera was set up at the side of the classroom in the mid-point between the front and back of the classroom to record teacher behavior. When it was not possible to set up the camera in the mid-point of the classroom, the camera was set up at the back of the classroom with the same focus—that is, on the teacher. The rear (second) camera was also used to record key classroom events. Both cameras were secured to a tripod for stability and to maximize participants' experience of conventional video-based classroom research. Since studies using this set-up aim to investigate teacher behavior as part of classroom-based research questions, the video-recording equipment was positioned to 'see' and capture teachers, which in turn means that the equipment was within teachers' visual field most of the time.

2.3 Analyses

2.3.1 Teacher and student questionnaire data analysis

Due to the small sample size, the teacher questionnaire data should be interpreted cautiously. For getting an impression whether the data of the present study is comparable to those of previous studies, paired t-tests (two-tailed) for dependent samples between the teacher ratings of the video condition and the control condition were conducted. We chose t-tests as they have higher power in detecting differences compared to non-parametrical tests. However, as the sample size was very small, we additionally conducted non-parametrical Wilcoxon-U tests for the variables that revealed to show significantly different means between the video and the control condition in the t-tests; thus, we could test to what extent the t-test results can be replicated with a non-parametrical test.

The student data were analyzed using MIMIC models in the context of structural equation modelling (see e.g., Miner, Dowson, & Sterland, 2010). The information on whether a student answered the items in the video vs. control condition therefore was added as a predictor to the model. For each area of potential reactivity (e.g., teaching quality or

student perception and behavior), a separate model was estimated. All scales were included as latent variables in the model. As the teaching practice measures were single items rather than scales, these were included as manifest variables. The hierarchical structure of the data (students nested in classes) was taken into account using the type = complex command in Mplus. All student data analyses were conducted with Mplus 7 (Muthén & Muthén, 1998-2012).

2.3.2 Teacher eye-tracking data analysis

The eye-tracking data was coded using BeGaze software. The semantic gaze-mapping facility was used in which fixations were displayed as a still image for the researcher to code. Fixations were coded for teacher gaze target, which we coded as *research* versus *non-research* targets. Research targets included *researcher, researcher and camera, research other* (i.e., in the region of the researcher). Non-research targets included *students, resources* (e.g., student desk, student materials, PowerPoint), *school staff* (i.e., teaching assistant or another teacher), *non-research other.* We coded fixation targets during the first 20 minutes and the final 10 minutes of each lesson.

Eye-tracking analyses were conducted using three approaches. The first analysis was to make research–non-research comparisons of teacher attention. That is, what category (i.e., research vs. non-research) of classroom regions was gazed at more by teachers during video classroom research (i.e., video condition)? The second analysis was to make pre-/post-test comparisons of the duration of teacher gaze at specified targets at the start of the lesson compared with the end of the lesson (Gaze Analysis 2). To do this, repeated measures analyses of variance was used to compare mean gaze durations towards research versus non-research targets during the first and last 10 minutes of eye-tracking. We also compared gaze durations of research versus non-research targets at the start versus the end of data collection. The third analysis of the eye-tracking data was time series analysis to assess whether significant changes occurred in teachers' attention during their teaching (Gaze Analysis 3). The first stage of the time series analysis involved auto-regressive integrated moving average (ARIMA) analysis. Repeated measures ANOVA analysis informed whether the changes were statistically significant. The second stage of time series analysis involved interrupted time series (ITS), which we used to trace where the teacher was primarily interested in at regular intervals of time. As suggested by its name, ITS analysis involves the identification of interruptions to a linear trend: interruptions are points where participant behavior changes dramatically (Figure 1). Thus, in ITS, new variables for each potential interruption are generated where dummy codes indicated the start and end observation for that. For each interruption, a statistically significant change in slope indicated that the interruption could potentially be interpreted as an onset of reactivity depletion. The interruption that finally qualified as the onset of reactivity depletion was the interruption immediately preceding zero-level research gaze, that is when teachers' gaze towards *research targets* stopped. Following the guidance given by Todman and Dugard (2001), we divided the 20 minute eye-tracking data into 20-second intervals, yielding 60 observations in total thereby satisfying the requirement of having 50 observations or more (cf. Schmidt, Perels & Schmitz, 2010). The mean durations (in seconds) of all participants' gaze towards research and non-research areas of interest (AOI) were obtained.

Throughout, we used relativized measures of gaze duration in analysis. To achieve relativized gaze measures for each participant, we first allocated every event to one observation number from 1 to 60. We then counted the number of each gaze event occurring within each observation number. The duration of each research gaze was summed and divided by the total count of that gaze type during that observation. Measures of duration per visit were thus computed for each participant. We then obtained the mean duration per visit across all participants for each observation number: these whole-sample mean duration per visit measures were the bases of our statistical analyses. Although we had individual gaze targets (i.e., individual, uncategorized targets; e.g., non-research resources), we constrained the present analyses to research versus non-research targets in order to yield meaningful results. Since the time series analysis of gaze target sub-categories yield significant differences throughout the full duration of the coded part of the lesson, we dropped this detail from the data to obtain broader—more informative—changes in teacher attention. Individual gaze targets were only used in Gaze Analysis 1.

3 Results

3.1 Descriptive Analyses Regarding Teacher and Student Data

The descriptive statistics (means, standard deviations, Cronbach's alpha, ICC's) for the teacher as well as the student data on reactivity can be found in Table 2.

3.2 Teacher-Reported Reactivity

The t-tests for the teacher-reported data showed only few reactivity effects when the observed and the non-observed conditions were compared. For the teaching practices (6 items) as well as the teaching quality items (6 items), none of the t-tests indicated significant differences. Regarding the lesson preparation, one difference (namely thinking through the concept of the lesson more thoroughly; out of four possible differences) revealed to be significant ($t = 1.92$, $df = 11$, $p = .04$, Cohen's $d = 1.16$). Regarding teacher variables, none of the two motivation variables revealed significant differences; for teachers' negative emotions, both items showed significant differences (worried: $t = 2.84$, $df = 19$, $p = .01$, Cohen's $d = 1.30$; nervous: $t = 3.26$, $df = 19$, $p = .004$, Cohen's $d = 1.50$). Finally, for the student variables, none of the items related to motivation (2 items), cognition (3 items), and behavior (2 items) showed significant differences; for students' negative emotions, one (out of three) differences were significant (nervous: $t = 2.16$, $df = 11$, $p = .03$, Cohen's $d = 1.30$). All occurring differences indicate less positive values in the observed compared to the non-observed condition.

