Heartbeat to Data – Using Wearable Fitness Trackers as an Affordable Approach to Assess Teacher Stress

Mandy Klatt1, Christin Lotz1,Peer Keßler2,3, Gregor Kachel1 and Anne Deiglmayr1

1 Institute of Educational Sciences, Leipzig University

2 Institute of Psychology, University of Greifswald

3 German Center for Child and Adolescent Health (DZKJ), Greifswald

# Author Note

Preliminary findings from the research reported in this manuscript have been presented at a conference with published 600-words abstracts (GEBF conference 2023). We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Mandy Klatt, Division of Empirical School and Classroom Research, Institute of Educational Sciences, Leipzig University, Marschnerstr. 29, Leipzig, 04109, Germany. Email: [mandy.klatt@uni-leipzig.de](mailto:mandy.klatt@uni-leipzig.de)

# Abstract

Past research on physiological indicators of teacher stress often had to rely on expensive and obtrusive assessment methods. Modern fitness trackers represent a non-invasive and convenient alternative. This study investigated the use of wrist-worn fitness trackers to assess teacher heart rate (HR) as an indicator of stress during teaching. In a laboratory study, we used a Fitbit® fitness tracker to assess teachers´ HR before, during, and after a potentially stressful micro-teaching session. Our results demonstrate that the fitness tracker was indeed useful for mapping teachers’ stress, with the data showing how teachers’ HR increased before, peaked during, and progressively decreased after the micro-teaching session. Moreover, we related the fitness tracker data to retrospective teacher self-reports. We found that teachers’ subjective stress appraisals, together with their teaching experience, explained only small amounts of variance in HR data. We discuss the potential of fitness trackers as an affordable and ubiquitous assessment tool for research on teacher stress in the classroom and provide advice for practical implementation.

*Keywords:* teacher stress, fitness tracker, heart rate, classroom disruptions, wearable technology, physiological stress measurement

# Introduction

The teaching profession is one of the most stressful professions, with teachers facing a host of stressors during their everyday work (Herman et al., 2020; Schult et al., 2014; Smith, 2000). To better understand mechanisms in teacher stress, there is a growing research interest in physiological measures such as heart rate (HR) as online measures of teachers’ stress during teaching (Kärner & Höning, 2021; Wettstein et al., 2020). For example, it has been shown that teacher-centered activities and typical classroom-related stressors increase teacher HR during teaching activities (Donker et al., 2018; Huang et al., 2022; Junker et al., 2021; Scheuch & Knothe, 1997; Sperka & Kittler, 1995). However, previous studies have often relied on expensive and obtrusive electrocardiographs (ECG). Modern fitness trackers represent a non-invasive and convenient alternative (Ferguson et al., 2015).

Classroom disruptions are a major stressor in teachers’ daily work (Aloe et al., 2014; Boyle et al., 1995), and learning how to deal with them is an important aspect of professional expertise (Wolff et al., 2015). According to Lazarus (1990) transactional model of stress and coping, the experience of stress in response to stressors such as classroom disruptions depends on the teacher´s subjective appraisal, which, in turn, depends on their coping resources, such as their professional knowledge. The resulting stress response has a psychological, physiological, or behavioral dimension (Kyriacou & Sutcliffe, 1978). Therefore, in order to better understand how classroom stressors affect teachers’ stress response, subjective self-reports should be accompanied by objective, physiological measures (Wettstein et al., 2021). Teachers’ use of wrist-worn fitness trackers in educational research provides fine-grained, in vivo data, allowing researchers as well as teachers themselves to monitor their physiological stress response continuously during teaching, across settings, and at low costs. Being able to monitor, and eventually counteract, teacher stress levels appear particularly relevant given the profession’s generally high stress levels and associated negative health effects (Johnson et al., 2005; Montgomery & Rupp, 2005). To harness this potential, the present study explored the use of wrist-based fitness trackers as a tool to assess teachers’ HR, as an indicator of stress, before, during, and after a teaching session during which typical, potentially stressful, classroom disruptions occurred. Further, we explored to what extent teachers’ subjective appraisals of classroom disruptions and their teaching experience predicted teacher stress as assessed by the fitness tracker.

## Fitness trackers as a ubiquitous, low-cost tool for assessing physiological stress responses

Fitness trackers provide data on physical activity and cardiovascular parameters such as HR, supporting personalized fitness goals (Nuss et al., 2021) and stress management (Hao et al., 2017). They can be used as ubiquitous, low-cost, and unintrusive data collection instruments (Godfrey et al., 2018), and their wide-spread use and everyday availability align with the increasing popularity and acceptance of wearables among the general population (Peng et al., 2022). In contrast to self-reported questionnaires on stress (Chaplain, 2008; Liu & Yan, 2020) that are prone to biases like social desirability (Razavi, 2001) or recall errors (Van den Bergh & Walentynowicz, 2016), fitness trackers, as ambulatory assessment methods (Trull & Ebner-Priemer, 2013; Wettstein et al., 2020), offer more objective insights into teachers’ stress levels by monitoring teachers’ physiological stress responses without disrupting teaching (Donker et al., 2018; Runge et al., 2020).

One important health parameter assessed by nearly all wrist-worn fitness trackers is heart rate (Scalise & Cosoli, 2018). HR indicates the number of heartbeats within one minute and is typically expressed as beats per minute (BPM; Berntson et al., 2007; Hottenrot, 2007). At rest, the average HR of adults typically ranges between 60 and 80 BPM (Sammito et al., 2015). HR can be detected and measured in different ways using sensors, such as electrocardiography (ECG) or photoplethysmography (PPG; Mukhopadhyay & Islam, 2017). While ECG sensors offer precise measurements by detecting the heart’s electrical activity, their intrusive nature and requirement of direct skin contact may limit their suitability (Kranjec et al., 2014), particularly in educational settings. In contrast, photoplethysmography (PPG) is a rather uncomplicated and inexpensive technique to measure HR, commonly found in commercially available fitness trackers (Castaneda et al., 2018). This optical method assesses HR by flashing green or red lights to measure changes in blood volume in the capillaries of the skin (Allen, 2007).

Physiologically, HR is regulated by the sympathetic and parasympathetic nervous systems (Pham et al., 2021). An increase in sympathetic activity results in HR being sped up (“fight or flight” response; Taelman et al., 2009), whereas an increase in parasympathetic activity results in HR being slowed down (“rest and digest” response; Battipaglia & Lanza, 2015). Stress or mental and physical strain directly increases HR (Custodis et al., 2014; Sachs, 2014). Therefore, an increase in HR can be regarded as an indicator of increasing stress, and a decrease as an indicator of relaxation and ease (Kyriacou & Sutcliffe, 1978). Thus, fitness trackers offer low-cost and unobtrusive access to psychological stress data.

## HR in teaching-learning contexts

Prior research, typically using ECG methods, has shown that changes in teachers’ HR can be mapped onto stressors they experience during teaching. For example, teachers’ HR tends to increase when teachers take an exposed position in student-teacher interaction (Donker et al., 2018; Junker et al., 2021; Scheuch & Knothe, 1997; Sperka & Kittler, 1995). Sperka & Kittler (1995) for example recorded the HR of 16 pre-service teachers during their first lesson and showed that teachers’ HR increased significantly during teaching. The activation was particularly prominent at the beginning of the lesson and decreased over the course of the lesson. The authors suggested that pre-service teachers’ proactive coping strategies, such as actively managing student interactions, helped lower their HR levels. Other ECG studies identified typical stressors predicting increases in HR, such as class size (Huang et al., 2022), or low student engagement and motivation (Junker et al., 2021). Junker et al. (2021) recorded the HR of 40 teachers during a real classroom lesson. Again, teacher stress, induced by factors such as low student engagement (e.g., lack of motivation or interest in tasks) or teacher-centered activities (e.g., teacher-focused classroom activities), resulted in elevated HR.

