From heartbeat to data – Using wearable fitness trackers as an affordable approach to assess teacher stress

Mandy Klatt[a,1,*], Christin Lotz[a,2], Gregor Kachel[b,3], Peer Keßler[c,4], Anne Deiglmayr[a,5]

[a] *Division of Empirical School and Classroom Research, Institute of Educational Sciences, Marschnerstr. 29, Leipzig, 04109, Germany*
[b] *Department, Deutscher Pl. 6, Leipzig, 04103, Germany*
[c] *Division of Health and Prevention, Institute of Psychology, Robert-Blum-Str. 13, Greifswald, 17489, Germany*

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

1. Introduction

The teaching profession is one of the most stressful professions, with teachers facing a host of stressors during their everyday work [**?** **?** **?**]. 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 [**?** **?**]. For example, it has been shown that teacher-centered activities and typical classroom-related stressors increase teacher HR during teaching activities [**?** **?** **?** **?** **?**]. However, previous studies have often relied on expensive and obtrusive electrocardiographs (ECG). Modern fitness trackers represent a non-invasive and convenient alternative [**?**].

Classroom disruptions are a major stressor in teachers' daily work [**?** **?**], and learning how to deal with them is an important aspect of professional expertise [**?**]. According to **?**] transactional model of stress and coping, the experience of stress in response to stressors such as classroom disruptions depends on the

*Corresponding author
Email addresses: mandy.klatt@uni-leipzig.de (Mandy Klatt), christin.lotz@uni-leipzig.de (Christin Lotz), gregor@example.com (Gregor Kachel), peer.kessler@uni-greifswald.de (Peer Keßler), anne.deiglmayr@uni-leipzig.de (Anne Deiglmayr)

[1] Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft
[2] Conceptualization, Methodology, Writing - Review & Editing, Supervision
[3] Formal analysis, Data Curation, Writing - Review & Editing
[4] Conceptualization, Methodology, Writing - Review & Editing
[5] Conzeptualization, Methodology, Writing - Review & Editing, Supervision

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 [**?**]. Therefore, in order to better understand how classroom stressors affect teachers' stress response, subjective self-reports should be accompanied by objective, physiological measures [**?**]. 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 appears particularly relevant given the profession's generally high stress levels and associated negative health effects [**?** **?**]. 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.

1.1. 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 [**?**] and stress management [**?**]. They can be used as ubiquitous, low-cost, and unintrusive data collection instruments [**?**], and their wide-spread use and everyday availability align with the increasing popularity and acceptance of wearables among the general population [**?**]. In contrast to self-reported questionnaires on stress [**?** **?**] that are prone to biases like social desirability [**?**] or recall errors [**?**], fitness trackers, as ambulatory assessment methods [**?** **?**], offer more objective insights into teachers' stress levels by monitoring teachers' physiological stress responses without disrupting teaching [**?** **?**].

One important health parameter assessed by nearly all wrist-worn fitness trackers is heart rate [**?**]. HR indicates the number of heartbeats within one minute and is typically expressed as beats per minute (BPM) [**?** **?**]. At rest, the average HR of adults typically ranges between 60 and 80 BPM [**?**]. HR can be detected and measured in different ways using sensors, such as electrocardiography (ECG) or photoplethysmography (PPG) [**?**]. 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 [**?**], 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 [**?**]. This optical method assesses HR by flashing green or red lights to measure changes in blood volume in the capillaries of the skin [**?**].

Physiologically, HR is regulated by the sympathetic and parasympathetic nervous systems [**?**]. An increase in sympathetic activity results in HR being sped up ("fight or flight" response) [**?**]; whereas an increase in parasympathetic activity results in HR being slowed down ("rest and digest" response, **?**]). Stress or mental and physical strain directly increases HR [**?** **?**]. Therefore, an increase in HR can be regarded as an indicator of increasing stress, and a decrease as an indicator of relaxation and ease [**?**]. Thus, fitness trackers offer low-cost and unobtrusive access to psychological stress data.

1.2. 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 [**?** **?** **?** **?**]. **?**] 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 [**?**], or low student engagement and motivation [**?**]. **?**] 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 [**?** **?**]. **?**] 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. **?**] used HR data from a fitness tracker to identify physiological changes during stress-inducing tasks (i.e., the Trier Social Stress Test; **?**]). Average HR increased significantly from the resting to the stress inducing phases. Even though the participants of these previous studies [**?** **?**] 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: **?**] 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 [**?** **?**]. For teachers, teacher-centered phases led to an increase in HR [**?** **?** **?** **?**]. 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.

1.3. A model of teacher stress

According to **?**], 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 **?**] and, for a more recent adaptation, **?**]) is based on the transactional stress model by Lazarus and colleagues [**?** **?**], 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.

Fig.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 **?**]: 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 Fig.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 [**?**], 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 [**?** **?**]. Novices in the teaching profession, on the other hand, face considerable stress and often feel overwhelmed by the demands of teaching [**?** **?** **?**] with many leaving the profession within the first five years [**?**]. 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 [**?** **?** **?**], highlighting the importance of professional expertise for managing teacher stress [**?**].

1.4. 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

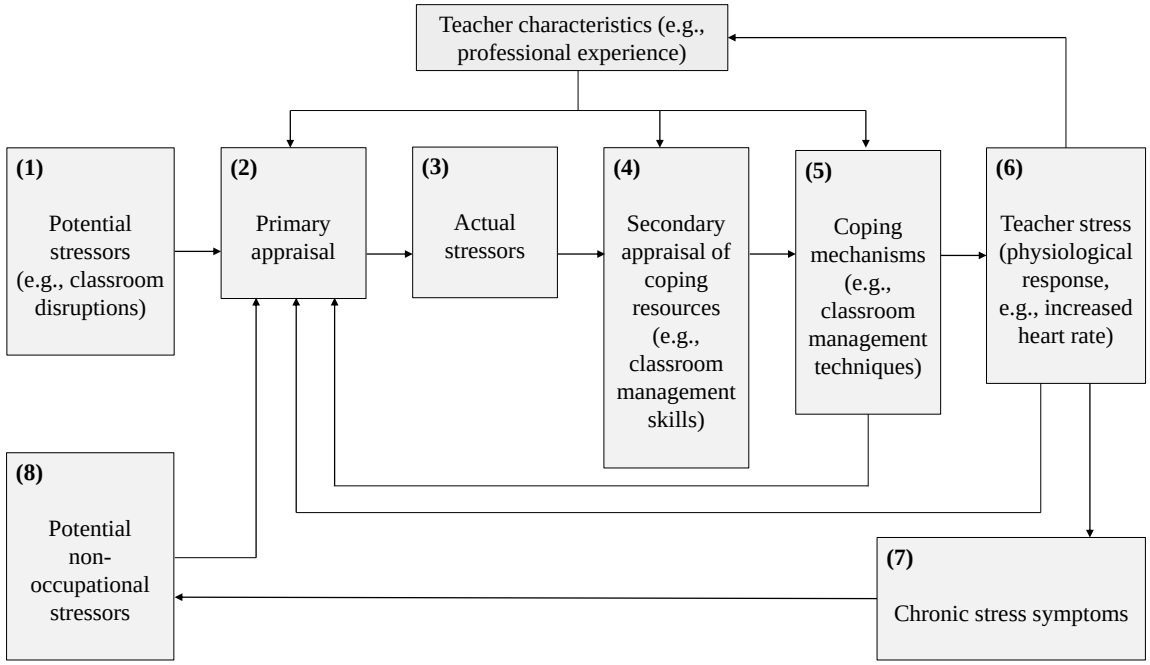


Figure 1: A model of teacher stress (adapted from van Dick 2006, p.37, modified by the authors).