Conducting non-parametrical Whitney-U tests for the variables that have shown significant differences between both conditions revealed no significant differences for any of the variables (lesson preparation: $U = -1.41$, exact $p = .25$; teacher worry: $U = -1.63$, exact $p = .13$; teacher nervousness: $U = -1.90$, exact $p = .07$; student nervousness: $U = -1.63$, exact $p = .11$).

3.3 Student-Reported Reactivity

For the student-reported data, four MIMIC models were estimated. All models had sufficient fit indices (model including all student variables: $\chi2 = 63.46$, $df = 35$; CFI = .96; RMSEA = .04; SRMR = .04; model including all teacher variables: $\chi2 = 5.87$, $df = 5$; CFI =

.99; RMSEA = .02; SRMR = .03; model including all teaching quality variables: $\chi 2 = 8.43$, *df* = 8; CFI = 1.00; RMSEA = .01; SRMR = .02; for the model containing the teaching practices, no fit indices exist as it was a saturated model, see section 2.4.1).

The factor loadings, correlations, and path coefficients for all models can be found in Table 3. Neither for the teaching quality model nor for the teaching practices model, any of the path coefficients of the video condition variable was significant; thus, for these aspects, it can be assumed that no reactivity occurred. In the student as well as the teacher model, the path coefficient on negative emotions was significant; furthermore, in the student model, additionally student cognition and student behavior seemed to be influenced by reactivity effects, as can be seen based on the significant path coefficients. All of the significant coefficients indicate less positive values in the observed compared to the non-observed condition.

3.4 Reactivity in Teacher Gaze

3.4.1 Preliminary teacher gaze analysis

Initial comparisons were made between teacher attention towards research targets and non-research targets. According to repeated measures MANOVA, teachers looked significantly more at non-research targets than research targets, $F(1,59) = 1478.50$, partial $eta^2 = .96$, $p < .001$. Comparisons among individual targets were then made using repeated measures ANOVA; teacher attention towards individual non-research targets were still significantly greater than towards individual research targets, $F(1,56) = 1015.51$, partial $eta^2 = .99$, $p < .001$ (Figure 2). Paired samples t-tests further supported these individual comparisons. Namely, teachers looked more at non-research targets in general than research targets (or area of interest, AOI1; Table 3), $t(59) = 37.87$, $p < .001$, $d = 8.30$. Teachers looked more at students than the researcher, $t(59) = 28.38$, $p < .001$, $d = 5.41$. Teachers looked more at learning resources than the cameras, $t(59) = 32.06$, $p < .001$, $d = 6.04$. Teachers looked more at non-research other than research other, $t(59) = 14.43$, $p < .001$, $d = 2.57$. However, teachers did not look more at school staff than the researcher and camera ($p = .95$).

3.4.2 Pre-/post-test teacher gaze comparisons

To begin exploring how teacher attention changes over the course of the video-recorded lesson, we compared what teachers looked at during the start of the lesson with what they looked at at the end of the lesson. A Pre-/Post-Test approach was taken, whereby repeated measures ANOVA was run on the lesson start–end difference in teacher attention towards each, research and non-research, targets. Teachers gazed at *research* targets significantly less at the end of the data collection, $F(1,11) = 117.76$, $p < .001$, partial $eta^2 = .92$. Gaze at *non-research* targets was not significantly different at the start ($M = .25$s) compared with the end ($M = .27$s) of data collection ($p = .18$; Figure 3).

3.4.3 Time series analysis of teacher gaze

3.4.3.1 Overall attentional changes

Next, we explored whether teacher attention significantly changed over time. To do this, ARIMA analysis was run on teacher gaze towards each, research and non-research, targets. According to ARIMA time series analyses, teacher attention towards *research* targets did significantly change over time, $b = -.66$, s.e. = .11, $t = -5.92$, $p < .001$. However, teacher attention towards *non-research* targets did not change over time ($p = .72$; Figure 4).

3.4.3.2 Onset of reactivity depletion in teacher attention

Visual inspection of the time series graphs (Figure 4) shows clear depletion of reactivity in teacher attention (i.e., gaze towards research targets). A simultaneous surge of gaze directed solely at non-research targets is also visible from these graphs. Visual inspection also suggested observation number four (i.e., 1min 20s) to be the interruption which led to zero-level reactivity, making 1min 20s the likely 'onset of reactivity depletion'.

We therefore ran interrupted time series (ITS) analyses to identify the important 'interruptions', which in turn were interpreted as onset of reactivity depletion in teacher attention. Two stages were involved in ITS analyses. First, a minute-by-minute approach was taken to identify which interruptions should be included in the final ITS model. Since, in their own individual models, the first ten out of the 20 minutes and the fourth observation (1min 20s) were significant interruptions to teacher gaze towards research targets (our measure of reactivity), these specific interruptions were included as parameters in the second stage of ITS analysis when the full ITS model was run (Table 5).

The full ITS model of *research gaze* durations revealed two significant interruptions in gaze towards research targets, namely the 2-minute mark, $b = .66$, s.e. = .27, $t = 2.43$, $p = .02$, and the 1min 20s mark, $b = 1.43$, s.e. = .28, $t = 25.07$, $p < .001$. The full ITS model *non-research gaze* durations revealed only one significant interruption, namely the 1min 20s mark, $b = -3.25$, s.e. = .1.04, $t = -3.12$, $p = .004$. The interruption of research gaze at 1min 20s was thus supported by the sole significant interruption of non-research gaze at the same time (Table 5) as well as the visual inspection of the time series graphs (Figure 4), whereas the interruption of research gaze at 2 mins was not. We therefore interpreted 1min 20s to be the onset of reactivity depletion, that is when teachers' attentional reactivity ends.

4. Discussion

Existing evidence in instructional research is largely based on video-based ratings of external, trained observers (Brophy, 2006). To draw meaningful conclusions from these studies, it is necessary to assume that the presence of a video camera or an observer does not change what is happening in the classroom. Therefore, the present study investigated whether this assumption is true. Specifically, we explored the presence of reactivity in response to video-based classroom research and asked whether reactivity changes over time during a researched lesson.