More recent studies have begun using wrist-worn fitness trackers to investigate HR trends in instructional settings (Chalmers et al., 2021; Darnell & Krieg, 2019). Darnell & Krieg (2019) measured the HR of 15 medical college students listening to lectures, using wrist-worn devices. The analysis revealed a constant decrease in HR throughout the lecture, with HR peaks during more interactive learning phases. Chalmers et al. (2021) used HR data from a fitness tracker to identify physiological changes during stress-inducing tasks (i.e., the Trier Social Stress Test; Kirschbaum et al., 1993). Average HR increased significantly from the resting to the stress inducing phases. Even though the participants of these previous studies (Chalmers et al., 2021; Darnell & Krieg, 2019) were not teachers but learners, the results are relevant for studying teacher stress as they demonstrated that HR can be effectively recorded using fitness trackers over the course of a whole learning unit, and HR changes are in line with the occurrence of activating or stress-inducing tasks.

To the best of our knowledge, only one study has directly assessed teachers’ HR during teaching using a wrist-worn fitness tracker: Runge et al. (2020) assessed HR as an indicator of stress in four in-service teachers during authentic lessons. They used fitness trackers’ recordings to create a profile for each teacher, with the aim of differentiating between teachers reporting higher vs. lower levels of stress. It was found that the combination of a high HR, a high number of steps, and short sleep duration was characteristic for teachers reporting high stress levels. However, it should be noted that the generalizability of these results is limited due to the small sample size.

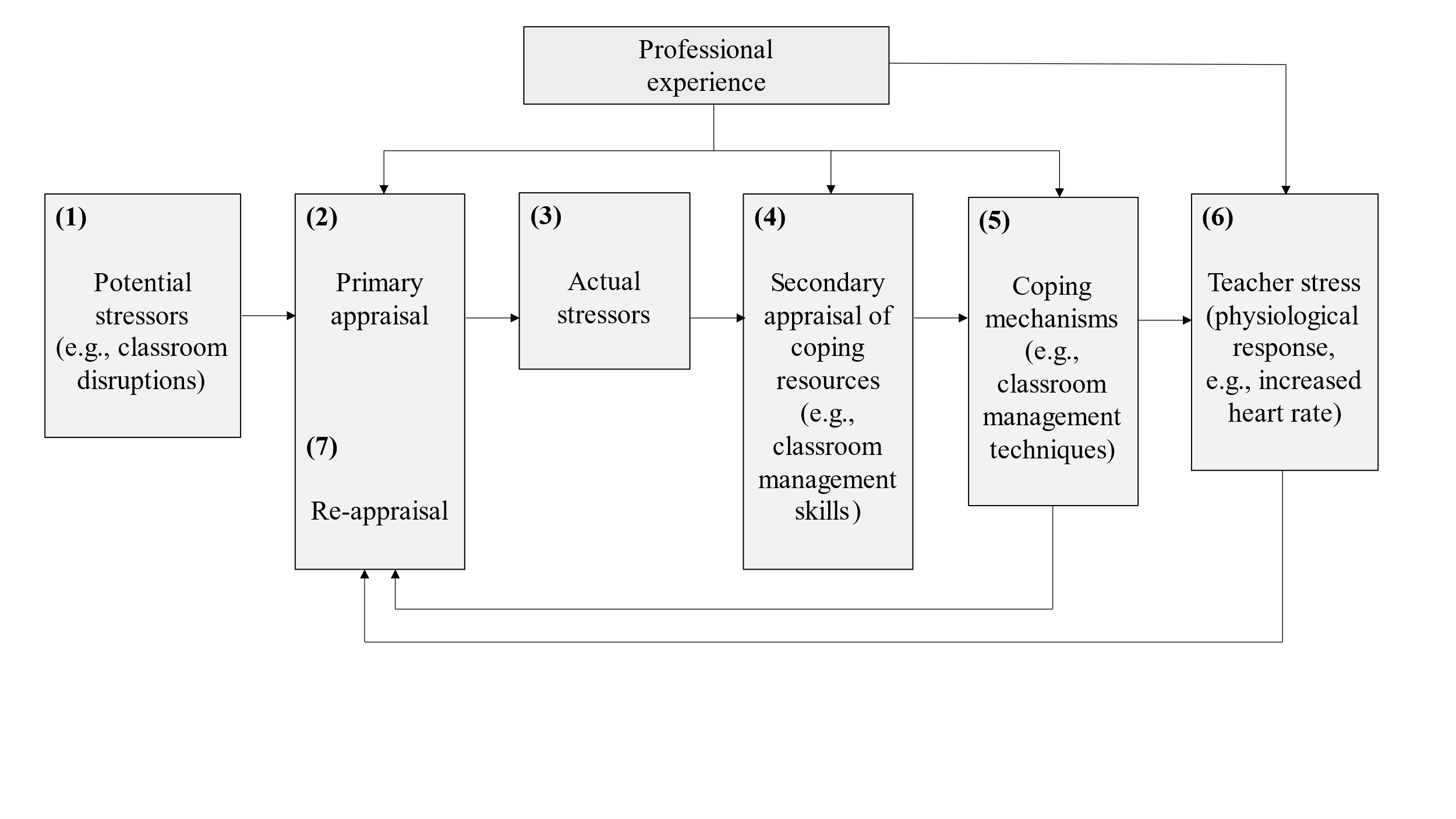
In summary, previous studies have revealed that teachers’ (and students’) HR changes depend on their activities and the stressors they experience, with an increase in HR already before the expected stressors occur, and with peaks in activating phases (Chalmers et al., 2021; Darnell & Krieg, 2019). For teachers, teacher-centered phases led to an increase in HR (Donker et al., 2018; Junker et al., 2021; Scheuch & Knothe, 1997; Sperka & Kittler, 1995). However, there is a lack of studies using teacher-worn fitness trackers in larger samples, exploring the feasibility of this convenient tool for researching links between teachers’ HR and stress inducing events and mechanisms.

## A model of teacher stress

According to Kyriacou & Sutcliffe (1978), teacher stress can be defined as a negative affective response, typically accompanied by physiological changes such as increased HR, triggered by job-related demands, and perceived as threatening to one’s self-esteem or well-being. Coping mechanisms help to reduce the perceived threat. Kyriacou’s definition of teacher stress (see Kyriacou & Sutcliffe, 1978; and, for a more recent adaptation, Van Dick, 2006) is based on the transactional stress model (Lazarus, 1966, 1990), which highlights the interaction between an individual and the environment, whereby stress refers to a person’s subjective reaction to an event (a stressor) that exceeds the person’s adaptive resources.

**Figure 1**

*A Model Of Teacher Stress (Adapted From Van Dick 2006, p.37, Modified by The Authors)*

 Figure 1 shows, in a simplified way, how classroom events affect teachers’ stress level, according to the adaptation of the Lazarus model to teacher stress proposed by Van Dick (2006): When potential stressors (e.g., classroom disruptions) occur during teaching (1), teachers intuitively judge how disruptive the event is in a primary appraisal (2). If potential stressors are judged as threatening, i.e., as actual stressors (3), teachers consider whether they have sufficient resources for coping with the stressors (4). Teachers utilize these resources in trying to cope with the stressors, e.g., by employing classroom management strategies (5). In cases when coping fails, stress ensues, often accompanied by physiological reactions like increased HR (6). As part of the coping process, and dependent on its outcomes, teachers re-evaluate the stressor (7).

As shown in Figure 1, teachers’ primary and secondary appraisal as well as coping attempts are influenced by professional experience. As professional experience grows, teachers develop cognitive scripts for managing classroom events, resulting in more complex and problem-focused classroom management skills (Wolff et al., 2021), and thus more effective coping and less stress. Empirically, classroom management skills and problem-focused coping styles are linked to fewer instances of emotional exhaustion (Clunies-Ross et al., 2008; Maslach et al., 2001). Novices in the teaching profession, on the other hand, face considerable stress and often feel overwhelmed by the demands of teaching (Klusmann et al., 2012; Ophardt & Thiel, 2017; Wolff et al., 2015) with many leaving the profession within the first five years (Ingersoll & Smith, 2003). Accordingly, when resources are lacking and coping fails, negative consequences for health (e.g., burnout) and for work (e.g., high turnover rates) can arise (Aloe et al., 2014; Jalongo & Heider, 2006; Unterbrink et al., 2007), highlighting the importance of professional expertise for managing teacher stress (Fisher, 2011).