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 Fig.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).

(2) 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 (Hypothesis 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).

2. Method

2.1. 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 $n_{total} = 81$ ($n_{total} = 52$ women, $n_{total} = 29$ 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).

2.2. 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.

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 Tab.A1 in the supplementary material for an overview and categorization of all events; and Fig.A2 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.

2.3. Measures

2.3.1. 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 [**?**], 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[6]. 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 [**?**]. The intervals were selected based on the following rules: The *pre-teaching interval* (I_1) comprised the first 10 minutes after the fitness tracker had been put on. The *teaching interval* (I_2) 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 [**? ?**]. The *post-teaching interval* (I_3) started immediately after the end of the teaching unit. The *interview interval* (I_4) 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 (I_5) comprised the last 10 minutes before the fitness tracker was taken off.

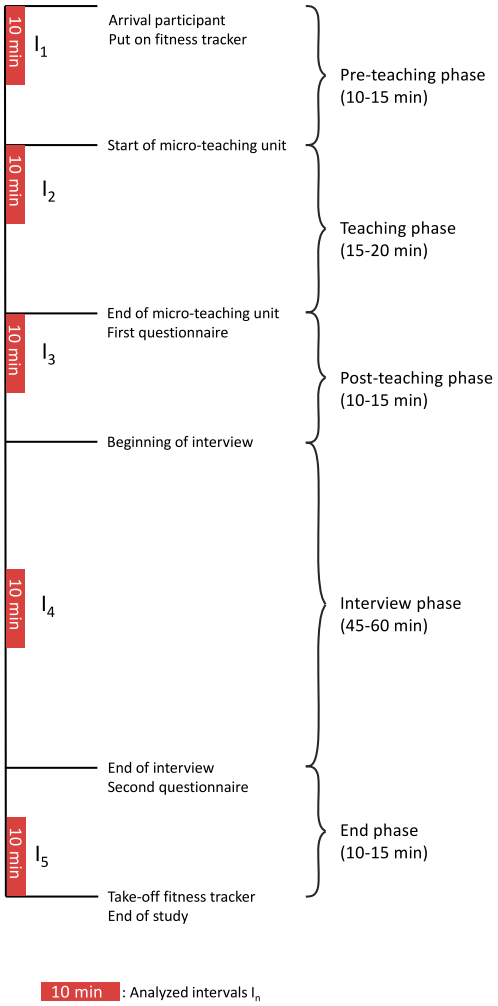


Figure 2: Procedure of the two-hour-long study consisting of five phases with five representative 10-minute intervals.

2.3.2. Teaching Experience

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

[6]The fluctuations in the number of seconds in which the HR was measured are due to the participants' movements, meaning that the device could not measure the HR every second.

2.3.3. 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.

2.4. Data analysis

We conducted all analyses with R [**?**]. Graphics were created using ggplot2 [**?**].

To enable visual inspection of HR trends, we displayed smoothed teacher HR over the course of the recording.[7] We visually compared unstandardized and standardized HR trends over the two-hour recording period.[8] 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[9], 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* (I_2) and the other four intervals by planned contrasts and inspection of effect size d [**?**].

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 [**?**] for each interval to estimate intercepts and linear slopes for each individual, which were then averaged across individuals.[10] 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).[11]

3. 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 Tab.1. Fig.3 displays 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.

Fig.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, $F(4, 400) = 260.62$, $p < .05$, $d = 1.60$ (large effect). Planned contrasts indicated that, as hypothesized (Hypothesis 1a), mean

[7] The curve was smoothed using the geom_smooth() function from the ggplot2 package in R [**?**] 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.

[8] 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.

[9] 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 ($x = 0$).

[10] 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 Fig.A3 to A7 in the supplementary material).

[11] Please note: HR levels and changes were not regressed on the disruption and confidence appraisals in the *pre-teaching interval* (I_1), because the appraised classroom events had not yet taken place in that phase.

standardized HR was significantly higher in the *teaching interval* (I_2) than in all other intervals, specifically, the *pre-teaching interval* (I_1; $t(400) = -10.08$, $p < .05$, $d = 1.034$; large effect), the post-teaching interval (I_3; $t(400) = -6.94$, $p < .05$, $d = 1.37$; large effect), the interview interval (I_4; $t(400) = 15.00$, $p < .05$, $d = 3.29$; large effect), and the end interval (I_5); $t(400) = 22.54$, $p < .05$, $d = 4.64$; large effect).

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 Tab.2. The mean slope of the *pre-teaching interval* (I_1) was significantly positive, indicating an increase in HR, as hypothesized. Further, the mean slopes of the *teaching interval* (I_2), *post-teaching interval* (I_3), and *interview interval* (I_4) were significantly negative, indicating a decrease in HR. For the last interval, the *end interval* (I_5), the mean slope was negative, but did not differ significantly from zero.

Interval	*M* HR	*SD* HR	Min	Max
Overall Course of 2h	90.09/0.04[12]	15.76/0.991	51/-4.03	164/4.56
Pre-teaching interval (I_1)	96.28/0.48	14.11/0.88	56/-3.56	139/3.24
Teaching interval (I_2)	100.80/0.85	16.23/0.77	63/-2.18	164/4.37
Post-teaching interval (I_3)	93.61/0.27	14.01/0.76	60/-2.17	150/3.06
Interview interval (I_4)	82.32/-0.72	11.85/0.74	51/-2.51	132/4.39
End interval (I_5)	77.95/-1.07	11.14/0.57	50[13]/-2.68	120/2.96

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* (*SD*)		*p*	
	Intercept	Slope	Intercept	Slope
Pre-teaching interval	0.052 (0.820)	0.085 (0.133)	.57	$< .05$ (I_1)
Teaching interval	1.025 (0.690)	-0.039 (0.108)	$< .05$	$< .05$ (I_2)
Post-teaching interval	0.549 (0.547)	-0.060 (0.101)	$< .05$	$< .05$ (I_3)
Interview interval	-0.617 (0.614)	-0.022 (0.070)	$< .05$	$< .05$ (I_4)
End interval	-1.004 (0.500)	-0.012 (0.074)	$< .05$	.14 (I_5)

Table 2: Analysis (*M, SD, p*-values) for the mean intercepts and the mean slopes for the five intervals.