4.1 The Extent of Reactivity Effects and Its Explanation

The present investigation differentiated between dimensions of reactivity that could potentially occur. Based on a review of aspects that were assumed to be related to reactivity as well as based on theoretical considerations (see Carter, 2008; Clausen, 2002; Stigler et al., 1999), we distinguished teaching practices, teaching quality, and lesson preparation on the

classroom level. We also included teacher aspects on the classroom level such as his or her emotions and motivation. On the individual student level, we included aspects related to student emotions, motivation, cognition, and behavior. Our analyses showed that it is important to differentiate between different dimensions to investigate reactivity effects as reactivity occurred only with respect to teacher emotions as well as student emotions, cognition, and behavior. No reactivity effects were found regarding aspects related to teaching, namely teaching practices, teaching quality, and lesson preparation (with the exception of thinking through the concept of the lesson more thoroughly, if the t-test results are referred to; the Whitney-U test showed no effect for this aspect). These findings are in line with the quantitative ratings as well as qualitative investigations of videotaped teachers in the TIMSS Video and the IPN Video studies (see section 1.3.1); in both studies, teachers reported that their lessons were typical to a large extent with respect to their teaching as well. Negative emotions of teachers were reported to a large extent; additionally, qualitative findings for small number of teachers in the IPN Video study also indicated that the students' behavior was partly different. As reactivity seems to differ depending on the aspects investigated, we recommend that future studies use differentiated dimensions of reactivity in keeping with the present study. Interestingly, all existing differences between the video and control condition point towards a negative impact of videotaping, confirming the assumptions of Carter (2008) and Clausen (2002). That the observed lesson is idealized compared to regular instruction (see the assumption if Stigler et al., 1999) could therefore not be supported based on the data of the current study.

The central contribution of the present investigation is that we not only used teacher reports but also student reports and eye-tracking data to measure the extent of reactivity. It is noteworthy that teacher and student reports were consistent to a large extent. Both students and teachers perceived that video-based observation is connected to more negative emotions for students as well as teachers (at least based on the t-test results; again no effects for teachers were found based on the Whitney-U test). Students and teachers in both conditions did also report similar values regarding all teaching-related aspects such as teaching practices and teaching quality. However, students in the video condition compared to the control condition perceived more reactivity regarding aspects that are related to themselves than the teachers did; whereas the teacher data only indicated that students have more negative emotions (based on the t-test results; using the Whitney-U test, no such effect was found), the student data also showed differences with respect to cognition (e.g., less concentrated) and behavior (e.g., less engaged). Teacher reports might therefore rather lead to an underestimation of reactivity effects related to the students.

Taking both perspectives together, we would conclude that reactivity effects do not seem to distort video-based results regarding teaching practices and teaching quality, as some researchers hypothesized (e.g., Helmke, 2009; Stigler et al., 1999). We should, however, be careful not to over-interpret single behaviors or face expressions of students and teachers, as their emotions, and for students also their cognition and behavior, might not be as usual. Suggestions in the literature to extend investigations of teachers' emotions in the classroom to observer ratings (e.g., Becker, Keller, Goetz, Frenzel, & Taxer, 2015) have therefore to be seen critically as the observed emotions might not be valid.

One explanation why reactivity effects occurred only to such a limited degree is the sample used. Teachers who agree to be videotaped are probably a selective group (e.g., Borich, 2008; Liang, 2015). It can be easily imagined, for example, that these teachers are rather confident with respect to their teaching quality. How reactivity would look like for teachers who do not participate in such research willingly, remains open. However, the representativeness of teacher samples in video-based classroom research is a more general problem. The current study allows conclusions regarding these non-representative samples that are usually investigated in classroom research.

So far, we do not know much about reasons why reactivity occurs or not. Focusing future investigations on potential determinants for reactivity would therefore be highly interesting. One possible group of determinants is teacher characteristics like teaching experience or, more specifically, experience with being observed during lessons. Another group of determinants is the conditions of observation like, for example, the video equipment used. For the study at hand the recording equipment was chosen so that it represents conventional video-based classroom research equipment. Current development in video technology would, however, also allow using very small recording equipment and managing the equipment from outside of the classroom. In such a case, the salience of being observed would be much smaller. We then would not necessarily expect the same patterns as reported in this paper—especially where teacher gaze is concerned. The extent of reactivity could then be even smaller compared to the findings in the present study.

4.2 Change in Teacher Reactivity Over Time

Teacher eye-tracking enabled us to focus on the reactivity impact of video-based classroom research on teachers in particular. Teacher attention was focused undeniably and unequivocally on classroom instruction during video-based research. This focus on instruction was demonstrated by the longer durations of teacher gaze towards non-research regions in contrast to research regions of the classroom. As expected, participating teachers' gaze demonstrated that the complexity (Chisholm, Caird & Lockhart, 2008), difficulty (Rötting, 2001) and importance (Mackworth & Morandi, 1967; Schumann et al., 2008) of the non-research areas dramatically outweigh those of research areas in the classroom. We were not surprised that video research imposed minimal attentional distraction on teachers, given the pressured, sophisticated and unpredictable nature of classroom teaching (e.g., Berliner, 2001). Compared with the research targets, non-research classroom regions possessed unmatched relevance (Charness, Reingold, Pomplun & Stampe, 2001; Foulsham & Kingstone, 2012) to teachers as they carried out their tasks of classroom instruction.

Yet, a definite change in teacher attention was found over the course of the lesson: teachers directed their attention towards research areas significantly more at the start compared with the end of the lesson. Teacher gaze towards research targets also significantly changed over time. Accordingly, we cannot ignore that teachers do give regard to research regions of the classroom during video-based research—especially at the start of a researched lesson. On this matter, some classroom observation researchers (e.g., Helmke, 2009; Masling & Stern, 1969) have proposed that, though reactivity occurs, these effects disappear after a negligible period of time. Others have specified that video research

participants need an acclimatization period of five to ten minutes (e.g., Kingstone, 2013). Our teacher gaze analysis supports this recommended acclimatization time frame, in that teachers' attentional reactivity indeed terminated within the first five to ten minutes of classroom observation. However, our results further suggest that an even shorter acclimatization period is sufficient before teaching behavior can be assumed to be 'ordinary' and regarded as unrelated to the presence of an external observer or video technology. Namely, teachers cease to attend to research targets in the classroom within one minute and 20 seconds: at this point, classroom instruction receives the teacher's full (gaze) attention.

4.3 Limitations and Further Directions

The study at hand allows more differentiated statements regarding the effects of reactivity in video-based classroom research than this was possible in prior studies. However, the conclusions that can be drawn based on the study are limited due to several reasons. The most important ones are mentioned in the following; these restrictions at the same time give hints how to investigate reactivity in future studies in a more optimal way.