## Present Study

The present study aimed to explore the relations between teachers’ HR response, and their subjective appraisals of stress during a micro-teaching unit, and to relate their self-reported appraisals and physiological stress responses to their teaching experience. We analyzed data from in-service and pre-service teachers who participated in a laboratory study as part of a larger project targeting the development of classroom management skills. Participants came to the lab individually and taught a short lesson to a class of three actors (i.e., trained student assistants) who performed several typical and possibly disruptive classroom events. The micro-teaching unit was thus potentially stressful for the participants

The aims of the present study were twofold:

1. The first research goal was to investigate whether HR measures assessed by a wrist-based fitness tracker were a suitable and effective method for mapping teachers’ HR over the course of the lab study, with a total duration of approximately 2 hours, including phases before, during, and after the stressful micro-teaching unit.

Looking at HR measures globally, we expected the participants to show an initial increase in their HR, followed by a peak during the micro-teaching unit and a decrease for the remaining phases. In addition, we examined whether z-standardization of the participants’ HR could serve as a useful method to account for individual differences in baseline HR: We expected to observe the same trends in both standardized and non-standardized HR values.

In addition, we selected five representative 10-minute intervals from the five phases of the lab study (see Figure 2) in order to test the hypotheses that, regarding HR levels, teachers’ HR would be the highest during the micro-teaching unit, compared to all other phases (\*\*Hypothesis 1a\*\*), and, regarding HR slopes, that teachers’ HR would increase while they were preparing for teaching (*pre-teaching interval*), but decrease in all of the following intervals, i.e. when they were habituating to and recovering from the stressful micro-teaching unit (\*\*Hypothesis 1b\*\*).

1. We further explored whether teaching experience made a difference in how teachers’ HR reacted to the classroom disruptions. We expected more experienced teachers to be less stressed by the classroom events (\*\*Hypothesis 2a\*\*). In addition, we examined the relations between teachers’ subjective appraisals of the classroom events (specifically, the disruptiveness of the events, and their confidence in dealing with them) and teachers’ HR level, beyond the explanatory power of teaching experience. We expected higher HR levels for teachers who felt more disrupted, regardless of their teaching experience (\*\*Hypotheses 2b\*\*), and lower HR levels for teachers who felt more confident, regardless of teaching experience (\*\*Hypothesis 2c\*\*). We hypothesized that each of the three predictors (teaching experience, disruption appraisal, confidence appraisal) uniquely contributed to explaining variance in teachers’ HR levels (\*\*Hypothesis 2d\*\*). In addition, we exploratively ran analogous analyses for the changes in HR (i.e., slopes).

# Method

## Participants

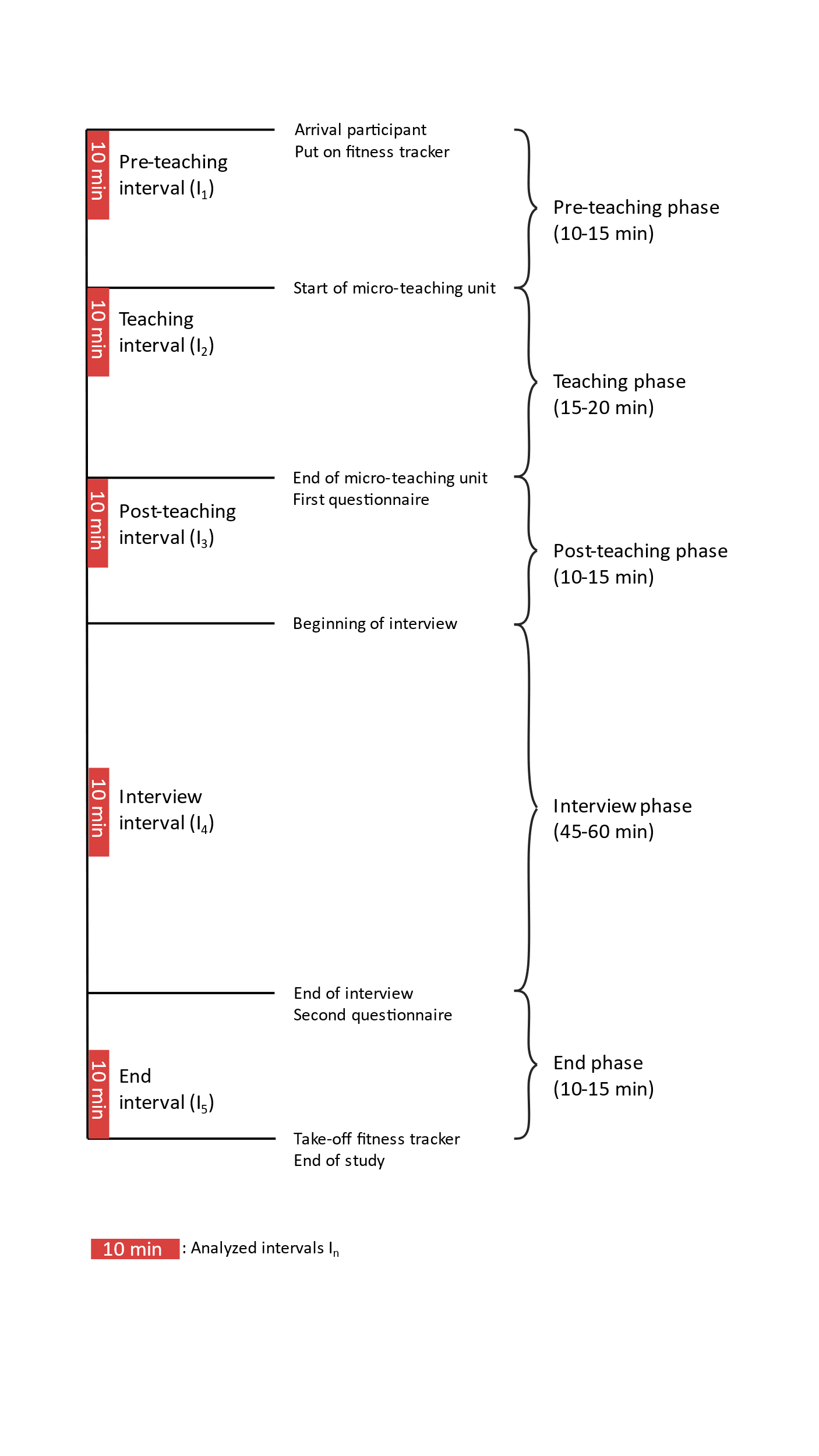
The sample consisted of *N* = 84 pre- and in-service teachers from Germany, who were recruited via personal contacts, email lists, and flyers. The data of three participants was lost due to failed data transmission, yielding an analysis sample of ntotal = 81 (ntotal = 52 women, ntotal = 29 men), including 40 pre-service and 41 in-service teachers. Participants had a mean age of 30.95 years (*SD* = 10.90; range: 19-60) and an average teaching experience of 5.64 years (*SD* = 9.46; range: 0-37).

## Setting and Procedure

The study was carried out following the ethical standards and the approval of the University’s Institutional Review Board. All participants were informed in detail about the aims of the study prior to testing. Participation was voluntary, not incentivized, and only took place after written consent had been given.

