Hier fancy Titel überlegen für Expertise-Paper

# 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 demonstrated that the fitness tracker was 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 and 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:* classroom management, professional vision, eye-tracking, expertise differences, managing classroom disruptions

# Introduction

## Classroom Management

## Teachers’ Professional Knowledge

## Expertise Differences in Teachers’ Professional Vision

## Eye-Tracking to Assess Teachers’ Professional Vision

## Present Study

The present study aimed to investigate differences between experienced and inexperienced teachers in their gaze behavior during a micro-teaching unit involving classroom events, their self-evaluations of competencies in classroom disruption prevention and management, their subjective assessments of how disruptive the events were, and how confident they felt in dealing with them, and their strategic knowledge of classroom management. To address these objectives, data were analyzed from both in-service (experienced) and pre-service (inexperienced) teachers who participated in the laboratory-based study “Professional Vision of Novice and Expert Teachers” (ProVisioNET), which focused on developing classroom management skills.

Participants individually attended a lab session and conducted a brief micro-teaching unit for a small “class” of three trained actors simulating typical and potentially disruptive classroom behaviors. During the session, teachers’ gaze patterns were recorded using eye-tracking technology, while their evaluations of classroom management and strategic knowledge were assessed through self-report questionnaires, an interview, and a test.

The study had three primary aims:

1. The first research goal was to examine teachers’ gaze behavior during classroom interactions in a lab-based teaching task lasting approximately 15 minutes and to explore differences between experienced and inexperienced teachers in their visual attention. It was expected that all participants would focus more frequently on relevant areas, such as students, rather than less relevant areas, such as the classroom environment or teaching materials (**Hypothesis 1**). These differences were expected to be reflected in eye-tracking parameters, including the number and duration of fixations.
2. The second research goal was to investigate global gaze behavior, represented by the Gaze Relational Index (GRI), calculated as the ratio of fixation count to fixation duration. The GRI served as an indicator of effective scanning behavior. Experienced teachers were expected to demonstrate more frequent but shorter fixations, resulting in a smaller GRI compared to inexperienced teachers (**Hypothesis 2**).
3. The third research goal was to explore the relationship between gaze efficiency, as indicated by the GRI, and other classroom management characteristics. Teachers who felt more disrupted by classroom behaviors were expected to show higher GRI values, reflecting less effective gaze behavior (**Hypothesis 3a**). Teachers who felt more confident in dealing with the disruptions were expected to show lower GRI values, reflecting more effective gaze behavior (**Hypothesis 3b**). Furthermore, lower GRI values were hypothesized to correlate with higher self-assessed competencies in classroom disruption prevention and management (**Hypothesis 3c**). Additionally, lower GRI values were expected to correlate with greater strategic knowledge of classroom management (**Hypothesis 2d**).

# Method

## Participants

We recruited a total of 84 teachers from Germany (42 pre-service teachers and 42 in-service teachers) through personal contacts, email lists, and flyers. Pre-service teachers were required to be actively enrolled in a teacher education program and to have completed their first internship, while in-service teachers needed to have completed both a teacher training program and a traineeship and to be currently working in the teaching profession. Data from two in-service teachers were excluded due to low-quality eye-tracking data, resulting in a final sample of 82 teachers, comprising 42 pre-service teachers and 40 in-service teachers.

The pre-service teachers (*n*= 29 women, *n*= 13 men) had a mean age of 22.80 years (*SD* = 1.90; range: 19–27). On average, they were in their 6.70 semesters (*SD* = 2.60; range: 3–11) and had an average of 9.60 hours (*SD* = 7.20; range: 1–36) of teaching experience through internships completed during their studies. Of the pre-service teachers, 21% were preparing to become primary school teachers, 60% were secondary school teachers, and 19% were special education teachers. Additionally, 88% were involved in extracurricular teaching activities, such as tutoring or homework supervision.

The in-service teachers (*n*= 24 women, *n*= 16 men) had a mean age of 39.10 years (*SD* = 10.60; range: 26–60) and an average of 11.60 years (*SD* = 11.30; range: 1–38) of teaching experience. Among these teachers, 10% taught at primary schools, 85% at secondary schools, and 5% at vocational schools. Furthermore, 52% were also involved in secondary teaching roles, such as university lecturers, main training supervisors for trainee teachers, or subject advisers.

Data collection occurred between June 2021 and June 2023, primarily in a seminar room at the faculty. However, for a subset of experienced teachers, data was gathered in school classrooms, while maintaining the same controlled laboratory conditions.

The study adhered to ethical guidelines and received approval from the University’s Institutional Review Board. Participants were fully informed about the study’s objectives before testing. Their participation was voluntary, without incentives, and commenced only after written consent.

## Setting and Procedure

Participants individually attended the lab for approximately two hours, following a standardized procedure for which a seminar room was transformed into a classroom. Upon arrival, they were welcomed by the experimenter, introduced to the procedure, and asked to sign the data protection agreement. Participants were then fitted with eye-tracking glasses, adjusted for comfort and vision (up to +/- five diopters). After performing an initial one-point calibration of the glasses (for details of the calibration, see Eye-tracking apparatus and calibration), the experimenter activated and synchronized the recording devices (eye-tracking glasses, four cameras, and an audio recorder) using an auditory signal. This setup phase included a brief introductory game (“Name Juggling”) to acclimate participants to the eye-tracking equipment, which took approximately 10-15 minutes.

After the initial setup, a second nine-point calibration was done in a separate room. As soon as the teacher re-entered the classroom, the micro-teaching unit started. Therefore, participants were asked to prepare a 15-minute lesson on a topic and grade level of their choice. The only requirement was that the unit had to be an introductory lesson, and had to consist of supervised individual work and/or frontal teaching. During the unit, three trained actors (playing students) performed scripted classroom disruptions, which occurred approximately every 1.5 minutes on a screen only visible to the “class” (e.g., chatting with a neighbor, heckling, looking at the phone; see Table A1 in the supplementary material for an overview and categorization of all events; and Figure B1 and B2 in the supplementary material for a depiction of the laboratory setting of the micro-teaching unit). The order of the disruptions and the performing students were fully balanced using Latin Squares. To capture teachers’ gaze patterns, the whole micro-teaching unit was recorded using eye-tracking glasses. The micro-teaching unit lasted about 15-20 minutes.