One limitation is the small sample used in the present study. As eye-tracking was involved, a larger sample was not feasible; in fact, the sample in the present study can even be seen as a large sample in contrast to some eye-tracking studies (cf. MacDonald & Tatler, 2015; Tatler et al., 2013) and comparable with others (Cortina et al., 2015). Moreover, using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), post-hoc power analysis was conducted on the paired samples t-test for research versus non-research target gaze (d = 8.30), the sample size was determined to be b = 1.00, which satisfied the conventional b = .80 power requirement. However, regarding the teacher and student ratings, a larger sample size would be desirable. The teacher sample is actually too small for conducting trustworthy t-tests. We therefore also conducted non-parametrical Whitney-U tests which, however, have only limited power to detect existing differences compared to t-tests. Indeed, all statistically significant differences based on the t-tests could not be replicated using the Whitney-U test, leaving the reader not knowing whether existing differences could not be detected with the Whitney-U test or whether the t-test yielded unreliable results due to the small sample. For the student data, we did take the clustered structure of the data (students nested in classes) into account in the analyses, but we could not conduct two-level analyses. For the aspects of reactivity that are located on the student level (e.g., students' emotions), this is even more appropriate; for aspects on the classroom level, a multilevel-analysis, however, would be more appropriate. Additionally, as the teacher and the student sample differed so largely with respect to sample size, comparing teacher and student reports with respect to reactivity might not be trustworthy.

A second limitation is the assignment of students respective classes to the video vs. control condition as teachers were first asked to participate with a class of their choice in the video condition and second to use student and teacher questionnaires in another class at the same day. It can be hypothesized that reactivity should occur independent of specific class characteristics and that it thus would not be overly problematic even if classes in the video and control condition differed. However, a random assignment would be more appropriate for investigating reactivity.

The design used is a third limitation of the study. It was a partially within-subjects design as we had the same teachers in both conditions (with vs. without video) but not the same students. We did choose the design as it allowed us to have the surveys of both conditions filled out at the same day, therefore being able to reduce occasion-specific variation. Additionally, students did not need to fill out the questionnaire twice within a very short period of time, which could have unintended effects such as memory effects or decreased motivation in answering the items. However, the design has the drawback that the student populations for both conditions are different; existing differences between those can therefore not unambiguously be interpreted as differences regarding the two conditions. In future studies, ideally, both design options should be included in a study, so that the effects the design choice has can be evaluated. Regarding the design, it also needs to be mentioned that our study focused on reactivity effects in video-based classroom research. The effects found might be different from the ones one would find if investigating reactivity regarding live observation (for differences in research findings between live and video observation, see Casabianca, McCaffrey, Gitomer, Bell, Hamre, & Pianta, 2013).

The measures used in the study are the fourth limitation of the study. In terms of the student surveys, it can be criticized that the intra-class correlations, and thus the amount of shared variance on the class level, are low. This indicates that students in a class perceive reactivity in very different ways. This is plausible regarding their own emotions, motivation, cognition, and behavior, but is problematic if aspects on the class level (i.e., the teaching aspects and the teacher variables) should be assessed. Eye-tracking is assumed to be a more objective measure. However, eye-tracking data cannot be fully interpreted without supplementary verbal data. In line with hand movement research (McNeill, 1985), eye movements can be regarded as more valid indicators of viewer cognition when associated verbalizations are used as accompanying analysis (for simultaneous verbalizations, see Church, Kelly & Holcombe, 2014; for retrospective verbalizations, see van Gog, Paas, van Merriënboer & Witte, 2005). Additionally, though minimal (Glaholt & Reingold, 2011), eye-tracking will nonetheless have led to some reactivity itself. Third, eye-tracking is a behavior-oriented measure but it sheds light solely on visual attention, which is only one out of many non-verbal reflexive behaviors. Thus, the present measures are not comprehensive for investigating reactivity. In other research areas, reactivity is investigated with respect to performance differences in quality and quantity (e.g., working on a speed motor task, Becker & Marique, 2015). However, as teaching quality cannot be measured in an objective way (Clausen, 2002; Kunter & Baumert, 2006), just counting differences with respect to quantity is not an option as well. Optimal from a research perspective would be covert observation (e.g., Samph, 1976), but this approach is problematic due to ethical reasons. It therefore seems that there is no single optimal method for investigating reactivity. Using a combination of different approaches as done in the present study therefore is probably the best—and most realistic—alternative. To build on the current design, a positive next step would be to analyze the verbal data (simultaneous, retrospective, or both) associated with the teacher gaze data, to further interrogate the present conclusions. Other than asking students to wear eye-tracking glasses as well as teachers, subsequent research into reactivity in videoed classroom research should also consider integrating other intensive and reflexive

measures such as skin conductance, heart rate and cortisol measures: if time series analyses of those measures reveals the same timescale for onset of reactivity depletion, then the present study would be corroborated. A richer understanding of the nature of classroom reactivity would also be enabled by incorporating more reflexive measures in to address our research question. For example, we may find further support for stress as a primary emotion related to reactivity in video-based classroom research.

4.4 Conclusions

Reactivity effects can seriously impact research findings. The present study showed that, for most aspects, this seems not to be the case in video-based classroom research. Especially if researchers are mainly interested in teaching practices and teaching quality, reactivity seems to be negligible. Only if researchers focus on teacher emotions or student emotions, cognition, and behavior, reactivity needs to be taken into account to a greater extent. And even then this might not be too problematic as our findings indicate that reactivity effects disappear after a very short period of time. All in all, reactivity seems not to distort study results in video-based classroom research to a large degree—a finding that supports the credibility of an important research area in educational research.

References

Becker, T. E., & Marique, G. (2014). Observer effects without demand characteristics: An inductive investigation of video monitoring and performance. *Journal of Business and Psychology, 29*(4)*,* 541-553.

Becker, E. S., Keller, M. M., Goetz, T., Frenzel, A. C., & Taxer, J. L. (2015). Antecedents of teachers' emotions in the classroom: an intraindividual approach. *Frontiers in Psychology*, *6*, 635.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*(5), 463–482.

Blease, D. (1983). Observer effects on teachers and pupils in classroom research. *Educational Review, 35*(3), 213–217.

Borich, G. D. (2008). *Observation skills for effective teaching*. Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

Brophy, J. (2006). History of research on classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 17–43). Mahwah, NJ: Lawrence Erlbaum Associates.

Carter, V. (2008). Five Steps to becoming a better peer reviewer. *College Teaching, 56*(2), 85–88.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & cognition, 29*(8), 1146–1152.

Chisholm, S., Caird, J. K., & Lockhart, J. (2008). The effects of practice with MP3 players on driving performance. *Accident Analysis & Prevention, 40*(2), 704–713.

Church, R. B., Kelly, S., & Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition and Neuroscience, 29*(3), 345–354.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. R. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (No. 545). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clausen, M. (2002). *Qualität von Unterricht – Eine Frage der Perspektive? [Quality of instruction – A matter of perspective?].* Münster, Germany: Waxmann.

Codding, R. S., Livanis, A., Pace, G. M., & Vaca, L. (2008). Using performance feedback to improve treatment integrity of classwide behavior plans: An investigation of observer reactivity. *Journal of Applied Behavior Analysis, 41*(3), 417–422.