[12]Please note that standardized *M* and *SD* of the overall course were not exactly 0 and 1 due to rounding differences.

[13]Deviations of the minimum values in the overall course vs. the *end interval* (I_5) are due to data of a few participants who needed more than two hours to finish the study.

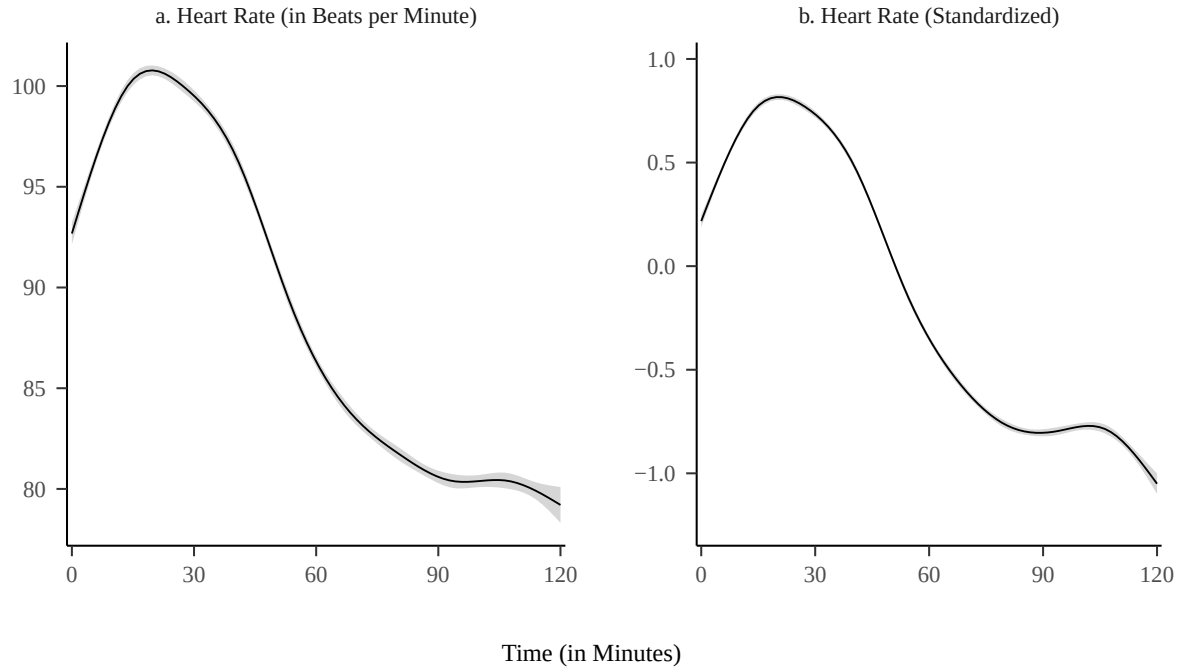
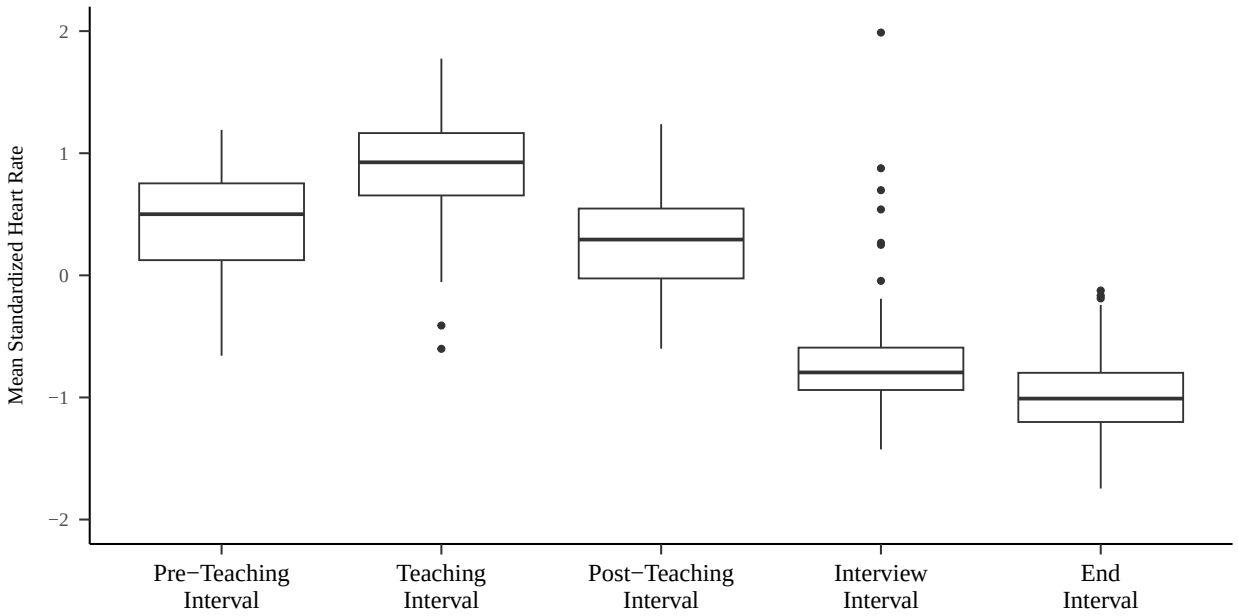


Figure 3: Overall course of the HR with the unstandardized HR in BPM shown in Fig.3a. and the z-standardized HR shown in Fig.3b. for the planned 2-hour study.



Note: N = 81 participants per interval. Fig. shows median (bold line), interquartile range (box) and outliers (dots).

Figure 4: Distribution of the standardized heart rate means in the five intervals.

3.2. Predicting mean standardized HR and mean slopes

Tab.3 shows the raw correlations among mean standardized HR/mean slopes (see Tab.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 (Tab.3). Correlations between teaching experience and appraisals (not shown in Tab.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 ($r = -.37$).

Tab.4 shows the results of the regression analyses. Teaching experience significantly predicted mean standardized HR only in the *interview interval* (Tab.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* (Tab.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* (Tab.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.

Variable	Pre-teaching interval	Teaching interval	Post-teaching interval	Interview interval	End interval
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 3: Correlations between mean standardized HR/mean slopes and the predictor variables of teaching experience, disruption appraisal, and confidence appraisal for the five intervals.