After the teaching session, participants repeated the nine-point calibration in the side room. During this time, the experimenter set up four letters A to C within the seminar room for a fixation task, which the participant performed after re-entering the room. Following this task, all recording devices were stopped, and participants filled out a brief computer-based questionnaire (~10-15 minutes) assessing sociodemographic data and a self-evaluation of their classroom management during the micro-teaching unit. A ten-minute break followed, concluding the first part of data collection.

In the study’s second phase, participants engaged in a Stimulated Recall Interview (SRI). They watched a video of their own teaching session, recorded through the eye-tracking glasses, while the experimenter paused the video at each classroom disruption. For each disruption, participants answered five open-ended questions and three rating questions, including confidence appraisal and disruption assessment (see Measures). The SRI lasted approximately 45-60 minutes. Finally, participants completed a digital Situational Judgment Test (SJT), assessing their strategic knowledge of classroom management. The questionnaire took approximately 15 minutes to complete, marking the end of the study.

## Eye-tracking apparatus and calibration

Teachers wore a binocular Tobii Pro Glasses 2 eye-tracker during the micro-teaching unit to record eye-tracking data. The system consisted of a wearable head unit and a recording unit. The head unit was a measuring device with different sensors. A high-definition scene camera captured a full HD video of the teacher’s field of vision. An integrated microphone recorded surrounding sounds. Infrared light illuminators supported the eye-tracking sensors which recorded the eye orientation to capture the teacher’s gaze point. The videos were recorded with a sampling rate of 50 Hz in a video resolution of 1920 x 1080 at 25 frames per second. The scene camera had a field of view of 90 degrees in 16:9 format (82 degrees horizontal and 52 degrees vertical) and a frame dimension of 179 x 159 x 57 mm (width x depth x height). The recording unit is a compact computer that manages the head unit. It captures and saves eye-tracking data, audio, and scene camera footage on a removable SD memory card.

The evaluation of the calibration process followed the guidelines outlined in Onkhar et al. (2024) for assessing calibration quality. Participants’ gaze was calibrated using a bullseye card that the participant held at arm’s length. A successful calibration was achieved when the participant’s gaze marker sufficiently overlapped with the bullseye for a specified period of time, based on criteria internally determined by the manufacturer’s software (Tobii Pro Glasses 2, 2023). All participants achieved successful calibration, and no participants were excluded due to calibration failure. The robustness of the calibration was further verified through a secondary nine-point calibration. During this step, participants were asked to read numbers from one to nine aloud and direct their gaze at specific fields corresponding to each number. The initial calibration and the verification were performed before and after each data collection session.

## Measures

***Gaze Behavior***

To examine gaze behavior, analyses were performed within predefined global Areas of Interest (AOIs; see Figure X). To structure the data meaningfully, two types of AOIs were defined: global AOIs and event-based AOIs. Global AOIs, such as the group of students or the teacher's materials, were defined to analyze gaze behavior over the entire duration of the video, with the teacher’s gaze points coded accordingly. In contrast, event-based AOIs, such as the "disruptive person," focused on specific disruption-related areas and were coded only during the duration of the disruption (Figure 6). These distinctions allowed for a detailed analysis of both overall monitoring behavior and targeted responses to disruptions.

#### Global Monitoring Gaze Behavior.

The global gaze monitoring behavior was assessed using three key eye-tracking metrics, recognized for their robustness and sensitivity to expertise in prior research (Grub et al., 2020). These metrics comprised fixation number, fixation duration, and the Gaze Relational Index (GRI; cf. Gegenfurtner et al., 2020). The GRI was calculated as the ratio of during micro-teaching unit (MTU), measuring cognitive focus or attention levels as a measure of expertise.

#### Event-related Gaze Behavior.

Event-related gaze behavior was measured using the number and duration of fixations on the AOI labeled “disruptive person.” Additionally, the time to first fixation on this AOI served as an indicator (see Figure 7). These metrics were assessed within specific AOIs representing teaching disruptions caused by students.

### Teaching Experience

Participants’ teaching experience was assessed as part of their sociodemographic data, with the duration of their work experience reported in years.

### Self-Evaluations of Competencies in Classroom Disruption Prevention and Management

After the micro-teaching unit, teachers answered a questionnaire using five items from a validated questionnaire (Helmke et al., 2013) and nine self-developed items derived from the research literature (Kiel et al., 2013; Kounin, 2006; Marzano, 2007) to assess teachers’ self-evaluated competencies in classroom disruption prevention and management. The questionnaire was a 4-point Likert scale (1 = strongly disagree; 4 = strongly agree).

### Subjective Disruption and Confidence Ratings of the Classroom Events

The subjective disruption and confidence ratings 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.

### Strategic Knowledge of Classroom Management

Teachers’ strategic knowledge of classroom management was assessed using a Situational Judgment Test (SJT; Gold & Holodynski, 2015) via an online questionnaire on SoSci Survey. Participants graded five to six action alternatives for twelve teaching scenarios in which classroom disruptions were discussed on a six-point Likert scale (grade 1 = “very good” to grade 6 = “unsatisfactory”). As the SJT was originally designed for primary schools, adjustments were made in order to be able to use the SJT for all types of schools in the *ProVisioNET* study. Due to their general applicability, all twelve scenarios and answer options were adopted and only the names of the class levels were removed from the questions - except for scenario 6, where this information was essential. Only fully completed questionnaires were included in the present study.

## Data analysis

The data were analyzed using R (RStudio Team, 2020) and IBM SPSS Statistics (Version 29). Graphics were created using ggplot2 (Wickham, 2016).

To test **Hypothesis 1a**, we averaged each person’s standardized HR over each of the five selected intervals[[1]](#footnote-2), resulting in one HR *level* measure per person per interval. We conducted a one-way Analysis of Variance with repeated measures as an omnibus test for HR level as dependent variable and the five intervals the repeated measures factor. Subsequently, we tested the mean differences between the *teaching interval (I2)* and the other four intervals by planned contrasts and computed the effect size *d* (Cohen, 1988).