Cortina, K. S., Miller, K., McKenzie, R., & Epstein, A. (2015). Where low and high inference data converge: Validation of CLASS assessment of mathematics instruction using mobile eye tracking with expert and novice teachers. *International Journal of Science and Mathematics Education, 13*(2), 389–403.

Dalehefte, I. M., Rimmele, R., Prenzel, M., Seidel, T., Labudde, P., & Herweg, C. (2009). Observing instruction "next-door": A video study about science teaching and learning in Germany and Switzerland. In T. Janík & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 83-101). Münster, Germany: Waxmann.

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition, 17*(6-7), 790–811.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction. Comparing student and teacher reports. *Educational Policy, 24*(2), 267–329.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 392–431). New York, NY: Macmillan.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. *Educational Psychologist, 42*(3), 123–137.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement, 33*(3), 613-619.

Foulsham, T., & Kingstone, A. (2012). *Goal-driven and bottom-up gaze in an active real-world search task.* Paper presented at the Proceedings of the Symposium on Eye Tracking Research and Applications.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision, 8*(2), 1–17.

Glaholt, M. G., & Reingold, E. M. (2011). Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics, 4*(2), 125–146.

Glaholt, M. G., Wu, M.-C., & Reingold, E. M. (2010). Evidence for top-down control of eye movements during visual decision making. *Journal of vision, 10*(5), 15–15.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts [Quality of instruction and teacher professionalism: diagnosis, evaluation, and improvement of instruction]*. Seelze, Germany: Klett-Kallmeyer.

Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Die Videostudie des Englischunterrichts. In DESI-Konsortium (Eds.), *Unterricht und Kompetenzerwerb zu Deutsch und Englisch. Ergebnisse der DESI-Studie [Instruction and competence development in German and English as a foreign language. Results of the DESI study]* (pp. 345-363). Weinheim, Germany: Beltz.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford: Elsevier.

Heyns, R. W., & Zander, A. F. (1953). Observation of group behavior. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 381–417). London: Staples Press Limited.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher, 41*(2), 56–64.

Hristova, E., Georgieva, S., & Grinberg, M. (2011). Top-down influences on eye-movements during painting perception: the effect of task and titles. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller, & G. Scarpetta (Eds.), *Toward autonomous, adaptive, and context-aware multimodal interfaces. Theoretical and practical issues* (pp. 104–115). Heidelberg: Springer.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains.* Retrieved from: http://files.eric.ed.gov/fulltext/ED540960.pdf

Kazdin, A.E. (1982). Observer effects: reactivity of direct observation. *New Directions for Methodology of Social & Behavioral Science*. *14*, 5–19.

Kerlinger, F. N. (1973). *Foundations of behavioral research.* New York, NY: Holt, Rinehart, and Winston.

Kingstone, A. (2013). *The cycle of social signaling*. Keynote presented at the European conference on eye movements, Lund.

Klieme, E. (2006). Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. Einführung in den Thementeil [Empirical instructional research: current developments, theoretical basis and subject-specific findings]. *Zeitschrift für Pädagogik/German Journal of Pedagogy, 51*(6), 765–773.

Kohut, G. F., Burnap, C., & Yon, M. G. (2007). Peer observation of teaching. *College Teaching*, 55(1), 19–25.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht [Multiple goals in math instruction]*. Münster, Germany: Waxmann.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*(3), 231–251.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research, 41*(25), 3559–3565.

Liang, J. (2015). Live video classroom observation: an effective approach to reducing reactivity in collecting observational information for teacher professional development. *Journal of Education for Teaching, 41*(3), 235–253.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131.

Macdonald, R. G., & Tatler, B. W. (2015). Referent expressions and gaze: Reference type influences real-world gaze cue utilization. *Journal of Experimental Psychology: Human perception and performance, 41*(2), 565–575.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & psychophysics, 2*(11), 547–552.

Masling, J. & Stern, G. (1969). Effect of the observer in the classroom, *Journal of Educational Psychology, 60*(5), 351–354.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*(3), 350–371.

Medley, D.M. & Mitzel, H.E. (1962). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of Research on Teaching* (pp. 247–328). Chicago, IL: Rand McNally.

Miner, M. H., Dowson, M., & Sterland, S. (2010). Ministry orientation and ministry outcomes: Evaluation of a new multidimensional model of clergy burnout and job satisfaction. *Journal of Occupational and Organizational Psychology, 83*, 167–188.

Muthén, B., & Muthén, L. (1998-2012). *Mplus* (Version 7.11). Los Angeles, CA: StatModel.

Newton, X.A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: a generalizability analysis. *Studies in Educational Evaluation, 36*, 1–13.

Petko, D., Waldis, M., Pauli, C., & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdedaktik. Ansätze der TIMSS 1999 Video Studie und ihrer schweizerischen Erweiterung [Methodological considerations about video-based research in mathematical didactics. Approaches of the TIMSS 1999 video study and its Swiss extension]. *Zentralblatt der Didaktik für Mathematik, 35*(6), 265–280.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7)*,* 3–14.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht - Unterricht aus der Perspektive von Lernenden und Beobachtern [Motivational support in math instruction – instruction from the learner and observer perspectives]*. Münster, Germany: Waxmann.

Rotting, M. (2001). *Parametersystematik der Augen- und Blickbewegungen fur arbeitswissenschaftliche Untersuchungen.* RWTH Aachen, Unpublished doctoral dissertation.

Samph, T. (1976). Observer effects on teacher verbal classroom behavior. *Journal of Educational Psychology, 68*(6), 736–741.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & Koenig, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*(14), 12–12.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Seidel, T., Prenzel, M. & Kobarg, M. (2005). *How to run a video study: Technical report of the IPN Video Study*. Münster, Germany: Waxmann.

Seidel, T., Prenzel, M., Duit, R., & Lehrke, M. (2003). *Technischer Bericht zur Videostudie "Lehr-Lern-Prozesse im Physikunterricht“ [Technical report of the video study “teaching and learning processes in Physics instruction“]*. Kiel, Germany: IPN.

Stigler, J. (1998). Video surveys: New data for the improvement of classroom instruction. In S. G. Paris & H. M. Wellmann (Eds.), *Global prospects for education. Development, culture and schooling* (pp. 129–168). Washington, DC: American Psychological Association.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology, 27*(2)*,* 165–175.

Tatler, B. W., Gilchrist, I. D., & Land, M. F. (2005). Visual memory for objects in natural scenes: From fixations to object files. *The Quarterly Journal of Experimental Psychology Section A, 58*(5), 931–960.