Dependent variable:	Model 1				Model 2				Model 3				Model 4			
	Mean standardized HR and mean slopes															
	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 (I_1)																
Teaching Experience	−.17 (.005)	.12	−.27* (.002)	< .05												
R^2	.030		.071													
Teaching interval (I_2)																
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								
R^2	.012		.000		.040		.015		.012		.010		.042		.031	
Δ R^2			.028		.015		.000		.010		.030		.031			

Continued on next page

Dependent variable:	Model 1		Model 2		Model 3		Model 4	
	Mean standardized HR and mean slopes							
	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p
Post-teaching interval (I_3)								
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								
R^2	.002		.001		.043		.020		.006		.002		.058		.023	
Δ R^2			.041		.019		.004		.001		.056		.022			
Interview interval (I_4)								
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								

Continued on next page

Dependent variable:	Model 1		Model 2		Model 3		Model 4	
	Mean standardized HR and mean slopes							
	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p
Confidence Appraisal	−.02 (.050)	.85	−.13 (.007)	.29	−.04 (.052)	.76	−.16 (.007)	.20
R^2	.058		.040		.060		.050	
Δ R^2			.002		.010		.000	
End interval (I_5)								
Teaching Experience	.04 (.004)	.70	.11 (.001)	.32	.07 (.005)	.58	.18 (.001)	.13
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
R^2	.002		.013		.005		.053	
Δ R^2			.003		.040		.010	

Dependent variable:	Model 1		Model 2		Model 3		Model 4	
	Mean standardized HR and mean slopes							
	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p
Confidence Appraisal								
R^2	.058		.054		.061		.069	
Δ R^2			.014		.003		.029	
End interval (I_5)								
Teaching Experience	.09 (.005)	.46	.07 (.001)	.58	.10 (.005)	.43	.12 (.001)	.33
Disruption Appraisal								
Confidence Appraisal								
R^2	.012		.025		.013		.078	
Δ R^2	.012		.011		.065			

Table 5: Standardized regression coefficients of mean standardized heart rate and mean slopes predicted by teaching experience, disruption appraisal, and confidence appraisal for the five intervals.

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. * $p < .05$.

4. Discussion

4.1. 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 [**?** **?** **?** **?**], as well as with findings showing how HR changes align with activating events and stress-inducing tasks [**?** **?**].

Building on the model of teacher stress (**?**], 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 [**?**]. 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 [**?**]. Rather, honing skills from professional experience necessitates a deliberate practice of choosing to improve, learning through experience, and integrating new knowledge into future performances [**?**]. 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.

While we did not find a buffering effect of teaching experience on mean HR during teaching, we did, however, find a less steep HR increase in more experienced, compared to less experienced teachers during the *pre-teaching phase* ($\beta = -.27$), i.e., in preparation for the micro-teaching unit. This finding supports the idea that, even though teaching experience guarantees neither superior expertise nor stress resistance, the habits and routines formed by experienced teachers may at least lead to lower arousal levels (e.g., experienced as feeling less nervous and tense) when they anticipate potentially stressful teaching situations.

An interesting observation beyond our hypotheses was that teaching experience was predictive of HR differences, not during teaching, but in the *interview phase*: compared to less experienced teachers, more experienced teachers showed a higher mean standardized HR ($\beta = .24$) and, thus, probably experienced higher levels of physiological stress during the SRI. One possible explanation for this finding could again be the higher age of the more experienced teachers, along with slower recovery from stress in older teachers. For instance, **?**] observed that, compared to their younger colleagues, older teachers did not experience a decrease in their HR during periods of low stress levels, from which they concluded that recovery from stress was insufficient in the older teachers [**?**]. Alternatively, the finding could also be attributed to the fact that less experienced teachers, due to their ongoing or only recently concluded training, may have been more accustomed to reflecting on their work and receiving feedback as was the case during the SRI, whereas, these activities were less routine and possibly more face-threatening for experienced teachers. Therefore, it is possible that more experienced teachers found the interview itself to be more stressful and therefore showed a higher mean standardized HR during this phase.

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' mean 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 [**?**]. 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 ones 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 ($\beta = .25$) 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.

4.2. Limitations and future directions

While the laboratory setting of the study allowed for a controlled implementation of stressors and high internal validity, it was not an authentic classroom environment, raising questions about its external validity. Most importantly, the teacher and their students did not have a shared history, and only a very thin basis for establishing a positive teacher-student relationship, which is a core characteristic of effective classroom management [**?** **?**]. In addition, the micro-teaching unit was only about 15 minutes long, and thus much shorter than a regular school lesson, providing less opportunities for experienced teachers to build up an engaging lesson. Finally, student behavior was scripted, with classroom disruptions following the experimental schedule, irrelevant of the behavior of the teacher. Thus, the setting may have masked effects of teaching experience by providing too little opportunities of experienced teachers to demonstrate their true classroom management skills, in particular regarding the prevention of disruptions. In subsequent studies, it would therefore be insightful to assess teachers' HR in more authentic classroom settings over a longer period of time (e.g., days, weeks, or even months). Extended observation of teachers' HR in authentic classroom settings could reveal how factors such as student behavior, teaching methods, or organizational and administrative demands contribute to fluctuations in physiological arousal, uncovering insights into the sustained physiological demands of teaching that short-term studies may overlook. Finally, linking actual teacher behavior to potential stressors (e.g., classroom disruptions, noise level, etc.) would offer insights into teacher coping strategies and their links with physiological indicators of stress.

Another limitation concerns the assessment of teachers' HR. While our results demonstrate the usefulness of drawing upon easily available HR data from ubiquitous, low-cost, un-intrusive fitness trackers to estimate teacher stress, there also are shortcomings of this type of assessment. First, while fitness trackers typically yield HR data, heart rate variability (HRV) has been demonstrated to be an even more accurate indicator of stress [**?**]. While standard fitness trackers did not provide this measure at the time of our data collection, more recent products do offer this function. Thus, future studies might consider assessing HRV instead of HR. Second, we did not record participants' resting HR, which is generally considered an important baseline for determining inter- and intrapersonal differences in cardiovascular health and reactivity [**?** **?**]. A clean baseline HR requires a resting phase without physical movement or emotional stress, ideally fifteen minutes before the beginning of the activity, which is very difficult to achieve in practice [**?**], e.g., when assessing teacher HR before and during teaching. Thus, our study explored the possibility of substituting baseline HR measurement via z-standardization within participants. As a result, the absolute standardized values of each participant must always be interpreted in the context of the standardization sample, and thus are less interpretable than individual BPM values together with a baseline HR. However, for statistical analyses based on the whole sample, the standardization fulfills the aim of controlling for differences in individual HR due to, for example, age-related differences. Finally, depending on the brand and model of fitness trackers used, the precision of the HR measurement varies. Research on the reliability of our deployed Fitbit® device has proven that this brand is generally accurate in controlled settings and for moderate activity levels [**?** **?** **?** **?**], as in our study. For example, the Fitbit® fitness tracker has previously shown good HR measurement accuracy during resting phases [**?** **?**] and for activities such as walking, jogging, and running [**?**]. At higher exercise intensities such as cycling, the Fitbit® tracker may underestimate HR [**?** **?** **?** **?**] but is

still within an acceptable range according to systematic reviews [**?**]. Nevertheless, **?**] stressed that Fitbit® trackers cannot replace ECG when precision is paramount. Despite these considerations, the Fitbit® model appears suitable for our study purposes, as physical strain was moderate.