**Figure 2**

*Procedure of The Two-Hour Study, Consisting of Five Phases With Five Representative 10-Minute Intervals as The Basis of Our Analysis.*

**

Each participant came to the lab for a period of approximately two hours in total, and each participant underwent the same phases (see Fig. 2): In the pre-teaching phase, the experimenter welcomed the participants and helped them put on the fitness tracker. This was followed by a warm-up session to familiarize the participants with the laboratory setting and the class. This phase took about 10-15 minutes and participants spent this time mostly standing or slowly walking around. During the teaching phase, the participants held their self-prepared micro-teaching unit to a class of three trained actors who performed nine, potentially disruptive, classroom events (e.g., chatting with a neighbor, heckling, looking at the phone; see Table ## in the supplementary material for an overview and categorization of all events; and Fig## in the supplementary material for a depiction of the laboratory setting of the micro-teaching unit). The topic and class level of the teaching unit could be freely chosen by the teachers with the only requirement that the unit had to be an introductory lesson, and had to consist of supervised individual work and / or frontal teaching. The micro-teaching unit lasted about 15-20 minutes. Participants spent this time mostly standing or slowly walking around. While teaching, participants wore eye-tracking glasses, and their lesson was video-recorded. After having completed the micro-teaching unit, in the post-teaching phase, participants filled in questionnaires for approximately 10-15 minutes: a brief computer-based survey of sociodemographic data (e.g., teaching experience, gender, studied school type, studied school subjects, extracurricular teaching activities), and a short knowledge test that was irrelevant to the present study. In the interview phase, participants engaged in a Stimulated Recall Interview (SRI). During the SRI, participants sat in front of a computer monitor and watched the video of their own lesson from the ego perspective, as recorded through the eye-tracking glasses. The experimenter stopped the video each time one of the nine classroom events happened, and asked five open-ended, and three rating questions per event. Two of the rating questions are relevant to the present study: the disruption and the confidence appraisal ratings (see Measures). The interview lasted about 45-60 minutes. Finally, in the end phase, participants filled in another questionnaire irrelevant to the present study, which lasted about 10-15 minutes.

## Measures

### Heart Rate Data and Heart Rate Intervals

To measure teachers’ HR, we used the wrist-based fitness tracker Fitbit® Charge 4. In line with the manufacturer's instructions [@fitbitnd], the device was attached to the participants’ nondominant hand, a finger’s width above the wrist bone. The tracker works by flashing green LEDs hundreds of times per second, using light-sensitive photodiodes to catch the reflected light, to calculate the volume changes in the capillaries. From this, the tracker calculated the heart beats per minute. HR measurements were generated at least every 15 seconds. The raw data contained the estimated HR in BPM for each time stamp. To account for individual differences in the baseline HR, we also calculated z-standardized HR values based on individual means, i.e., at the subject level of n = 81 participants (standardized HR).

Since we aimed to keep measurement intervals comparable between study phases, we aggregated HR over a representative 10-minute interval within each phase (cf. Fig. 2). Previous research has indicated that 10-minute intervals are a useful duration for analyzing PPG data [@lu2008can]. The intervals were selected based on the following rules: The *pre-teaching interval (I1)* comprised the first 10 minutes after the fitness tracker had been put on. The *teaching interval (I2)* started two minutes after the lesson had started. This interval was of the highest relevance to our study. We explicitly chose an early 10-minute interval within the teaching phase, as previous studies revealed that the beginning of a lesson is most demanding and potentially stressful with regards to teacher-student interaction [@donker2018quantitative; @claessens2017positive]. The *post-teaching interval (I3)* started immediately after the end of the teaching unit. The *interview interval (I4)* was defined as the mid-10 minutes between the end of the teaching unit and the time point when the fitness tracker was taken off. All participants were being interviewed during this interval. The *end interval (I5)* comprised the last 10 minutes before the fitness tracker was taken off.

### Teaching Experience

Participants’ teaching experience was assessed as a part of their sociodemographic data. Participants stated their work experience in years.

### Subjective appraisal of the classroom events and coping processes

The subjective disruption and confidence appraisals were assessed during the SRI on an 11-point rating scale, ranging from 0 (not at all disrupting/confident) to 10 (extremely disrupting/confident). Ratings were averaged across the nine classroom events for each participant, as we were interested in the general stressfulness of the teaching phase for each participant.

## Data analysis

To enable visual inspection of HR trends, we displayed smoothed teacher HR over the course of the recording.[[1]](#footnote-2) We visually compared unstandardized and standardized HR trends over the two-hour recording period.[[2]](#footnote-3) For all further analyses, we used standardized rather than unstandardized HR values.

We averaged each person’s standardized HR over each of the five selected intervals[[3]](#footnote-4), resulting in one measure per person per interval. To test Hypothesis 1a, we initially conducted a one-way ANOVA with repeated measures as an omnibus test and then tested the mean differences between the *teaching interval* (I2) and the other four intervals by planned contrasts and inspection of effect size *d* [@cohen1988new].

For testing Hypothesis 1b, concerning HR changes within each interval, we first conducted a linear estimation of the increase or decrease in standardized HR values over time for each participant. To this end, we used fixed intercept fixed slope regression models [@gelman2006data] for each interval to estimate intercepts and linear slopes for each individual, which were then averaged across individuals.[[4]](#footnote-5) We tested Hypothesis 1b based on the unstandardized estimates of mean slopes (one estimate per participant per interval).

Addressing our second research goal, we ran linear regression analysis with teaching experience and subjective appraisals as predictors. To test Hypothesis 2a, we examined the effect of teaching experience on participants’ HR levels (i.e., mean standardized HR) for each of the five intervals, using linear regression models with teaching experience as the sole predictor. To test Hypotheses 2b and 2c, we separately augmented the model by either teachers’ disruption appraisal (Hypothesis 2b) or confidence appraisal (Hypothesis 2c) as predictors, while controlling for teaching experience. To test Hypothesis 2d, we examined the effects of all three predictors in one regression model. Furthermore, we repeated these steps to explore the effects of teaching experience and subjective appraisals on *changes* in teachers’ HR (i.e., mean slopes). Please note: HR levels and changes were not regressed on the disruption and confidence appraisals in the *pre-teaching interval* (I1), because the appraised classroom events had not yet taken place in that phase.

# Results

## 3.1 Mapping teachers’ HR over the course of the study phases

Means, standard deviations, and range of teachers’ unstandardized and standardized HR for the entire study period, and for the five intervals, are shown in Table 1. Fig. 3 a. and b. display the unstandardized and standardized HR trends, respectively, over the course of the entire study period. HR initially increased, peaked, and then decreased, with the unstandardized and standardized HR graphs showing high similarity. Thus, for all further analyses, we used participants’ standardized HR values.

**Table 1**

*Mean HR (*M*), Standard Deviations HR (*SD*), And Range of Teachers’ HR Over The Course of The Entire Study And The Five Intervals (Unstandardized in BPM/z-Standardized)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Interval | *M* HR | *SD* HR | Min | Max | |
| Overall Course of 2h | 90.09/0.041 | 15.76/0.991 | 512/-4.03 | 164/4.56 | |
| *Pre-teaching interval (I1)* | 96.28/0.48 | 14.11/0.88 | 56/-3.56 | 139/3.24 | |
| *Teaching interval (I2)* | 100.80/0.85 | 16.23/0.77 | 63/-2.18 | 164/4.37 | |
| *Post-teaching interval (I3)* | 93.61/0.27 | 14.01/0.76 | 60/-2.17 | 150/3.06 | |
| *Interview interval (I4)* | 82.32/-0.72 | 11.85/0.74 | 51/-2.51 | 132/4.39 | |
| *End interval (I5)* | 77.95/-1.07 | 11.14/0.57 | 502/-2.68 | 120/2.96 | |
| 1 Please note that standardized *M* and *SD* of the overall course were not exactly 0 and 1 due to rounding differences.  2 Deviations of the minimum values in the overall course vs. the *end interval (I5)* are due to data of a few participants who needed more than two hours to finish the study. | | | | |

**Figure 3**

*Overall Course of The HR With The Unstandardized HR in BPM Shown in Fig. 3a. And The z-Standardized HR Shown in Fig. 3b. Over The Course of The 2-Hour Study*

Ein Bild, das Text, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

Figure 4 shows the distribution of teachers’ mean standardized HR for the five intervals. Repeated measures ANOVA revealed significant differences in mean standardized HR between intervals, *F*(4, 400) = 260.62, *p* < .05, *f* = 1.60 (large effect). Planned contrasts indicated that, as hypothesized (Hypothesis 1a), mean standardized HR was significantly higher in the *teaching interval* *(I2)* than in all other intervals, specifically, the *pre-teaching interval* (*I1*; *t*(400) = -10.08, *p* < .05, *d* = 1.03; large effect), the *post-teaching interval* (I3; *t*(400) = -6.94, *p* < .05, *d* = 1.37; large effect), the *interview interval* (*I4*; *t*(400) = 15.00, *p* < .05, *d* = 3.29; large effect), and the *end interval* (*I5*; *t*(400) = 22.54, *p* < .05, *d* = 4.64; large effect).