Tatler, B. W., Hirose, Y., Finnegan, S. K., Pievilainen, R., Kirtley, C., & Kennedy, A. (2013). Priorities for selection and representation in natural tasks. *Philosophical Transactions of the Royal Society B: Biological Sciences, 368*(1628), 20130066.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs. A practical guide to randomization tests.* Mahwah, NJ: Erlbaum.

van Gog, T., Paas, F., van Merrienboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology (Applied), 11*(4), 237–244.

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli, & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht [Practices and quality of instruction. Findings of an international and Swiss video study on math instruction]* (pp. 171–208). Münster, Germany: Waxmann.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*(1), 47–58.

Yarbus, A. L. (1967). *Eye movements and vision*. New York, NY: Plenum Press.

Table 1
Overview of Items Used in the Teacher and Student Questionnaires

Compare this lesson to normal lessons. Compared to other lessons, … (poles: much less/ much more)

Dimension: student negative emotions
- my students felt nervous/ I felt nervous
- my students were angry/ I was angry
- my students felt helpless/ I felt helpless

Dimension: student motivation
- my students were interested/ I was interested
- my students believed they could handle challenges/ I believed I could handle challenges

Dimension: student cognition
- my students were listening to me/ I was listening to my teacher
- my students were concentrating on the lesson/ I was concentrating on the lesson
- my students were thinking about the lesson topic/ I was thinking about the lesson topic

Dimension: student behavior
- my students were interacting with me/ I was interacting with the teacher
- my students were engaging with the lesson/ I was engaging with the lesson

Dimension: teacher negative emotions
- I was nervous/ the teacher was nervous
- I was worried/ the teacher was worried

Dimension: teacher motivation
- I was enthusiastic/ the teacher was enthusiastic
- I believed in my own abilities/ the teacher believed in his/her own abilities

Dimension: Teaching quality

Sub-dimension: student support
- I made sure that every student was learning/ the teacher made sure that every student was learning
- I took time to help students when they were stuck/ the teacher took time to help us when we were stuck

Sub-dimension: classroom management
- I got the students listen to me/ the teacher got us to listen to him/her
- I kept the class in order/ the teacher kept the class in order

Sub-dimension: cognitive activation
- I asked questions that made the students think/ the teacher asked questions that made us think
- I gave problems that can be solved in several different ways/ the teacher gave problems that can be solved in several different ways

Dimension: teaching practices
- I lectured/ the teacher lectured
- I asked questions/ the teacher asked questions

the students had whole-class discussion/ we had whole-class discussion
the students did work on their own/ we did work on our own
the students did work in groups/pairs/ we did work in groups/pairs
the students moved around the classroom/ we moved around in the classroom

Dimension: lesson preparation

I thought about the concept
I carefully developed the material for the lesson
I thought about the abilities of the students in preparing the lesson
I thought about the motivation of the students in preparing the lesson

Table 2

Descriptive Statistics of the Scales/Items Used in Student and Teacher Questionnaires

Scale/Item	Teacher Questionnaire		Student Questionnaire			
	M	*SD*	*M*	*SD*	α	ICC
Student negative emotions	3.24/2.90/2.90	.54/.30/.31	2.86	.60	.82	.06
Student motivation	3.10/3.05	.44/.67	3.21	.51	.84	.03
Student cognition	3.05/3.00	.59/.55	3.13	.49	.86	.05
Student behavior	3.10/3.24	.44/.48	3.08	.49	.84	.04
Teacher negative emotions	3.19/3.33	.51/.66	2.96	.47	.88	.04
Teacher motivation	2.95/2.95	.22/.24	3.14	.37	.52	.09
Teaching quality						
Student support	3.11/3.00	.32/.00	3.09	.42	.80	.06
Classroom management	3.00/3.00	.00/.00	3.08	.37	.84	.08
Cognitive activation	3.05/2.83	.23/.38	3.05	.40	.65	.03
Teaching practices						
Teacher lectured	2.89	.58	2.92	.12	–	.06
Teacher asked questions	3.16	.37	3.11	.12	–	.09
Had whole-class discussion	2.82	.53	2.98	.18	–	.11
Students did work on their own	3.00	.34	3.15	.44	–	.07
Students did work in groups/pairs	2.75	.79	2.91	.29	–	.18
Students moved around	2.65	.75	2.79	.17	–	.06
Lesson preparation			–	–	–	–
Thought about the concept	3.14	.36	–	–	–	–
Carefully developed the material	3.05	.50	–	–	–	–
Thought about the abilities of the students	3.14	.48	–	–	–	–
Thought about the motivation of the students	3.10	.30	–	–	–	–

Note.− = This scale/item was not assessed from the student perspective. For the teacher perspective, Cronbach's α is not displayed as the analyses were conducted on the item level. The values for teachers indicate means and standard deviations for the single items used.

Table 3

MIMIC Models For the Student Questionnaire Data

Dimension	Item loadings	β video condition (SE)	*p*	*r* dimensions
Teaching quality model				.81-.93
Classroom management	.64; .79	-.06 (.09)	.51	
Student support	.61; .80	-.15 (.10)	.14	
Cognitive activation	.74; .62	.00 (.08)	.98	
Teaching practices model	−	-.12-.05 (.04-.10)	.10-.62	-.06-.45
Student model				-.43-.88
Negative emotions	.65-.87	.19 (.07)	.01	
Motivation	.64-.68	-.04 (.06)	.55	
Cognition	.70-.82	-.11 (.06)	.04	
Behavior	.53-.79	-.15 (.05)	.01	
Teacher model				-.33
Negative emotions	.79; 1.00	.11 (.05)	.04	
Motivation	.43; 1.00	-.11 (.07)	.14	

Note.− = In this model, no latent variable was estimated; *r* dimensions = correlation between the dimensions included in the respective model. For the teaching quality model, a correlation between the error loadings of two single items of the dimensions student support and classroom management had to be set free ($r = .31$, $p < .01$).

Table 4

Descriptive Statistics for Durations of Teacher Gaze Towards Research and Non-Research Targets

Gaze Target	AOI no.	*M*	*SD*	Min.	Max.
Research	AOI 1	233.15	1067.01	.00	6707.00
Researcher	AOI 2	127.66	522.37	.00	2993.58
Camera	AOI 3	29.54	98.97	.00	593.25
Researcher And Camera	AOI 4	69.04	311.34	.00	2023.58
Research other	AOI 5	51.75	196.32	.00	1001.75
Non-Research	AOI 1	11862.18	1670.51	4481	13579
Students	AOI 2	4747.22	1088.54	2432.50	7220.25
Resources	AOI 3	6289.74	1463.23	2224.67	8941.58
School Staff	AOI 4	71.39	150.09	.00	802.58
Non-Research other	AOI 5	740.24	323.89	117.08	1587.25

Note. *M* = mean duration in seconds. For repeated measures ANOVA and paired samples t-tests, research–non-research pairings were made between targets (or areas of interest, AOI) sharing the AOI label. For example, the AOI2 comparison was between teacher gaze towards the researcher (Research AOI2) and that towards students (Non-Research AOI2).