Furthermore, while we assessed teachers' appraisals of the stressful classroom disruptions using a SRI in which they could review the exact situation, these appraisal ratings were still post-hoc self-reports, which limits the interpretation of our results. One of the main issues with post-hoc self-reports is that they rely on the teachers' memories and subjective interpretations of past events, which may be prone to various biases such as social desirability [**?**] or recall errors [**?**]. Moreover, stress is not a fixed or stable construct; it is a dynamic, constantly evolving affective response that can vary depending on context, individual disposition, and prior experiences, making it particularly challenging to pinpoint valid and reliable process markers for how individuals appraise stress in real-time [**?**]. While SRIs provide a more detailed and reflective understanding of the stressor in question, the delayed nature of the response makes it difficult to capture the immediate, in-the-moment appraisal that occurs when the stressful event actually takes place.

4.3. Hands-on advice for using wrist-worn fitness trackers for research

For researchers aiming to use fitness trackers to collect data, there are practical aspects to consider concerning the design, data collection, and data analysis phases of research projects [for an additional overview, see **?**]:

1) Choosing a suitable model:

Before data collection, researchers need to decide which model of fitness tracker best suits their research question. One important point to consider is whether the study will be conducted in the laboratory, in a clinical environment, or under real-world conditions. Conventional fitness trackers should not be used if the focus is on measurement accuracy, such as in medical contexts, as they cannot replace the accuracy of ECG measurements [**?**]. Moreover, researchers should consider that measurement accuracy also depends on the intensity of the movements performed by the participants during data collection. Fitbit® fitness trackers, for example, underestimate HR at higher exercise intensities such as cycling [**?** **?** **?** **?**]. For reference, the systematic review by **?**] provides a detailed overview of studies that used wrist-worn fitness trackers between 2000 and 2019 and discusses their validity and reliability. Another point to consider is the price, which at the time of writing ranged between €30 for the cheapest models and up to €1.700 for medical wristbands. Currently, models assessing HRV in addition to HR are becoming more and more affordable and widespread. Still, Fitbit® fitness trackers might be a good choice for teams operating with moderate budgets or if larger groups of participants need to be tracked at the same time. Further, before conducting any study, it should be considered that the data collected with fitness trackers is health data, and therefore very sensitive. Researchers have to ensure that their chosen model of fitness tracker allows them to collect and store data in line with relevant ethical and legal requirements, for example, guaranteeing participants' anonymity and secure data storage.

2) Operating the fitness tracker:

In planning the operation of their chosen model of fitness tracker, researchers need to specify the circumference and attachment of the wrist band and the placement of the fitness tracker on participants. In particular, researchers conducting studies with children should take into account their smaller wrist size. When putting on a fitness tracker, attention must also be paid to whether it is attached to the dominant or non-dominant wrist, as this can influence HR measurements. Different models of fitness trackers need to be placed differently and in line with the manufacturer's instructions. It is also important to check that the battery is fully charged each time, that the latest software version is loaded, and that the fitness tracker has been synchronized before recording data to avoid unnecessary loss of data. Finally, if researchers want to accurately investigate parameters during specific time intervals, such as HR during lessons versus breaks, it is crucial to synchronize the fitness tracker with other time-keeping devices, such as watches. This synchronization allows researchers to precisely determine the onset and offset of particular activities or intervals of

interest. By aligning the recorded data with specific time frames, researchers can ensure that the physiological measurements, such as HR, are accurately associated with the corresponding periods of interest. This process enhances the validity and reliability of the data analysis, enabling a more precise examination of variations in physiological responses across different time intervals.

3) Extracting and analyzing fitness tracker data:

As far as the procedure for processing the data is concerned, researchers should ensure that the raw data of the physiological measurements are available for further analysis. For the Fitbit® HR measurements, for example, the raw data can be downloaded from a website in the form of .csv files. However, these must be downloaded as soon as possible after data collection, as some platforms automatically delete or archive older data files after a certain period due to policies regarding data storage and retention. This can result in loss of access to critical data. Additionally, ensuring that data is collected at the intended sampling rate is crucial for accurate analysis. For instance, while our fitness tracker was designed to record HR every 1-5 seconds, we occasionally observed recordings only every 15 seconds, possibly due to participant movement and tracker placement.

4.4. Conclusion

This study investigated whether HR data collected from teacher-worn fitness trackers are suitable for exploring links between HR, subjective stressor appraisal, and individual teaching experience, to achieve a more profound comprehension of teacher stress. Results suggest that the widespread availability of HR data from wearable fitness trackers, moving "from heartbeat to data", presents opportunities both to teachers for self-monitoring stress levels, and to researchers for assessing physiological indicators of stress. For example, using fitness trackers could enable teachers to strengthen their self-awareness in stressful situations and allow for early self-intervention such as mindfulness techniques (e.g., deep breathing or body scans; **?**]). Integrating fitness trackers into teacher training and everyday practice could offer an affordable and practical method for assessing and managing teacher stress. In teacher training as well as in research, triangulating data from fitness trackers, lesson videos, and interviews could provide teachers with insights into their own stress management, and foster the implementation of effective stress and classroom management strategies. Taken together, our findings cater to **?**] call for the use of ambulatory assessment methods, particularly in the context of classroom disruptions, for gaining a deeper understanding of teacher stress and its impact on both psychological and physiological variables.

In summary, our study contributes to the understanding of stress in educational settings and underscores the potential of wearable fitness trackers in advancing research on teacher well-being. By harnessing the power of wearable technology, we can provide teachers with the tools needed to better understand and manage their stress, ultimately enhancing their overall well-being.

Appendix

Verbal disruptions	Physical disruptions	Lack of eagerness to learn
Heckling	Clicking pen	Looking at phone
Chatting	Snipping hands	Drawing
Whispering	Drumming hands	Head on table

Table A1: Classification of nine, typical classroom disruptions according to [**?**] performed in the micro-teaching unit by actors.