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 (Gelman & Hill, 2006) for each interval to estimate intercepts and linear slopes for each individual, which were then averaged across individuals.[[2]](#footnote-3) 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 of HR level and changes. 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

## 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. Figure 3 a. and b. display the unstandardized and standardized HR trends, respectively, over the course of the entire study period. Results showed that 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, Standard Deviations HR, 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 | 90.09/0.04a | 15.76/0.99a | 51b/–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 |
| *Note*. Mean heart rate (*M* HR), standard deviations (*SD* H) of HR, and HR range (minimum and maximum values) are presented for the entire study and across five intervals in both unstandardized beats per minute (BPM) and z-standardized scores.  a Please note that standardized *M* and *SD* of the overall course were not exactly 0 and 1 due to rounding differences.  b Deviations in the minimum values between the overall course and the *end interval (I5)* can be explained by the fact that the study duration was capped at two hours for the overall analysis, while the end intervals were determined individually for each participant. This discrepancy arose because a few participants required more than two hours to complete the study, leading to differences in the recorded data. | | | | |

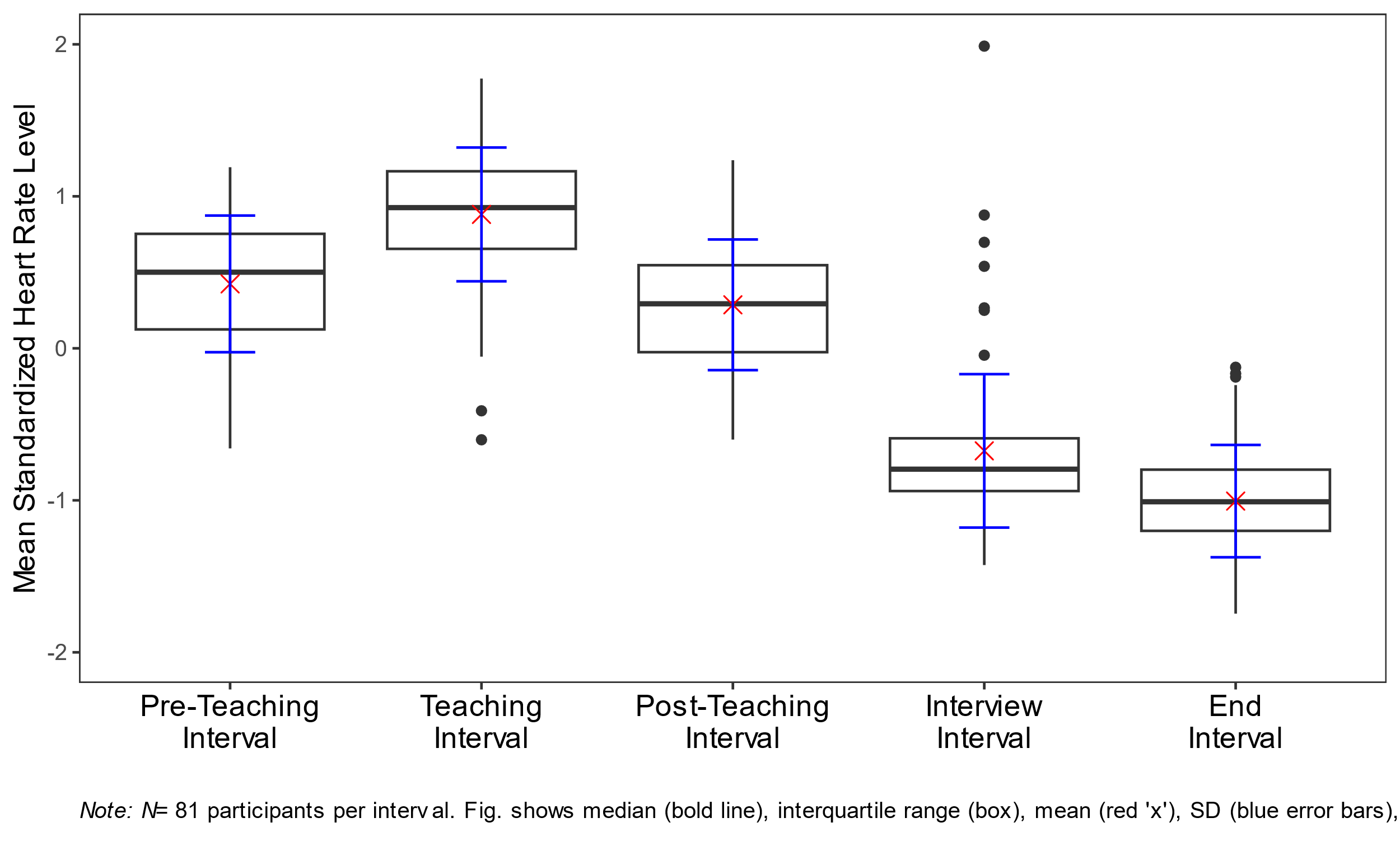
**Figure 3**

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

Figure 4 shows the distribution of teachers’ mean standardized HR *levels* 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 Mean Standardized Heart Rate Levels For The Five Intervals*



*Note*. *N* = 81 participants per interval. Figure 4 shows the median (bold line), interquartile range (the box spanning the 25th to 75th percentiles), the mean (red “x”), the standard deviation (blue vertical error bars), whiskers (lines extending to data points within 1.5 times the interquartile range), and outliers (individual dots beyond the whiskers).

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* *(Mean, Standard Deviation, and 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 |
| *Note. M* = Mean, *SD* = Standard Deviation, *p* = *p*-value.  \* *p* < .05. | | | | |

## Predicting mean standardized HR and mean slopes

Table 3 shows the correlations among mean standardized HR/mean slopes, 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). Correlations with HR measures were mostly very small and statistically non-significant. Correlations among teaching experience and appraisals (not shown in Table 3) were significant: more experienced teachers had lower disruption appraisals (*r* = –.36), and higher confidence appraisals (*r* = .44). Moreover, the two appraisal variables were negatively correlated (*r* = –.37).