**Figure 4**

*Distribution of The Standardized Heart Rate Means in The Five Intervals*

*Ein Bild, das Text, Diagramm, technische Zeichnung, Plan enthält.

Automatisch generierte Beschreibung*

Next, we examined HR changes (i.e., mean slopes) within each interval to test the hypothesis that HR increased during the *pre-teaching phase* and decreased during all other phases (Hypothesis 1b). The mean intercepts and mean slopes, complemented by their standard deviations for each interval, are shown in Table 2. The mean slope of the *pre-teaching interval* *(I1)* was significantly positive, indicating an increase in HR, as hypothesized. Further, the mean slopes of the *teaching interval (I2)*, *post-teaching interval (I3)*, and *interview interval* *(I4)* were significantly negative, indicating a decrease in HR. For the last interval, the *end interval* *(I5)*, the mean slope was negative, but did not differ significantly from zero.

**Table 2**

*Analysis* *(*M, SD, P*-Values)* *For The Mean Intercepts And The Mean Slopes For The Five Intervals*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Interval | *M (SD)* | | *p* | |
|  | Intercept | Slope | Intercept | Slope |
| *Pre-teaching interval (I1)* | 0.052  (0.820) | 0.085\* (0.133) | .57 | < .05 |
| *Teaching interval (I2)* | 1.025\* (0.690) | -0.039\* (0.108) | < .05 | < .05 |
| *Post-teaching interval (I3)* | 0.549\* (0.547) | -0.060\* (0.101) | < .05 | < .05 |
| *Interview interval (I4)* | -0.617\* (0.614) | -0.022\* (0.070) | < .05 | < .05 |
| *End interval (I5)* | -1.004\* (0.500) | -0.012 (0.074) | < .05 | .14 |

## Predicting mean standardized HR and mean slopes

Table 3 shows the raw correlations among mean standardized HR/mean slopes (see Table 2 for means and standard deviations), teaching experience (*M* = 5.64, *SD* = 9.46), disruption appraisal (*M* = 5.19, *SD* = 2.87), and confidence appraisal (*M* = 7.81, *SD* = 1.97). With a few notable exceptions, correlations with HR measures were mostly very small and statistically non-significant (Table 3). Correlations between teaching experience and appraisals (not shown in Table 3) were substantial: more experienced teachers gave lower disruption appraisals (*r*=-.36), and higher confidence appraisals (*r*=.44). Moreover, the two appraisal variables were negatively correlated (*r*=-.37).

Table 4 shows the results of the regression analyses. Teaching experience significantly predicted mean standardized HR only in the *interview interval* (Table 4, Interview interval, Model 1), indicating a higher mean standardized HR for teachers with more teaching experience. This relationship is, in fact, in the opposite direction as predicted by Hypothesis 2a. Neither adding disruption appraisal (\*\*Hypothesis 2b\*\*) nor adding confidence appraisal (\*\*Hypothesis 2c\*\*) increased the amount of explained variance to a statistically significant extent.

When considering the effects of the three predictors in concert (Hypothesis 2d), mean standardized HR was significantly predicted only by disruption appraisal, and only in the *post-teaching interval* (Table 4, Post-teaching interval, Model 4), indicating a higher mean standardized HR for teachers who felt more disrupted by the classroom events, when controlling for the other variables.

Concerning the explorative investigation of the effects of teaching experience and subjective appraisals on *changes* (i.e., mean slopes) in teachers’ HR, teaching experience significantly predicted the mean slope in the *pre-teaching interval* (Table 4, Pre-teaching interval, Model 1), indicating a less steep HR increase in teachers with more teaching experience. For all other intervals, no variable had significant predictive value.

**Table 3**

*Correlations Between Mean Standardized HR/Mean Slopes And The Predictor Variables Teaching Experience (TE), Disruption Appraisal (DA), And Confidence Appraisal (CA), For The Five Intervals*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Pre-teaching interval (I1) | Teaching  Interval (I2) | Post-teaching  Interval (I3) | Interview  Interval (I4) | End  Interval (I5) |
| Teaching Experience | − .17/− .27\* | .11/−.02 | − .04/−.03 | .24\*/−.20 | .04/.11 |
| Disruption Appraisal | − .01/.16 | − .20/.08 | .20/−.14 | − .13/.01 | .04/.12 |
| Confidence Appraisal | − .10/− .18 | .06/.09 | .04/−.03 | .09/−.19 | − .07/.13 |
| *Note.* \* *p* < .05. | | | | | |

**Table 4**

*Standardized Regression Coefficients of Mean Standardized Heart Rate And Mean Slopes Predicted by Teaching Experience, Disruption Appraisal, And Confidence Appraisal For The Five Intervals*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | *Dependent variable: mean standardized HR and mean slopes* | | | | | | | | | | | | | | | |
|  | Model 1 | | | | Model 2 | | | | Model 3 | | | | | Model 4 | | | |
|  | Mean std. HR | | Mean slopes | | Mean std. HR | | Mean slopes | | Mean std. HR | | Mean slopes | | | Mean std. HR | | Mean slopes | |
|  | β (SE) | *p* | β (SE) | *p* | β (SE) | *p* | β (SE) | *p* | β (SE) | *p* | | β (SE) | *p* | β (SE) | *p* | β (SE) | *p* |
| ***Pre-teaching interval (I1)*1** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| Teaching  Experience | -.17  (.005) | .12 | -.27\*  (.002) | <.05 |  |  |  |  |  |  | |  |  |  |  |  |  |
| R2 | .030 |  | .071 |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| ***Teaching interval (I2)*** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| Teaching  Experience | .11  (.002) | .34 | -.02  (.001) | .83 | .04  (.005) | .73 | .01  (.001) | .96 | .10  (.006) | .42 | | -.08  (.001) | .54 | .05  (.006) | .67 | -.05  (.001) | .72 |
| Disruption  Appraisal |  |  |  |  | -.18  (.041) | .13 | .08  (.010) | .50 |  |  | |  |  | -.19  (.042) | .13 | .12  (.010) | .34 |
| Confidence  Appraisal |  |  |  |  |  |  |  |  | .01  (.046) | .92 | | .12  (.011) | .34 | -.04  (.047) | .76 | .15  (.012) | .24 |
| R² | .012 |  | .000 |  | .040 |  | .015 |  | .012 |  | | .010 |  | .042 |  | .031 |  |
| ∆ R² |  |  |  |  | .028 |  | .015 |  | .000 |  | | .010 |  | .030 |  | .031 |  |
| **Post-teaching interval (I3)** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| Teaching  Experience | -.04  (.005) | .70 | -.03  (.001) | .80 | .04  (.005) | .76 | -.09  (.001) | .44 | -.08  (.006) | .55 | | -.02  (.001) | .89 | -.01  (.006) | .91 | -.07  (.001) | .61 |
| Disruption  Appraisal |  |  |  |  | .22  (.040) | .07 | -.18  (.009) | .14 |  |  | |  |  | .25\*  (.041) | <.05 | -.20  (.010) | .12 |
| Confidence  Appraisal |  |  |  |  |  |  |  |  | .08  (.045) | .55 | | -.03  (.011) | .83 | .14  (.046) | .27 | -.08  (.011) | .54 |
| R2 | .002 |  | .001 |  | .043 |  | .020 |  | .006 |  | | .002 |  | .058 |  | .023 |  |
| ∆ R2 |  |  |  |  | .041 |  | .019 |  | .004 |  | | .001 |  | .056 |  | .022 |  |
| **Interview interval (I4)** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| Teaching  Experience | .24\*  (.006) | <.05 | -.20  (.001) | .07 | .22  (.006) | .06 | -.23  (.001) | .06 | .25\*  (.006) | <.05 | | -.14  (.001) | .25 | .23  (.007) | .07 | -.17  (.001) | .18 |
| Disruption  Appraisal |  |  |  |  | -.05  (.045) | .66 | -.08  (.006) | .52 |  |  | |  |  | -.06  (.047) | .61 | -.12  (.007) | .34 |
| Confidence  Appraisal |  |  |  |  |  |  |  |  | -.02  (.050) | .85 | | -.13  (.007) | .29 | -.04  (.052) | .76 | -.16  (.007) | .20 |
| R2 | .058 |  | .040 |  | .060 |  | .050 |  | .058 |  | | .054 |  | .061 |  | .069 |  |
| ∆ R2 |  |  |  |  | .002 |  | .010 |  | .000 |  | | .014 |  | .003 |  | .029 |  |
| **End interval (I5)** |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  |  |
| Teaching  Experience | .04  (.004) | .70 | .11  (.001) | .32 | .07  (.005) | .58 | .18  (.001) | .13 | .09  (.005) | .46 | | .07  (.001) | .58 | .10  (.005) | .43 | .12  (.001) | .33 |
| Disruption  Appraisal |  |  |  |  | .06  (.035) | .60 | .19  (.007) | .12 |  |  | |  |  | .04  (.037) | .76 | .23  (.007) | .07 |
| Confidence  Appraisal |  |  |  |  |  |  |  |  | -.11  (.039) | .38 | | .10  (.008) | .43 | -.10  (.041) | .44 | .16  (.008) | .22 |
| R2 | .002 |  | .013 |  | .005 |  | .053 |  | .012 |  | | .025 |  | .013 |  | .078 |  |
| ∆ R2 |  |  |  |  | .003 |  | .040 |  | .010 |  | | .012 |  | .011 |  | .065 |  |
|  | *Note*. In Model 1, mean standardized HR and mean slopes were predicted only by teaching experience. In Model 2, solely disruption appraisal was added to teaching experience as a predictor. In Model 3, solely confidence appraisal was added to teaching experience as a predictor. In Model 4, all three predictors were considered in concert.  1 We calculated only Model 1 for the pre-teaching interval because the classroom events had not yet occurred in this interval.  \* *p* < .05. | | | | | | | | | | | | | | | | |