Note. The setting included three actors as the class (left) and a teacher (participant, right).

Figure A1: Laboratory setting of the micro-teaching unit.

Note. The experimenter and participant watched the previously taught micro-teaching unit on video.

Figure A2: Laboratory setting of the interview.

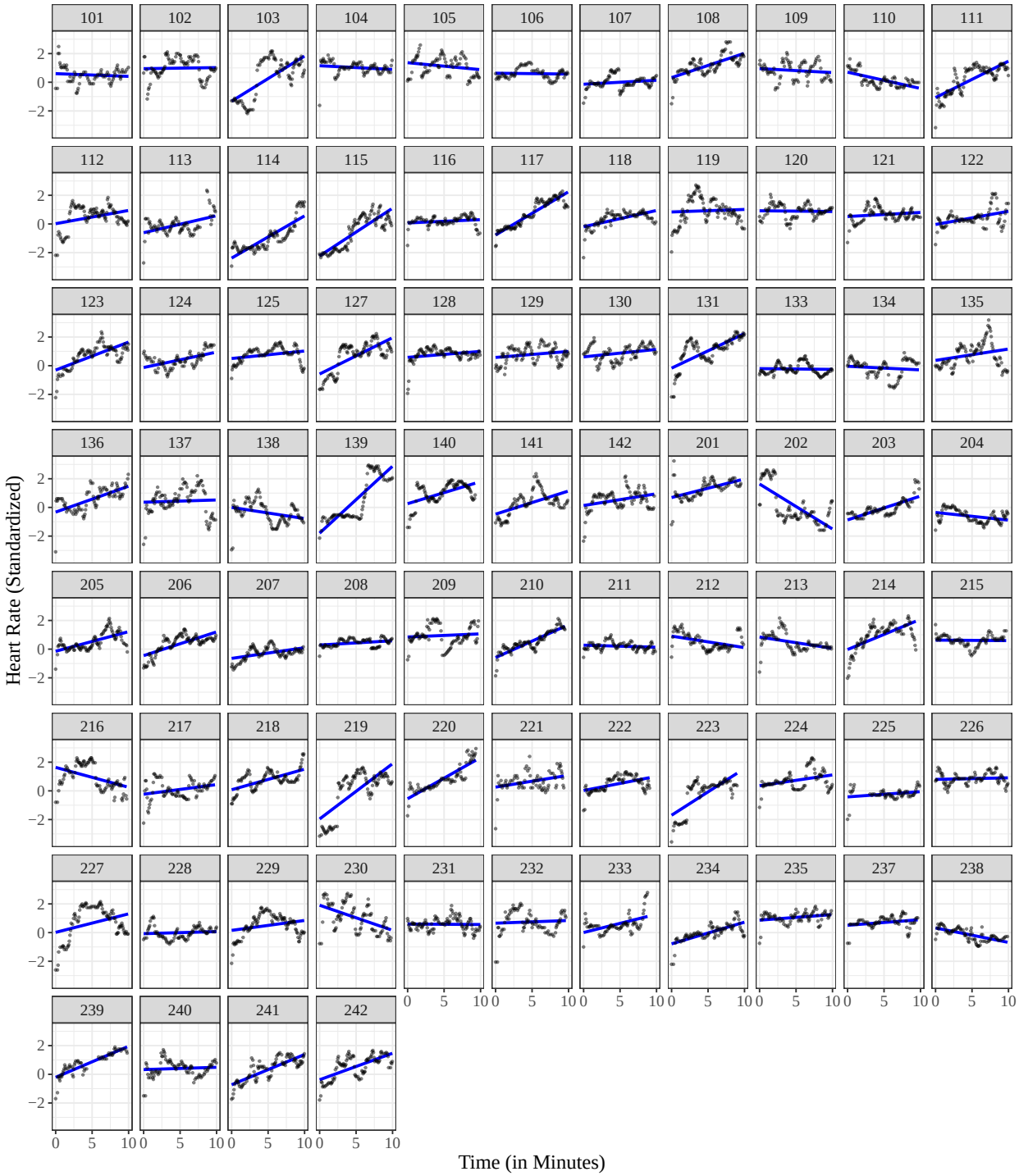


Figure A3: Linear estimation of individual HR changes over time during the preparation phase for $N = 81$ participants. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