**Table 3**

*Correlations Between Mean Standardized HR Level/Mean Slopes and the Variables Teaching Experience, Disruption Appraisal, and Confidence Appraisal, 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 |
| \* *p* < .05. | | | | | |

Table 4 shows the results of the regression analyses. Teaching experience significantly predicted mean standardized HR level only in the *interview interval* (Table 4, Interview interval, Model 1), indicating a higher mean standardized HR level for teachers with more teaching experience. This relationship is, in fact, in the opposite direction as predicted by **Hypothesis 2a**. For all intervals, neither adding disruption appraisal (**Hypothesis 2b**) nor 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 level 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 level 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 HR changes 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 4**

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Interval | Model 1 | | | | Model 2 | | | | | Model 3 | | | | | | Model 4 | | | |
|  | Mean std. HR level | | Mean slopes | | | Mean std. HR level | | Mean slopes | | | Mean std. HR level | | Mean slopes | | Mean std. HR level | | | 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)a** |  |  |  |  | |  |  |  |  | |  |  |  |  |  | |  |  |  |
| 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 level and mean slopes were predicted only by teaching experience. In Model 2, solely disruption appraisal was added to teaching experience as a second predictor. In Model 3, solely confidence appraisal was added to teaching experience as a second predictor. In Model 4, all three predictors were considered in concert.  a We calculated only Model 1 for the *pre-teaching interval* *(I1)* because the classroom events and their corresponding appraisals had not yet occurred in this interval.  \* *p* < .05. | | | | | | | | | | | | | | | | | | | |

# Discussion

## Key Findings

Our study investigated how data from a wrist-worn fitness tracker could reveal the effects of stressors, such as classroom disruptions, on teachers’ stress responses before, during, and after teaching sessions. Teachers’ HR was assessed using a Fitbit® in a five-phase lab study, including a micro-teaching unit with disruptive events. We also examined whether HR variance was explained by teaching experience and self-reported appraisals (disruption and confidence).

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 fitness tracker, we mapped teachers’ HR before, during, and after a stressful micro-teaching unit. HR increased in preparation for teaching, peaked while teaching, and decreased afterward, highlighting distinct physiological responses across different stages of the teaching process. Our findings are consistent with prior research that illustrates the variability of teachers’ HR in relation to their activities and the stressors they encounter. For instance, earlier studies demonstrated that HR levels increase when teachers are placed in exposed positions, such as when engaging in teacher-centered activities or managing challenging student behaviors (Donker et al., 2018; Junker et al., 2021; Scheuch & Knothe, 1997; Sperka & Kittler, 1995), as well as with findings showing how HR changes align with activating events and stress-inducing tasks (Chalmers et al., 2021; Darnell & Krieg, 2019).

Building on the model of teacher stress (Kyriacou & Sutcliffe, 1978; see Figure 1), we 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 effects of teaching experience on teachers’ HR, i.e., more experienced teachers did not show lower HR levels during the stressful teaching phase than less experienced teachers. 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 (Uchino et al., 2010). However, to avoid this kind of confounding influence, we had not used raw BPM data but standardized mean HR data for 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 (Kirschner et al., 2016). Rather, honing skills from professional experience necessitates a deliberate practice of choosing to improve, learning through experience, and integrating new knowledge into future performances (Dunn & Shriner, 1999). 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, our hypotheses were not supported, as neither confidence appraisals nor disruptiveness appraisals predicted teachers’ HR 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 (Kyriacou & Sutcliffe, 1978). In addition, while HR was assessed real-time during teaching, self-reported appraisals were given in retrospect during the interview following the teaching unit, 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 is consistent with the idea that differences in HR levels, as an indicator of the physiological stress response, can be linked to the cognitive appraisal of stressors.

## 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 (Beaty-O’Ferrall et al., 2010; Rüedi, 2014). 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, the onset of disruptive student behavior was scripted, following an experimental time schedule, which was not affected by 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 are some shortcomings of this type of assessment method. First, while fitness trackers typically yield HR data, heart rate variability (HRV) has been demonstrated to be an even more accurate indicator of stress (Wettstein et al., 2020). While standard fitness trackers did not provide this measure at the time of our data collection, more recent products do offer this function. Thus, we encourage future studies to consider assessing HRV in addition to 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 (Heneghan et al., 2019; Nanchen, 2018). 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 (Sammito et al., 2015), 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 fulfilled 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 might vary. Research on the reliability of the deployed Fitbit® device has proven that this brand is generally accurate in controlled settings and for moderate activity levels (Fuller et al., 2020; Hajj-Boutros et al., 2023; Jo et al., 2016; Wallen et al., 2016), as it was the case in our study. For example, the Fitbit® fitness tracker had previously shown good HR measurement accuracy during resting phases (Jo et al., 2016; Muggeridge et al., 2021) and for activities such as walking, jogging, and running (Hajj-Boutros et al., 2023). At higher exercise intensities such as cycling, the Fitbit® tracker may underestimate HR (Jachymek et al., 2022; Jo et al., 2016; Montoye et al., 2017; Thomson et al., 2019) but is still within an acceptable range according to systematic reviews (Chevance et al., 2022). Nevertheless, Gagnon et al. (2022) stressed that Fitbit® trackers cannot replace ECG when high precision measurement is paramount. Despite these considerations, the Fitbit® model appeared 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 (Razavi, 2001) or recall errors (Van den Bergh & Walentynowicz, 2016, p. 201). 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 (Lazarus, 1990). While SRIs provide a more detailed and reflective understanding of the stressor in question, the delayed nature of the response made it difficult to capture the immediate, in-the-moment appraisal that occurred when the stressful event actually took place.

## Hands-on advice for using wrist-worn fitness trackers for research

For researchers aiming to collect data using fitness trackers, there are practical aspects to consider regarding the design, data collection procedure, and data analysis phases of research projects (for an additional overview, see Nelson et al., 2020):

1. Choosing a suitable fitness tracker model:

Before data collection, researchers need to decide which model of fitness tracker best suits their research questions. 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 high measurement accuracy, such as in medical contexts, as they cannot replace the accuracy of ECG measurements (Gagnon et al., 2022). 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 (Jachymek et al., 2022; Jo et al., 2016; Montoye et al., 2017; Thomson et al., 2019). For reference, the systematic review by Fuller et al. (2020) 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 researchers operating with moderate budgets or if larger groups of participants are needed to be tracked at the same time. Furthermore, before conducting any study, it should be considered that the data collected with fitness trackers are health data, and therefore very sensitive personal data. 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.