# Discussion

## Key Findings

Overall, our findings indicate that wrist-worn fitness trackers are a useful tool for tracking teachers’ HR and identifying stressful periods during teaching. Using HR data from a commercially available and relatively low-cost Fitbit® fitness tracker, we were able to map teachers’ HR before, during, and after a stressful micro-teaching unit, with HR increasing in preparation for teaching, peaking during the teaching phase, and decreasing afterward.

These findings are in line with prior studies showing that teachers’ HR varies depending on their activities and encountered stressors with increases during phases where teachers are in an exposed position [@sperka1995; @scheuch1997psychophysische; @donker2018; @junker2021], as well as with findings showing how HR changes align with activating events and stress-inducing tasks [@Darnell2019; @chalmers2021].

Building on the model of teacher stress [@kyriacou1978, see Fig. 2], we had hypothesized that more experienced teachers, with better classroom management skills at their disposal, experience less physiological stress when dealing with classroom disruptions. Contrary to our expectations, we found no buffering effect of teaching experience on teachers’ HR, i.e., more experienced teachers did not show lower mean standardized HR during the stressful teaching phase than less experienced teachers. Rather, at least descriptively, we observed the opposite trend. There are several possible explanations for this finding. First, teaching experience is inherently confounded with age (the two variables correlated at *r* = .94 in our sample), and age has been shown to affect indicators of cardiovascular reactivity in various ways [@uchino2010older]. However, to avoid this kind of confounding influence, we had used not raw BPM but rather standardized mean HR for all our analyses, thus controlling at least for inter-individual differences in mean HR. Second, as research on teacher professionalization has repeatedly shown, professional experience is not a guarantee for higher professional knowledge and skills [@kirschner2016professionswissen]. Rather, honing skills from professional experience necessitates a deliberate practice of choosing to improve, learning through experience, and integrating new knowledge into future performances [@dunn1999deliberate]. Thus, rather than professional experience alone, more direct assessments of classroom management skills, such as objective behavior-based tests, would be a better indicator of expertise that future studies could explore. Finally, and most importantly, the highly controlled teaching situation that we created in the lab might not have provided sufficient resemblance to the expert teachers’ working conditions to let them effectively use their coping resources. In other words, since the situation was unfamiliar to both experienced and unexperienced teachers, their stress levels might have been more similar than they would have been in a more authentic classroom setting.

With regards to the predictive power of teachers’ subjective appraisals of the classroom disruption during teaching, we, first of all, have to conclude that our hypotheses were not supported, as neither confidence appraisal nor disruptiveness appraisal showed any notable correlations with teachers’ means standardized HR or any explanatory power over and beyond teaching experience. Possibly, teachers’ self-reported appraisals, and their actual physiological stress responses, tap into quite different phenomena, or at least, quite different aspects of the multifaceted stress response [@kyriacou1978]. In addition, while HR was assessed online during teaching, self-reported appraisals were given in retrospect during the SRI, and may be subject to biased (e.g., self-serving) reporting or simply an inability to recall one´s immediate stress reactions.

On the other hand, when controlling for all other factors, teachers who reported to have perceived the events as more disruptive showed a higher HR (β = .25) in the phase immediately following the micro-teaching unit. This finding would be consistent with the idea that differences in mean HR, as an indicator of the physiological stress response, can be linked to the cognitive appraisal of stressors.

In the present study we examined patterns of responses for online and paper-based SET scores at a midsized, regional, comprehensive university in the United States. We posed two questions: First, does the response rate or the average SET score change when an institution administers SET forms online instead of on paper? Second, what is the minimal response rate required to produce stable average SET scores for an instructor? Whereas much earlier research relied on small samples often limited to a single academic department, we gathered SET data on a large sample of courses (*N* = 364) that included instructors from all colleges and all course levels over 3 years. We controlled for individual differences in instructors by limiting the sample to courses taught by the same instructor in all 3 years. The university offers nearly 30% of course sections online in any given term, and these courses have always administered online SETs. This allowed us to examine the combined effects of changing the method of delivery for SETs (paper-based to online) for traditional classes and changing from a mixed method of administering SETs (paper for traditional classes and online for online classes in the first 2 years of data gathered) to uniform use of online forms for all classes in the final year of data collection.

# Method

## Sample

Response rates and evaluation ratings were retrieved from archived course evaluation data. The archive of SET data did not include information about personal characteristics of the instructor (gender, age, or years of teaching experience), and students were not provided with any systematic incentive to complete the paper or online versions of the SET. We extracted data on response rates and evaluation ratings for 364 courses that had been taught by the same instructor during three consecutive fall terms (2012, 2013, and 2014).

The sample included faculty who taught in each of the five colleges at the university: 109 instructors (30%) taught in the College of Social Science and Humanities, 82 (23%) taught in the College of Science and Engineering, 75 (21%) taught in the College of Education and Professional Studies, 58 (16%) taught in the College of Health, and 40 (11%) taught in the College of Business. Each instructor provided data on one course. Approximately 259 instructors (71%) provided ratings for face-to-face courses, and 105 (29%) provided ratings for online courses, which accurately reflects the proportion of face-to-face and online courses offered at the university. The sample included 107 courses (29%) at the beginning undergraduate level (1st- and 2nd-year students), 205 courses (56%) at the advanced undergraduate level (3rd- and 4th-year students), and 52 courses (14%) at the graduate level.

## Instrument

The course evaluation instrument was a set of 18 items developed by the state university system. The first eight items were designed to measure the quality of the instructor, concluding with a global rating of instructor quality (Item 8: “Overall assessment of instructor”). The remaining items asked students to evaluate components of the course, concluding with a global rating of course organization (Item 18: “Overall, I would rate the course organization”). No formal data on the psychometric properties of the items are available, although all items have obvious face validity.

Students were asked to rate each instructor as *poor* (0), *fair* (1), *good* (2), *very good* (3), or *excellent* (4) in response to each item. Evaluation ratings were subsequently calculated for each course and instructor. A median rating was computed when an instructor taught more than one section of a course during a term.

The institution limited our access to SET data for the 3 years of data requested. We obtained scores for Item 8 (“Overall assessment of instructor”) for all 3 years but could obtain scores for Item 18 (“Overall, I would rate the course organization”) only for Year 3. We computed the correlation between scores on Item 8 and Item 18 (from course data recorded in the 3rd year only) to estimate the internal consistency of the evaluation instrument. These two items, which serve as composite summaries of preceding items (Item 8 for Items 1–7 and Item 18 for Items 9–17), were strongly related, *r*(362) = .92. Feistauer and Richter (2016) also reported strong correlations between global items in a large analysis of SET responses.