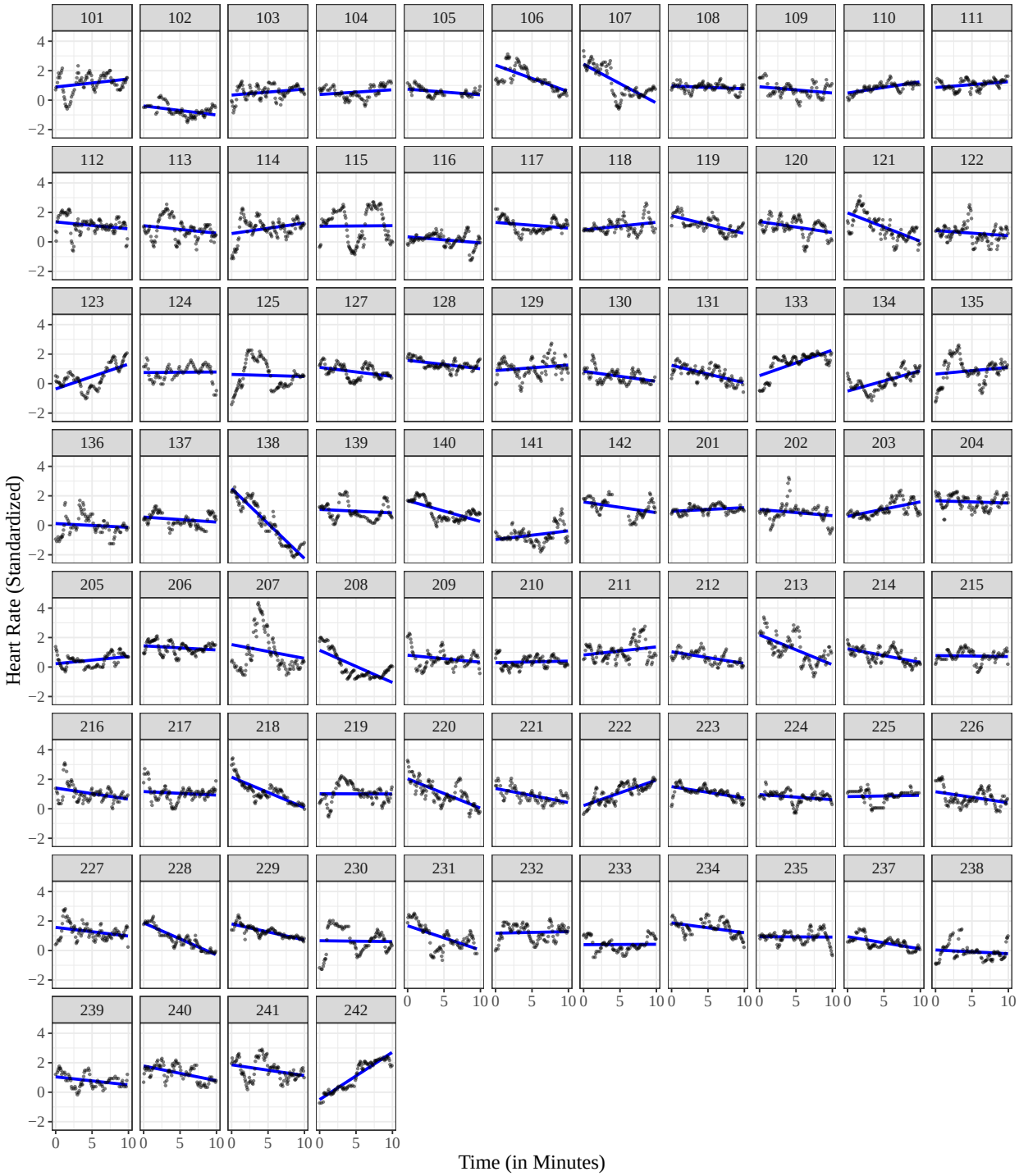


Figure A4: Linear estimation of individual HR changes over time during the teaching phase for $N = 81$ participants. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

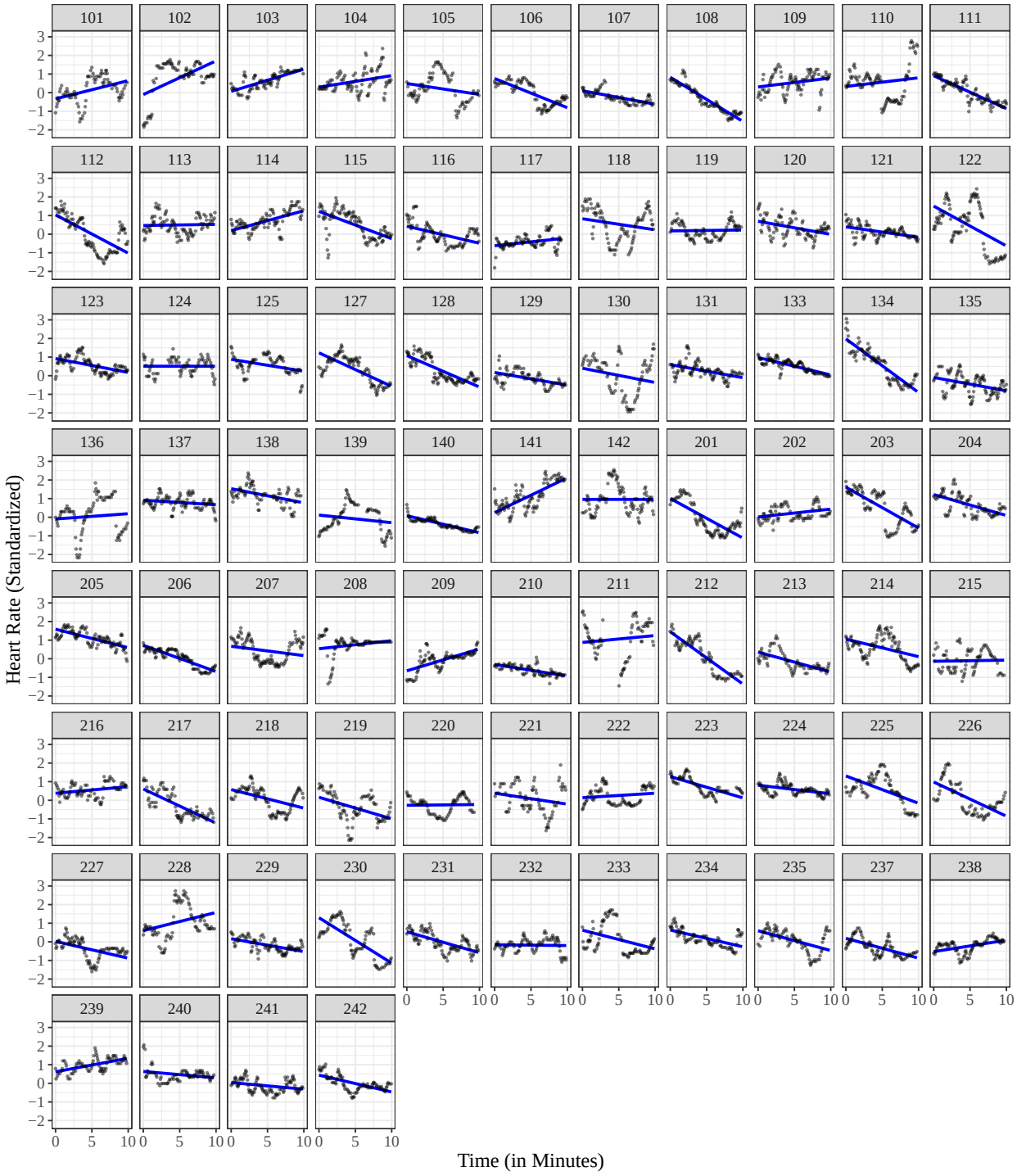


Figure A5: Linear estimation of individual HR changes over time during the post-teaching phase for $N = 81$ participants. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