1. Operating the fitness tracker:

When 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 sizes. 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 installed, 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 time 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.

1. 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 as .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 of time 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.

## 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 understanding of teacher stress. Results suggest that the widespread availability of HR data from wearable fitness trackers 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; Agyapong et al., 2023). 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 Wettstein's et al. (2021) call for the use of ambulatory assessment methods, particularly in the context of classroom management research, 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 stress. By moving “from heartbeat to data”, we can harness the power of wearable technology to provide teachers with the tools needed to better understand and manage their stress, ultimately enhancing their overall well-being.

# References

Agyapong, B., Brett-MacLean, P., Burback, L., Agyapong, V. I. O., & Wei, Y. (2023). Interventions to reduce stress and burnout among teachers: A scoping review. *International Journal of Environmental Research and Public Health*, *20*(9), 5625. https://doi.org/10.3390/ijerph20095625

Beaty-O’Ferrall, M. E., Green, A., & Hanna, F. (2010). Classroom Management Strategies for Difficult Students: Promoting Change through Relationships. *Middle School Journal (J1)*, *41*(4), 4–11.

Chalmers, T., Hickey, B. A., Newton, P., Lin, C., Sibbritt, D., McLachlan, C. S., Clifton-Bligh, R., Morley, J., & Lal, S. (2021). Stress watch: The use of heart rate and heart rate variability to detect stress: A pilot study using smart watch wearables. *Sensors*, *22*(1), 151. https://doi.org/10.3390/s22010151

Chevance, G., Golaszewski, N. M., Tipton, E., Hekler, E. B., Buman, M., Welk, G. J., Patrick, K., Godino, J. G., & others. (2022). Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing Fitbits against reference measures: Systematic review and meta-analysis. *JMIR mHealth and uHealth*, *10*(4), e35626. https://doi.org/10.2196/35626

Cohen, J. (1988). Statistical power for the behavioural sciences. Hilsdale. *NY: Lawrence Erlbaum*, *58*(1), 7–19.

Darnell, D. K., & Krieg, P. A. (2019). Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures. *PLOS ONE*, *14*(12), e0225709. https://doi.org/10.1371/journal.pone.0225709

Donker, M. H., van Gog, T., & Mainhard, M. T. (2018). A Quantitative Exploration of Two Teachers with Contrasting Emotions: Intra-Individual Process Analyses of Physiology and Interpersonal Behavior. *Frontline Learning Research*, *6*(3), 162–184.

Dunn, T. G., & Shriner, C. (1999). Deliberate practice in teaching: What teachers do for self-improvement. *Teaching and Teacher Education*, *15*(6), 631–651. https://doi.org/10.1016/S0742-051X(98)00068-7

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., Buote, R., Heerden, D. V., Luan, H., Cullen, K., Slade, L., & Taylor, N. G. A. (2020). Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. *JMIR mHealth and uHealth*, *8*(9), e18694. https://doi.org/10.2196/18694

Gagnon, J., Khau, M., Lavoie-Hudon, L., Vachon, F., Drapeau, V., & Tremblay, S. (2022). Comparing a Fitbit Wearable to an Electrocardiogram Gold Standard as a Measure of Heart Rate Under Psychological Stress: A Validation Study. *JMIR Formative Research*, *6*(12), e37885. https://doi.org/10.2196/37885

Gegenfurtner, A., Boucheix, J.-M., Gruber, H., Lehtinen, E., & Lowe, R. K. (2020). *The Gaze Relational Index as a Measure of Visual Expertise*. *3*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. https://doi.org/10.1017/CBO9780511790942

Gold, B., & Holodynski, M. (2015). Development and Construct Validation of a Situational Judgment Test of Strategic Knowledge of Classroom Management in Elementary Schools. *Educational Assessment*, *20*(3), 226–248. https://doi.org/10.1080/10627197.2015.1062087

Grub, Ann-Sophie, Biermann, Antje, & Brünken, Roland. (2020). *Process-based measurement of professional vision of (prospective) teachers in the field of classroom management. A systematic review*. https://doi.org/10.25656/01:21187

Hajj-Boutros, G., Landry-Duval, M., Comtois, A. S., Gouspillou, G., & Karelis, A. D. (2023). Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense. *European Journal of Sport Science*, *23*(2), 165–177. https://doi.org/10.1080/17461391.2021.2023656

Helmke, A., Schrader, F.-W., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., & Ade-Thurow, M. (2013). *Basisfragebogen EMU - Evidenzbasierte Methoden der Unterrichtsentwicklung*. Unterrichtsdiagnostik. http://www.unterrichtsdiagnostik.info/downloads/fragebogen/

Heneghan, C., Venkatraman, S., & Russell, A. (2019). Investigation of an estimate of daily resting heart rate using a consumer wearable device. *medRxiv*, 19008771. https://doi.org/10.1101/19008771

Jachymek, M., Jachymek, M. T., Kiedrowicz, R. M., Kaźmierczak, J., Płońska-Gościniak, E., & Peregud-Pogorzelska, M. (2022). Wristbands in Home-Based Rehabilitation—Validation of Heart Rate Measurement. *Sensors*, *22*(1), Article 1. https://doi.org/10.3390/s22010060

Jo, E., Lewis, K., Directo, D., Kim, M. J., & Dolezal, B. A. (2016). Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. *Journal of Sports Science & Medicine*, *15*(3), 540–547.

Junker, R., Donker, M. H., & Mainhard, T. (2021). Potential classroom stressors of teachers: An audiovisual and physiological approach. *Learning and Instruction*, *75*, 101495. https://doi.org/10.1016/j.learninstruc.2021.101495

Kiel, E., Frey, A., & Weiß, S. (2013). *Trainingsbuch Klassenführung*. Verlag Julius Klinkhardt.