Table 5

Interrupted Time Series Measures Regarding Each Potential Onset of Reactivity Depletion (ORD)

Gaze Target	ORD	Obs.	β	*SE*	*t*	*p*
Research	10 min	30	.00	.09	.00	1.00
	9 min	27	.00	.11	.00	1.00
	8 min	24	.00	.12	.00	1.00
	7 min	21	.00	.12	.000	1.00
	6 min	18	.000	.12	.003	1.00
	5 min	15	.002	.12	.016	.99
	4 min	12	.01	.12	.094	.93
	3 min	9	.06	.12	.511	.61
	2 min	6	.66	.27	2.43	.02
	1min 20s	**4**	**1.43**	**.28**	**5.07**	**<.001**
Non-Research	10 min	30	-.04	.61	-.06	.96
	9 min	27	-.42	.89	-.48	.64
	8 min	24	.94	.89	1.06	.30
	7 min	21	-.89	.89	-1.01	.32
	6 min	18	-.34	.89	-.39	.70
	5 min	15	.69	.89	.77	.45
	4 min	12	-.23	.89	-.26	.79
	3 min	9	-.40	.89	-.45	.66
	2 min	6	1.25	1.10	1.14	.26
	1min 20s	**4**	**-3.25**	**1.04**	**-3.12**	**.004**

Note. Obs = Observation number (cf. the time series graphs, Figure 4). SE = Standard Error.

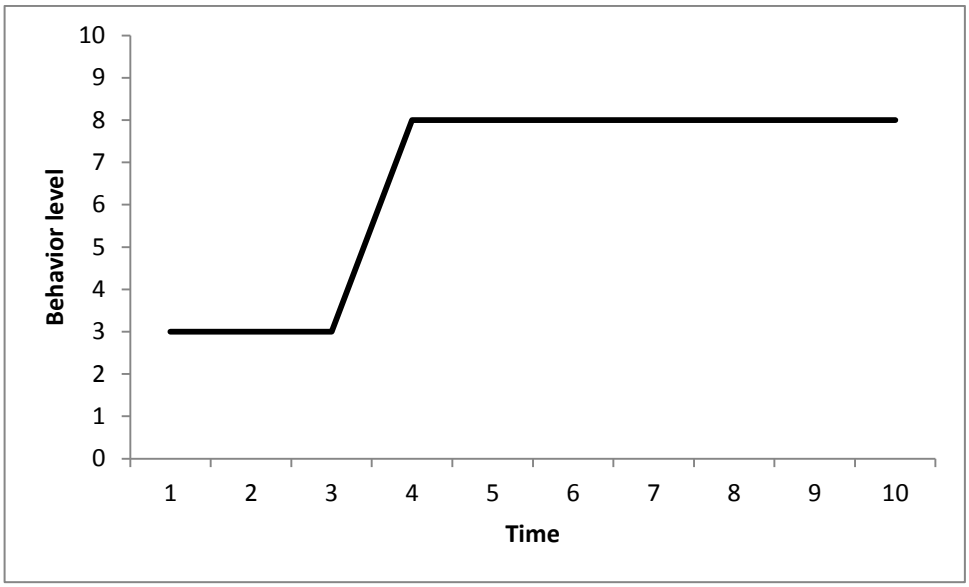

Figure 1. An example time series graph to illustrate interrupted time series analysis with the 'interruption' occurring at Time 3, which marks an abrupt change in behavior level.

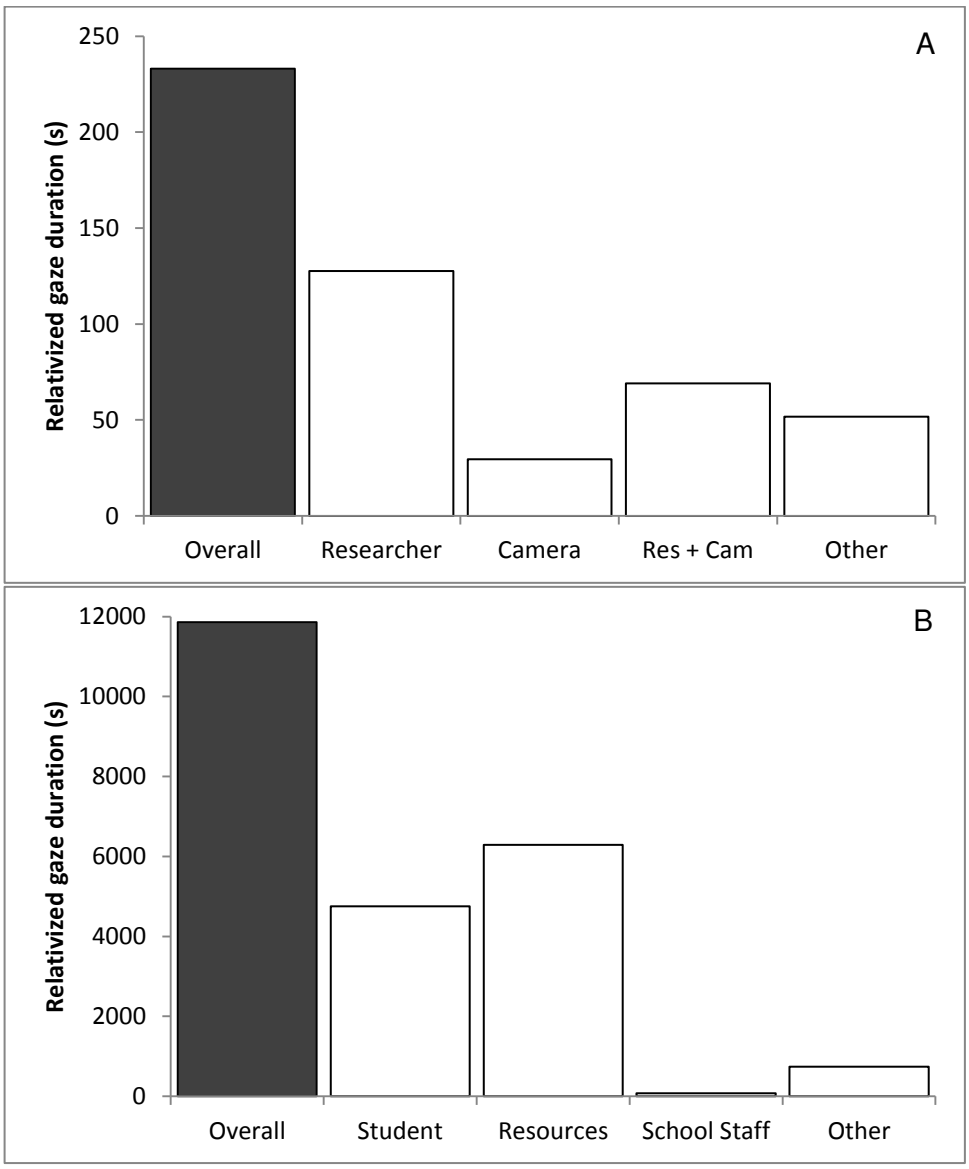

Figure 2. Relativized duration of teacher attention. Graph A shows research gaze; graph B shows non-research gaze.