## Design

This study took advantage of a natural experiment created when the university decided to administer all course evaluations online. We requested SET data for the fall semesters for 2 years preceding the change, when students completed paper-based SET forms for face-to-face courses and online SET forms for online courses, and data for the fall semester of the implementation year, when students completed online SET forms for all courses. We used a 2 × 3 × 3 factorial design in which course delivery method (face to face and online) and course level (beginning undergraduate, advanced undergraduate, and graduate) were between-subjects factors and evaluation year (Year 1: 2012, Year 2: 2013, and Year 3: 2014) was a repeated-measures factor. The dependent measures were the response rate (measured as a percentage of class enrollment) and the rating for Item 8 (“Overall assessment of instructor”).

Data analysis was limited to scores on Item 8 because the institution agreed to release data on this one item only. Data for scores on Item 18 were made available for SET forms administered in Year 3 to address questions about variation in responses across items. The strong correlation between scores on Item 8 and scores on Item 18 suggested that Item 8 could be used as a surrogate for all the items. These two items were of particular interest because faculty, department chairs, and review committees frequently rely on these two items as stand-alone indicators of teaching quality for annual evaluations and tenure and promotion reviews.

# Results

## Response Rates

Response rates are presented in Table 1. The findings indicate that response rates for face-to-face courses were much higher than for online courses, but only when face-to-face course evaluations were administered in the classroom. In the Year 3 administration, when all course evaluations were administered online, response rates for face-to-face courses declined (*M* = 47.18%, *SD* = 20.11), but were still slightly higher than for online courses (*M* = 41.60%, *SD* = 18.23). These findings produced a statistically significant interaction between course delivery method and evaluation year, *F*(1.78, 716) = 101.34, *MSE* = 210.61, *p* < .001.[[5]](#footnote-6) The strength of the overall interaction effect was .22 (ηp2). Simple main-effects tests revealed statistically significant differences in the response rates for face-to-face courses and online courses for each of the 3 observation years.[[6]](#footnote-7) The greatest differences occurred during Year 1 (*p* < .001) and Year 2 (*p* < .001), when evaluations were administered on paper in the classroom for all face-to-face courses and online for all online courses. Although the difference in response rate between face-to-face and online courses during the Year 3 administration was statistically reliable (when both face-to-to-face and online courses were evaluated with online surveys), the effect was small (ηp2 = .02). Thus, there was minimal difference in response rate between face-to-face and online courses when evaluations were administered online for all courses. No other factors or interactions included in the analysis were statistically reliable.

## Evaluation Ratings

The same 2 × 3 × 3 analysis of variance model was used to evaluate mean SET ratings. This analysis produced two statistically significant main effects. The first main effect involved evaluation year, *F*(1.86, 716) = 3.44, *MSE* = 0.18, *p* = .03 (ηp2 = .01; see Footnote 1). Evaluation ratings associated with the Year 3 administration (*M* = 3.26, *SD* = 0.60) were significantly lower than the evaluation ratings associated with both the Year 1 (*M* = 3.35, *SD* = 0.53) and Year 2 (*M* = 3.38, *SD* = 0.54) administrations. Thus, all courses received lower SET scores in Year 3, regardless of course delivery method and course level. However, the size of this effect was small (the largest difference in mean rating was 0.11 on a five-item scale).

The second statistically significant main effect involved delivery mode, *F*(1, 358) = 23.51, *MSE* = 0.52, *p* = .01 (ηp2 = .06; see Footnote 2). Face-to-face courses (*M* = 3.41, *SD* = 0.50) received significantly higher mean ratings than did online courses (*M* = 3.13, *SD* = 0.63), regardless of evaluation year and course level. No other factors or interactions included in the analysis were statistically reliable.

## Stability of Ratings

The scatterplot presented in Figure 1 illustrates the relation between SET scores and response rate. Although the correlation between SET scores and response rate was small and not statistically significant, *r*(362) = .07, visual inspection of the plot of SET scores suggests that SET ratings became less variable as response rate increased. We conducted Levene’s test to evaluate the variability of SET scores above and below the 60% response rate, which several researchers have recommended as an acceptable threshold for response rates (Berk, 2012, 2013; Nulty, 2008). The variability of scores above and below the 60% threshold was not statistically reliable, *F*(1, 362) = 1.53, *p* = .22.

# Discussion

Online administration of SETs in this study was associated with lower response rates, yet it is curious that online courses experienced a 10% increase in response rate when all courses were evaluated with online forms in Year 3. Online courses had suffered from chronically low response rates in previous years, when face-to-face classes continued to use paper-based forms. The benefit to response rates observed for online courses when all SET forms were administered online might be attributed to increased communications that encouraged students to complete the online course evaluations. Despite this improvement, response rates for online courses continued to lag behind those for face-to-face courses. Differences in response rates for face-to-face and online courses might be attributed to characteristics of the students who enrolled or to differences in the quality of student engagement created in each learning modality. Avery et al. (2006) found that higher performing students (defined as students with higher GPAs) were more likely to complete online SETs.

Although the average SET rating was significantly lower in Year 3 than in the previous 2 years, the magnitude of the numeric difference was small (differences ranged from 0.08 to 0.11, based on a 0–4 Likert-like scale). This difference is similar to the differences Risquez et al. (2015) reported for SET scores after statistically adjusting for the influence of several potential confounding variables. A substantial literature has discussed the appropriate and inappropriate interpretation of SET ratings (Berk, 2013; Boysen, 2015a, 2015b; Boysen et al., 2014; Dewar, 2011; Stark & Freishtat, 2014).

Faculty have often raised concerns about the potential variability of SET scores due to low response rates and thus small sample sizes. However, our analysis indicated that classes with high response rates produced equally variable SET scores as did classes with low response rates. Reviewers should take extra care when they interpret SET scores. Decision makers often ignore questions about whether means derived from small samples accurately represent the population mean (Tversky & Kahneman, 1971). Reviewers frequently treat all numeric differences as if they were equally meaningful as measures of true differences and give them credibility even after receiving explicit warnings that these differences are not meaningful (Boysen, 2015a, 2015b).

Because low response rates produce small sample sizes, we expected that the SET scores based on smaller class samples (i.e., courses with low response rates) would be more variable than those based on larger class samples (i.e., courses with high response rates). Although researchers have recommended that response rates reach the criterion of 60%–80% when SET data will be used for high-stakes decisions (Berk, 2012, 2013; Nulty, 2008), our findings did not indicate a significant reduction in SET score variability with higher response rates.

## Implications for Practice

### Improving SET Response Rates

When decision makers use SET data to make high-stakes decisions (faculty hires, annual evaluations, tenure, promotions, teaching awards), institutions would be wise to take steps to ensure that SETs have acceptable response rates. Researchers have discussed effective strategies to improve response rates for SETs (Nulty, 2008; see also Berk, 2013; Dommeyer et al., 2004; Jaquett et al., 2016). These strategies include offering empirically validated incentives, creating high-quality technical systems with good human factors characteristics, and promoting an institutional culture that clearly supports the use of SET data and other information to improve the quality of teaching and learning. Programs and instructors must discuss why information from SETs is important for decision-making and provide students with tangible evidence of how SET information guides decisions about curriculum improvement. The institution should provide students with compelling evidence that the administration system protects the confidentiality of their responses.

### Evaluating SET Scores

In addition to ensuring adequate response rates on SETs, decision makers should demand multiple sources of evidence about teaching quality (Buller, 2012). High-stakes decisions should never rely exclusively on numeric data from SETs. Reviewers often treat SET ratings as a surrogate for a measure of the impact an instructor has on student learning. However, a recent meta-analysis (Uttl et al., 2017) questioned whether SET scores have any relation to student learning. Reviewers need evidence in addition to SET ratings to evaluate teaching, such as evidence of the instructor’s disciplinary content expertise, skill with classroom management, ability to engage learners with lectures or other activities, impact on student learning, or success with efforts to modify and improve courses and teaching strategies (Berk, 2013; Stark & Freishtat, 2014). As with other forms of assessment, any one measure may be limited in terms of the quality of information it provides. Therefore, multiple measures are more informative than any single measure.