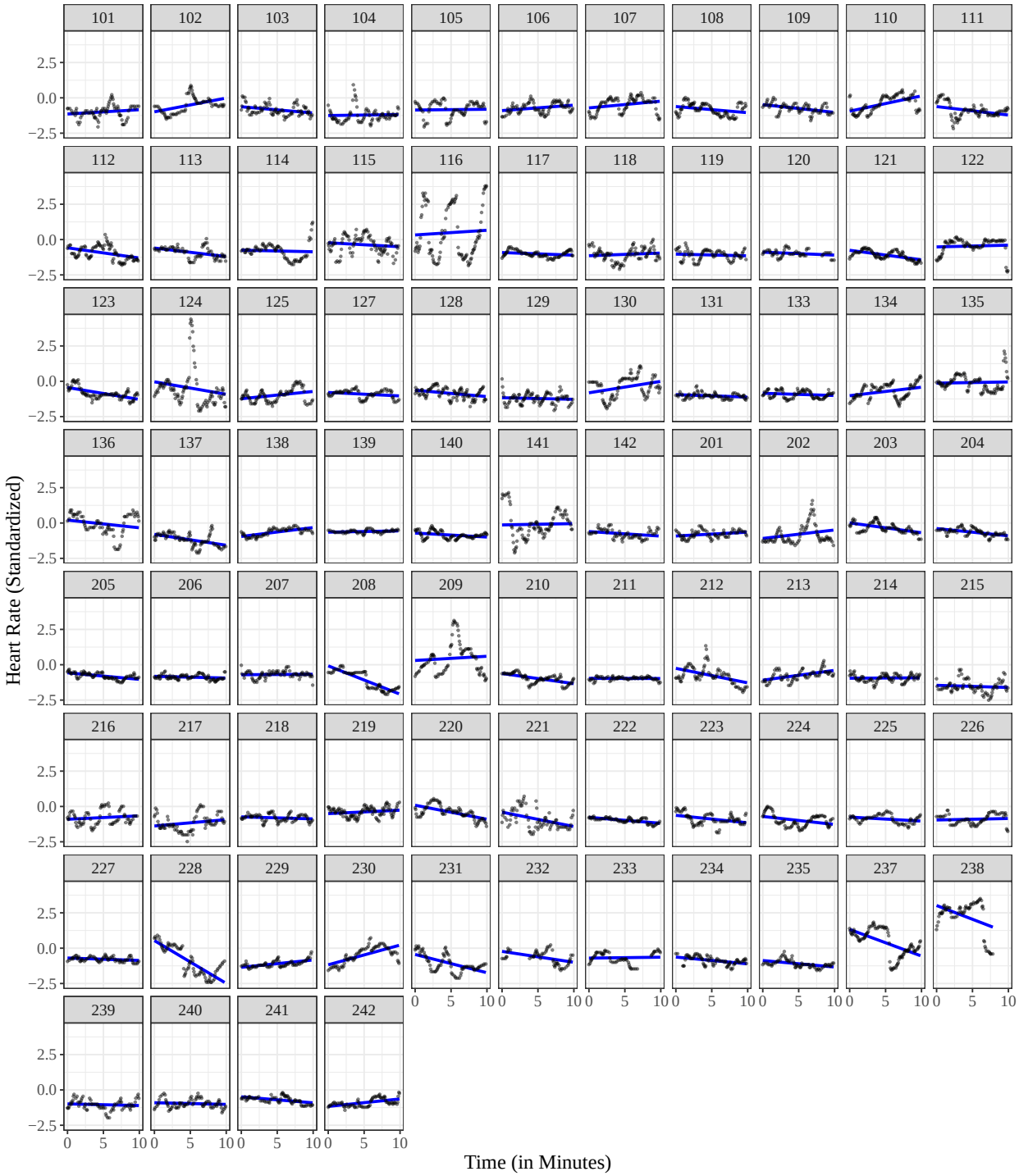


Figure A6: Linear estimation of individual HR changes over time during the interview phase for $N = 81$ participants. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

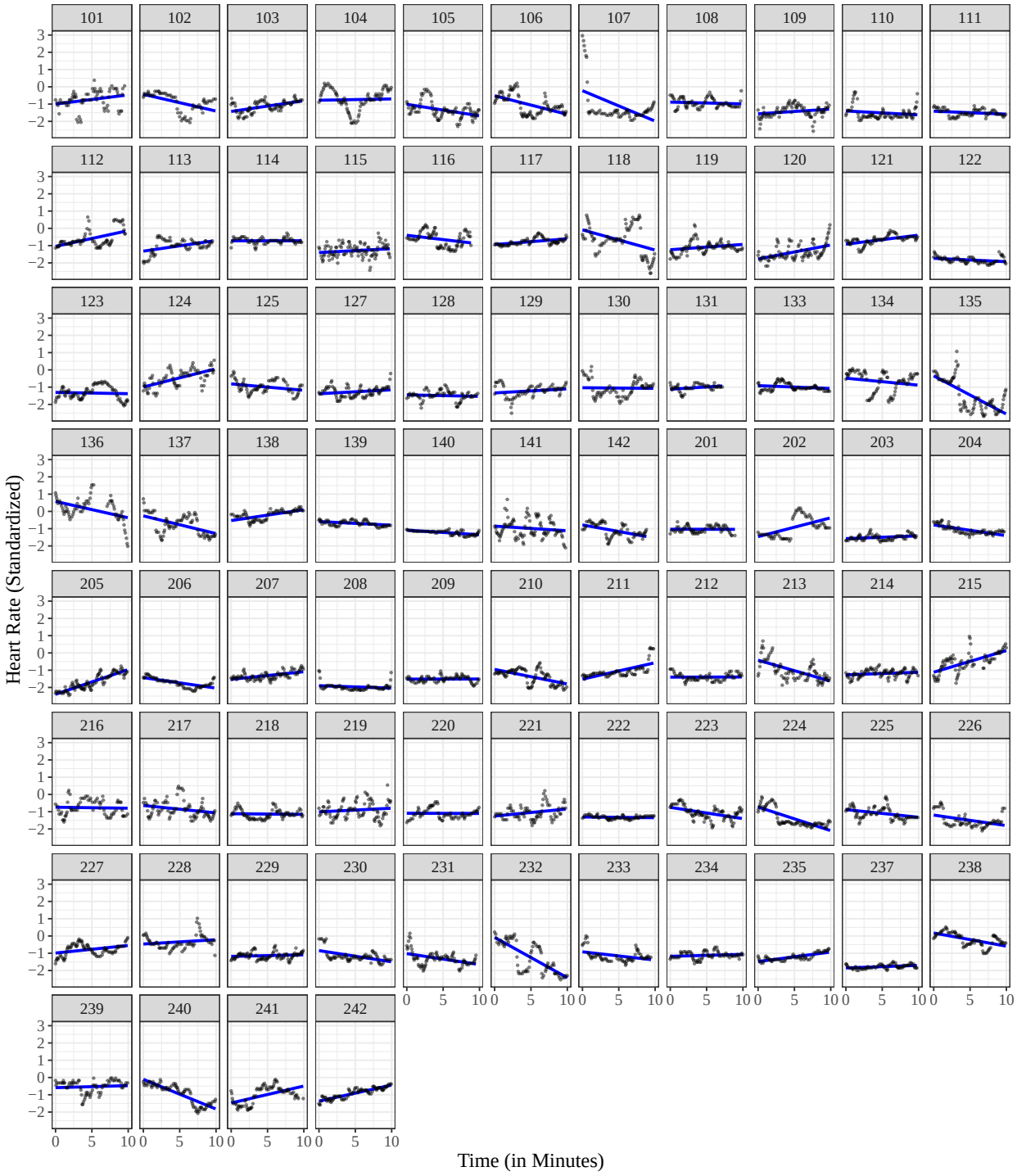


Figure A7: Linear estimation of individual HR changes over time during the end phase for $N = 81$ participants. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.