Kirschner, S., Sczudlek, M., Tepner, O., Borowski, A., Fischer, H. E., Lenske, G., Leutner, D., Neuhaus, B. J., Sumfleth, E., Thillmann, H., & others. (2016). Professionswissen in den Naturwissenschaften (ProwiN). In *Entwicklung von Professionalität pädagogischen Personals: Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (pp. 113–130). Springer. https://doi.org/10.1007/978-3-658-07274-2\_7

Kounin, J. S. (2006). *Techniken der Klassenführung*. Waxmann Verlag.

Kyriacou, C., & Sutcliffe, J. (1978). Teacher stress: Prevalence, sources, and symptoms. *British Journal of Educational Psychology*, *48*(2), 159–167. https://doi.org/10.1111/j.2044-8279.1978.tb02381.x

Lazarus, R. S. (1990). Theory-based stress measurement. *Psychological Inquiry*, *1*(1), 3–13. https://doi.org/10.1207/s15327965pli0101\_1

Lohmann, G., & Meyer, H. (2003). *Mit Sch"ulern klarkommen: Professioneller Umgang mit Unterrichtsst"orungen und Disziplinkonflikten*. Cornelsen-Scriptor.

Marzano, R. J. (2007). *The Art and Science of Teaching: A Comprehensive Framework for Effective Instruction*. ASCD.

Montoye, A. H. K., Mitrzyk, J. R., & Molesky, M. J. (2017). Comparative accuracy of a wrist-worn activity tracker and a smart shirt for physical activity assessment. *Measurement in Physical Education and Exercise Science*, *21*(4), 201–211. https://doi.org/10.1080/1091367X.2017.1331166

Muggeridge, D. J., Hickson, K., Davies, A. V., Giggins, O. M., Megson, I. L., Gorely, T., & Crabtree, D. R. (2021). Measurement of Heart Rate Using the Polar OH1 and Fitbit Charge 3 Wearable Devices in Healthy Adults During Light, Moderate, Vigorous, and Sprint-Based Exercise: Validation Study. *JMIR mHealth and uHealth*, *9*(3), e25313. https://doi.org/10.2196/25313

Nanchen, D. (2018). Resting heart rate: What is normal? In *Heart* (Vol. 104, Issue 13, pp. 1048–1049). BMJ Publishing Group Ltd and British Cardiovascular Society. https://doi.org/10.1136/heartjnl-2017-312731

Nelson, B. W., Low, C. A., Jacobson, N., Areán, P., Torous, J., & Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *NPJ Digital Medicine*, *3*(1), 90. https://doi.org/10.1038/s41746-020-0297-4

Onkhar, V., Dodou, D., & de Winter, J. C. F. (2024). Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. *Behavior Research Methods*, *56*(5), 4221–4238. https://doi.org/10.3758/s13428-023-02173-7

Razavi, T. (2001). *Self-report measures: An overview of concerns and limitations of questionnaire use in occupational stress research*.

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC.

Rüedi, J. (2014). Zur Bedeutung positive Beziehungen für die Klassenführung und den Umgang mit Unterrichtsstörungen. *Beziehungen in Schule Und Unterricht. Teil*, *3*, 105–126.

Sammito, S., Thielmann, B., Seibt, R., Klussmann, A., Weippert, M., & Böckelmann, I. (2015). Guideline for the application of heart rate and heart rate variability in occupational medicine and occupational science. *ASU Int*, *2015*(06), 1–29. https://doi.org/10.17147/ASUI.2015-06-09-03

Scheuch, K., & Knothe, M. (1997). Psychophysische beanspruchung von lehrern in der unterrichtstätigkeit. *Jahrbuch Für Lehrerforschung*, *1*(S 285), 299.

Sperka, M., & Kittler, U. . (1995). Psychophysiologische Beanspruchung von Lehramtskandidatinnen und -kandidaten im Schulunterricht. In K. Bräuer (Ed.), *Psychische Potentiale für eine interdisziplinäre Lehrerausbildung: Motivation—Kognition – Entwicklung* (pp. 182–197). Die Blaue Eule.

Thomson, E. A., Nuss, K., Comstock, A., Reinwald, S., Blake, S., Pimentel, R. E., Tracy, B. L., & Li, K. (2019). Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *Journal of Sports Sciences*, *37*(12), 1411–1419. https://doi.org/10.1080/02640414.2018.1560644

Uchino, B. N., Birmingham, W., & Berg, C. A. (2010). Are Older Adults Less or More Physiologically Reactive? A Meta-Analysis of Age-Related Differences in Cardiovascular Reactivity to Laboratory Tasks. *The Journals of Gerontology: Series B*, *65B*(2), 154–162. https://doi.org/10.1093/geronb/gbp127

Van den Bergh, O., & Walentynowicz, M. (2016). Accuracy and bias in retrospective symptom reporting. *Current Opinion in Psychiatry*, *29*(5), 302–308. https://doi.org/10.1097/YCO.0000000000000267

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., & Coombes, J. S. (2016). Accuracy of heart rate watches: Implications for weight management. *PloS One*, *11*(5), e0154420. https://doi.org/10.1371/journal.pone.0154420

Wettstein, A., Kühne, F., Tschacher, W., & La Marca, R. (2020). Ambulatory assessment of psychological and physiological stress on workdays and free days among teachers. A preliminary study. *Frontiers in Neuroscience*, *14*, 495378. https://doi.org/10.3389/fnins.2020.00112

Wettstein, A., Schneider, S., grosse Holtforth, M., & La Marca, R. (2021). Teacher Stress: A Psychobiological Approach to Stressful Interactions in the Classroom. *Frontiers in Education*, *6*. https://www.frontiersin.org/articles/10.3389/feduc.2021.681258

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-98141-3

# Appendix

# Appendix A

**Table A1**

*Classification of Nine Typical Classroom Disruptions According to Lohmann & Meyer (2003) Performed in The Micro-Teaching Unit by Actors*

|  |  |  |
| --- | --- | --- |
| 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 |

*Note.* Disruptions were classified based on the typology provided by Lohmann & Meyer (2003)*.* Categories include verbal, physical, and disengagement-related behaviors performed during the micro-teaching unit. The order of the performing actors and the disruptions was fully balanced using Latin squares.

# Appendix B

# Laboratory Setting of The Study

**Figure B1**

*Laboratory Setting of The Micro-Teaching Unit. Ein Bild, das Mobiliar, Stuhl, Kleidung, Schuhwerk enthält.