A portfolio of evidence can better inform high-stakes decisions (Berk, 2013). Portfolios might include summaries of class observations by senior faculty, the chair, and/or peers. Examples of assignments and exams can document the rigor of learning, especially if accompanied by redacted samples of student work. Course syllabi can identify intended learning outcomes; describe instructional strategies that reflect the rigor of the course (required assignments and grading practices); and provide other information about course content, design, instructional strategies, and instructor interactions with students (Palmer et al., 2014; Stanny et al., 2015).

## Conclusion

Psychology has a long history of devising creative strategies to measure the “unmeasurable,” whether the targeted variable is a mental process, an attitude, or the quality of teaching (e.g., Webb et al., 1966). In addition, psychologists have documented various heuristics and biases that contribute to the misinterpretation of quantitative data (Gilovich et al., 2002), including SET scores (Boysen, 2015a, 2015b; Boysen et al., 2014). These skills enable psychologists to offer multiple solutions to the challenge posed by the need to objectively evaluate the quality of teaching and the impact of teaching on student learning.

Online administration of SET forms presents multiple desirable features, including rapid feedback to instructors, economy, and support for environmental sustainability. However, institutions should adopt implementation procedures that do not undermine the usefulness of the data gathered. Moreover, institutions should be wary of emphasizing procedures that produce high response rates only to lull faculty into believing that SET data can be the primary (or only) metric used for high-stakes decisions about the quality of faculty teaching. Instead, decision makers should expect to use multiple measures to evaluate the quality of faculty teaching.

# References

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *The Journal of Economic Education*, *37*(1), 21–37. <https://doi.org/10.3200/JECE.37.1.21-37>

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching and Learning*, *8*(2), 98–107.

Berk, R. A. (2013). *Top 10 flashpoints in student ratings and the evaluation of teaching*. Stylus.

Boysen, G. A. (2015a). Preventing the overinterpretation of small mean differences in student evaluations of teaching: An evaluation of warning effectiveness. *Scholarship of Teaching and Learning in Psychology*, *1*(4), 269–282. <https://doi.org/10.1037/stl0000042>

Boysen, G. A. (2015b). Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation. *Scholarship of Teaching and Learning in Psychology*, *1*(2), 150–162. <https://doi.org/10.1037/stl0000017>

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education*, *39*(6), 641–656. <https://doi.org/10.1080/02602938.2013.860950>

Buller, J. L. (2012). *Best practices in faculty evaluation: A practical guide for academic leaders*. Jossey-Bass.

Dewar, J. M. (2011). Helping stakeholders understand the limitations of SRT data: Are we doing enough? *Journal of Faculty Development*, *25*(3), 40–44.

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College students’ attitudes toward methods of collecting teaching evaluations: In-class versus on-line. *Journal of Education for Business*, *78*(1), 11–15. <https://doi.org/10.1080/08832320209599691>

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611–623. <https://doi.org/10.1080/02602930410001689171>

Feistauer, D., & Richter, T. (2016). How reliable are students’ evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279. <https://doi.org/10.1080/02602938.2016.1261083>

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511808098>

Griffin, T. J., Hilton, J., III, Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: Taking a closer look. *Assessment & Evaluation in Higher Education*, *39*(3), 339–348. <https://doi.org/10.1080/02602938.2013.831809>

Jaquett, C. M., VanMaaren, V. G., & Williams, R. L. (2016). The effect of extra-credit incentives on student submission of end-of-course evaluations. *Scholarship of Teaching and Learning in Psychology*, *2*(1), 49–61. <https://doi.org/10.1037/stl0000052>

Jaquett, C. M., VanMaaren, V. G., & Williams, R. L. (2017). Course factors that motivate students to submit end-of-course evaluations. *Innovative Higher Education*, *42*(1), 19–31. <https://doi.org/10.1007/s10755-016-9368-5>

Morrison, R. (2011). A comparison of online versus traditional student end-of-course critiques in resident courses. *Assessment & Evaluation in Higher Education*, *36*(6), 627–641. <https://doi.org/10.1080/02602931003632399>

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education*, *35*(4), 463–475. <https://doi.org/10.1080/02602930902862875>

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314. <https://doi.org/10.1080/02602930701293231>

Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014). Measuring the promise: A learning-focused syllabus rubric. *To Improve the Academy: A Journal of Educational Development*, *33*(1), 14–36. <https://doi.org/10.1002/tia2.20004>

Reiner, C. M., & Arnold, K. E. (2010). Online course evaluation: Student and instructor perspectives and assessment potential. *Assessment Update*, *22*(2), 8–10. <https://doi.org/10.1002/au.222>

Risquez, A., Vaughan, E., & Murphy, M. (2015). Online student evaluations of teaching: What are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education*, *40*(1), 210–234. <https://doi.org/10.1080/02602938.2014.890695>

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598–642. <https://doi.org/10.3102/0034654313496870>

Stanny, C. J., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and learning through a syllabus review. *Assessment & Evaluation in Higher Education*, *40*(7), 898–913. <https://doi.org/10.1080/02602938.2014.956684>

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. <https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1>

Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education*, *37*(4), 465–473. <https://doi.org/10.1080/02602938.2010.545869>

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. <https://doi.org/10.1037/h0031322>

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>

Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, *35*(1), 101–115. <https://doi.org/10.1080/02602930802618336>

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally.

**Table 1**

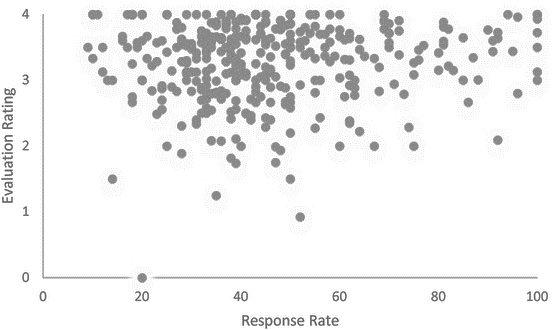
*Means and Standard Deviations for Response Rates (Course Delivery Method by Evaluation Year)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Administration year | Face-to-face course | | Online course | |
| *M* | *SD* | *M* | *SD* |
| Year 1: 2012 | 71.72 | 16.42 | 32.93 | 15.73 |
| Year 2: 2013 | 72.31 | 14.93 | 32.55 | 15.96 |
| Year 3: 2014 | 47.18 | 20.11 | 41.60 | 18.23 |

*Note.* Student evaluations of teaching (SETs) were administered in two modalities in Years 1 and 2: paper based for face-to-face courses and online for online courses. SETs were administered online for all courses in Year 3.

**Figure 1**

*Scatterplot Depicting the Correlation Between Response Rates and Evaluation Ratings*



*Note.* Evaluation ratings were made during the 2014 fall academic term.

1. The curve was smoothed using the geom\_smooth() function from the ggplot2 package in R (v3.3.3; Wickham, 2016) based on the smoothing method LOESS (Locally Estimated Scatterplot Smoothing). This method fits a polynomial surface determined by one or more numerical predictors, using local fitting. [↑](#footnote-ref-2)
2. Note that the study exceeded the planned duration of two hours for a few participants. To avoid distortions when mapping the HR over the course of the study (see Fig. 3), the endpoint was set at two hours for all participants, even though data from later time points was used in the end interval for a few participants. [↑](#footnote-ref-3)
3. We used the mean standardized HR instead of the mean intercept as we wanted to explain the mean HR of the entire intervals and not the HR at the very beginning of the interval (x = 0). [↑](#footnote-ref-4)
4. Although this procedure does not account for nonmonotonic progressions in individual HR, a graphical evaluation revealed that the linear estimates corresponded well to the majority of the cases (see XX in the supplementary material). [↑](#footnote-ref-5)
5. A Greenhouse–Geisser adjustment of the degrees of freedom was performed in anticipation of a sphericity assumption violation. [↑](#footnote-ref-6)
6. A test of the homogeneity of variance assumption revealed no statistically significant difference in response rate variance between the two delivery modes for the 1st, 2nd, and 3rd years. [↑](#footnote-ref-7)