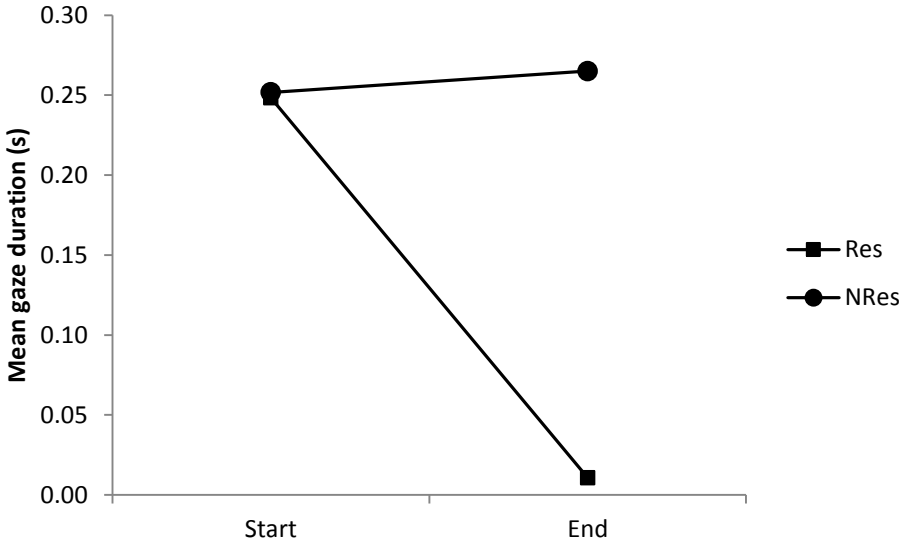

Figure 3. Start versus End comparisons of teacher gaze towards research and non-research classroom areas. Res = Research target gaze; NRes = Non-Research target gaze.

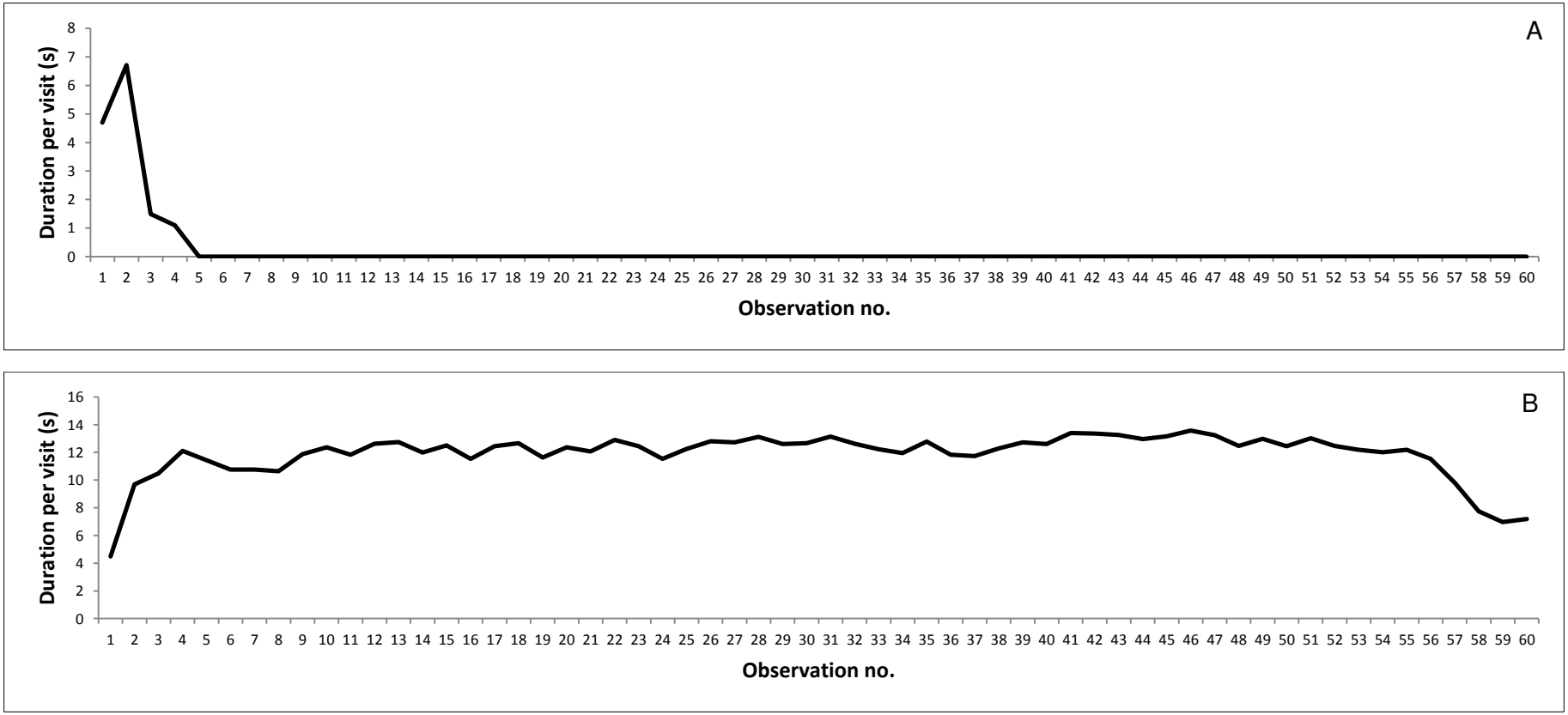

Figure 4. Time series comparisons of relativized gaze durations at research versus non-research classroom areas during the first 20 minutes. For analytic purposes, observations were taken every 20s. Graph A shows research target gaze; graph B shows non-research gaze.