Automatisch generierte Beschreibung*

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

**Figure B2**

*Laboratory Setting of The Interview.*

Ein Bild, das Mobiliar, Zeichnung, Entwurf, Tisch enthält.

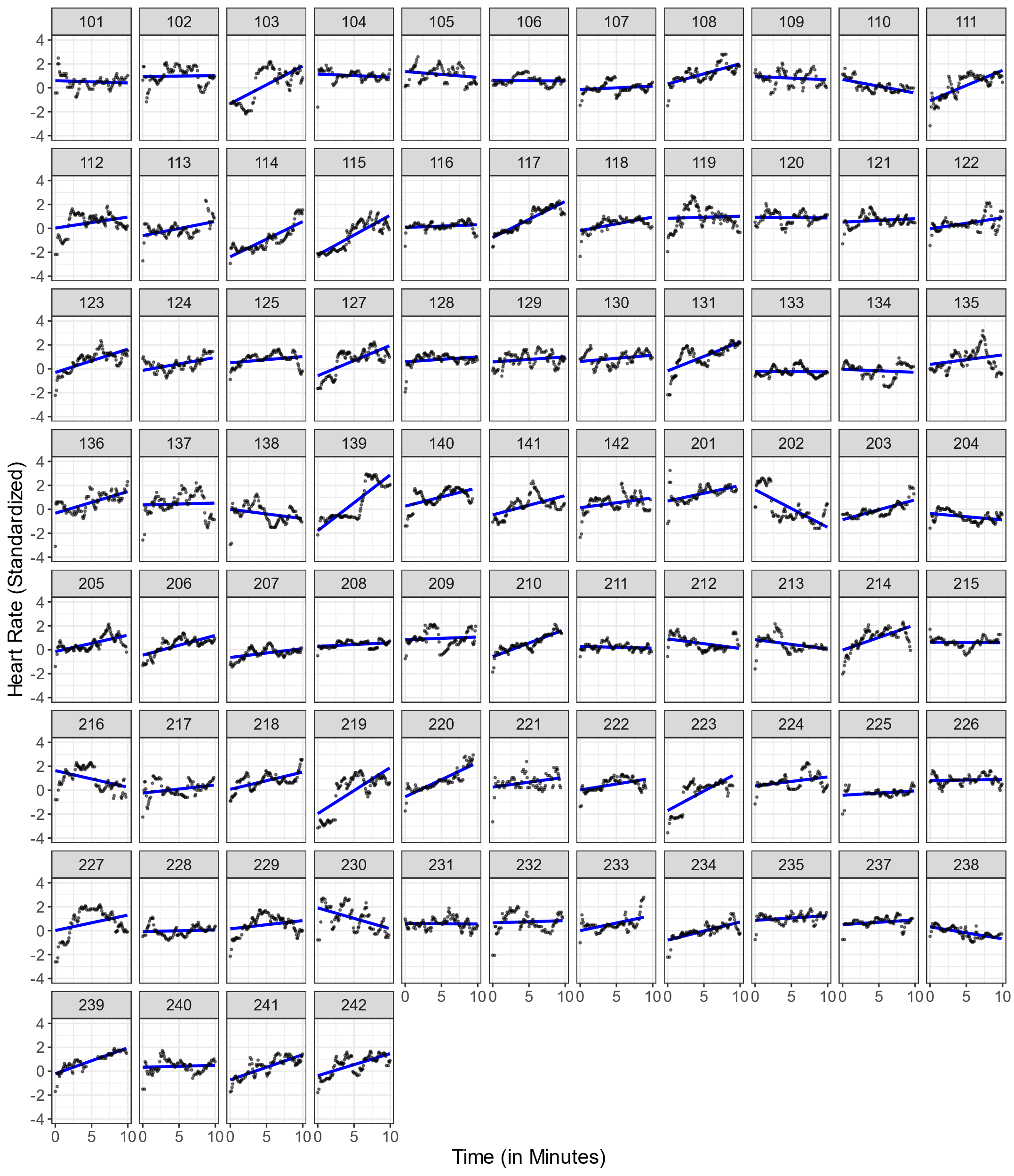
Automatisch generierte Beschreibung

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

# Appendix C

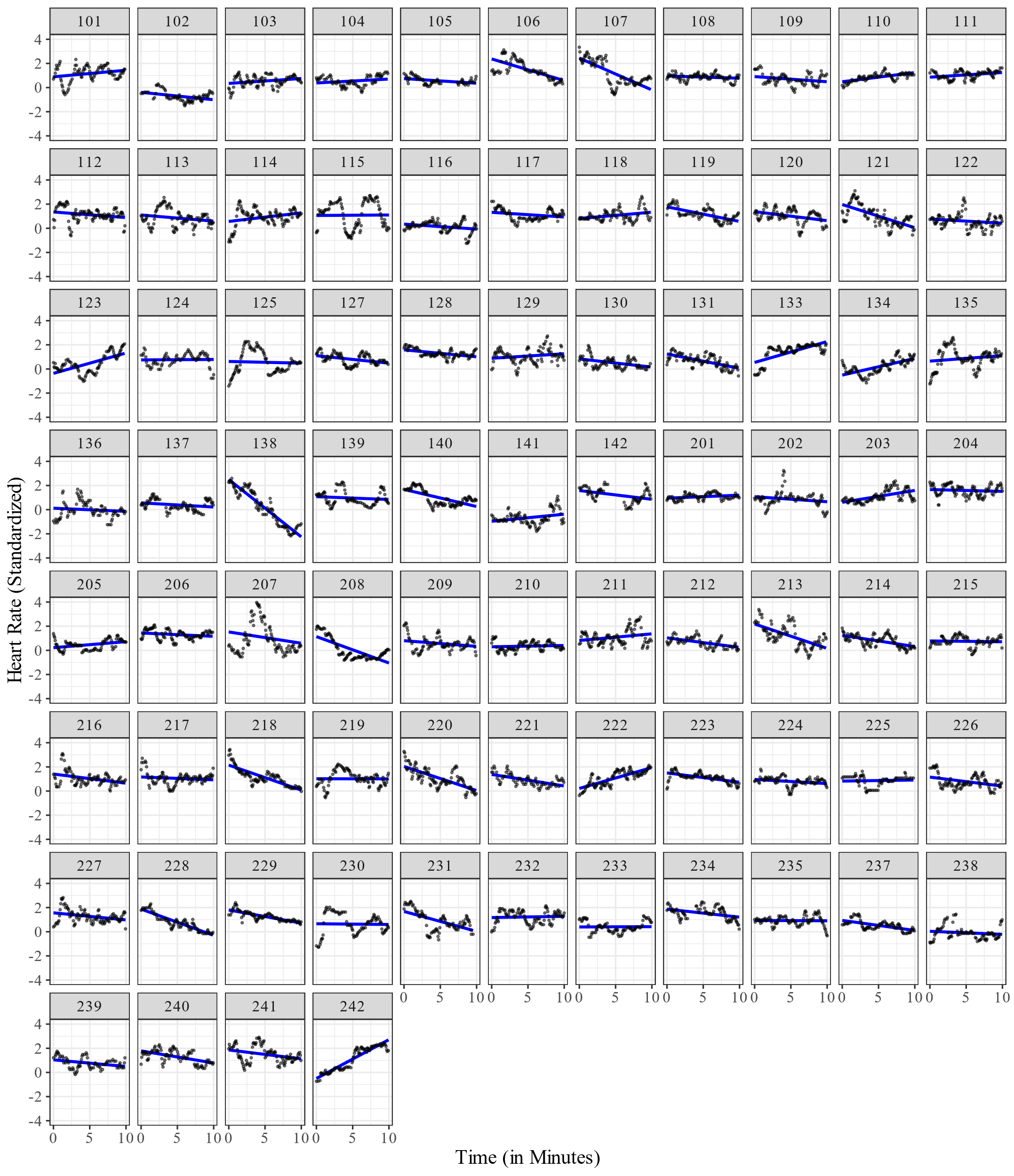
# Linear Estimation of Individual HR Changes For Five Intervals

**Figure C1**

*Linear Estimation of Individual HR Changes Over Time During The* Pre-Teaching Phase *For* N *= 81 Participants.*

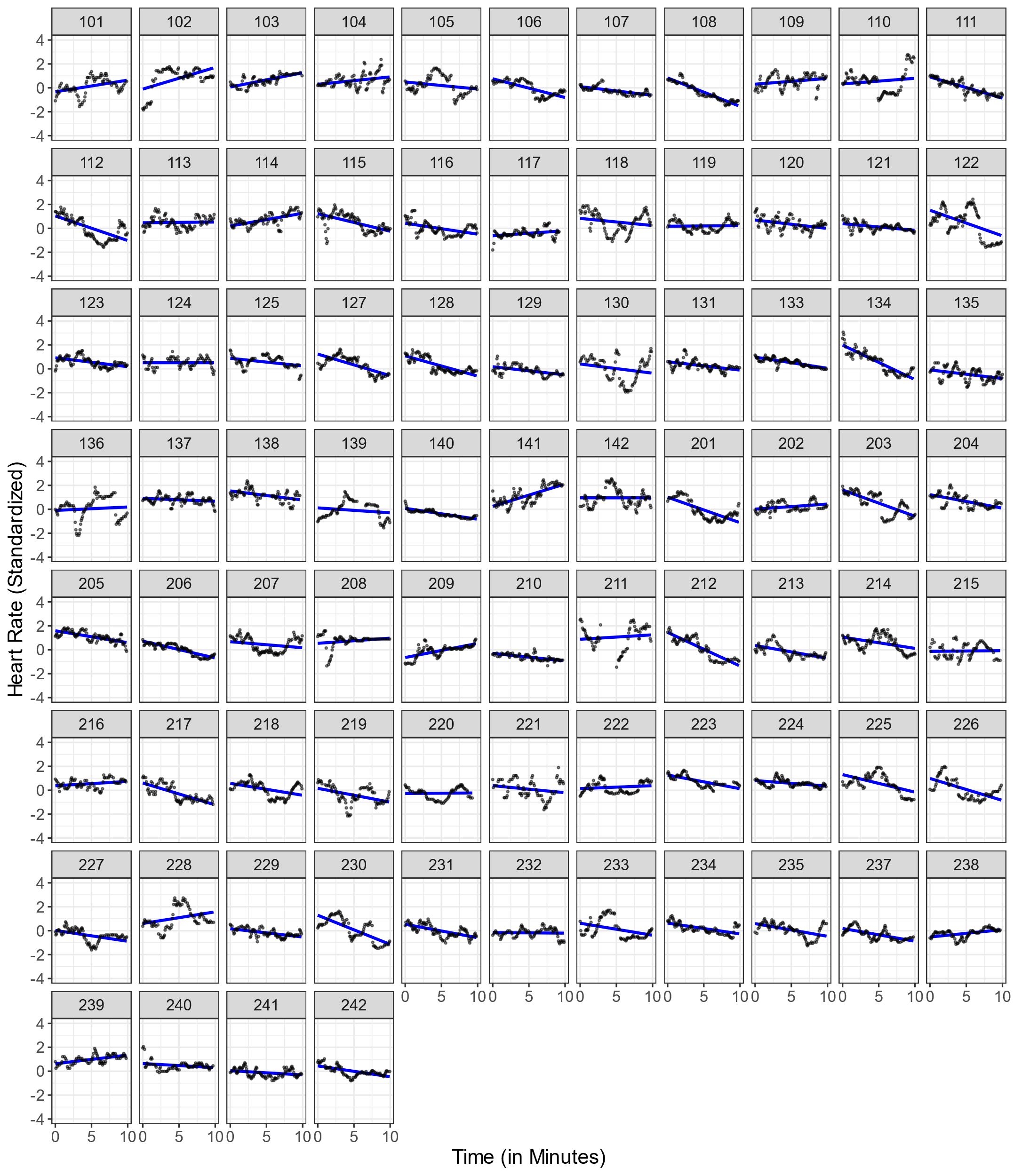
*Note.* Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

***Figure C2***

*Linear Estimation of Individual HR Changes Over Time During The Teaching Phase For N = 81 Participants.*

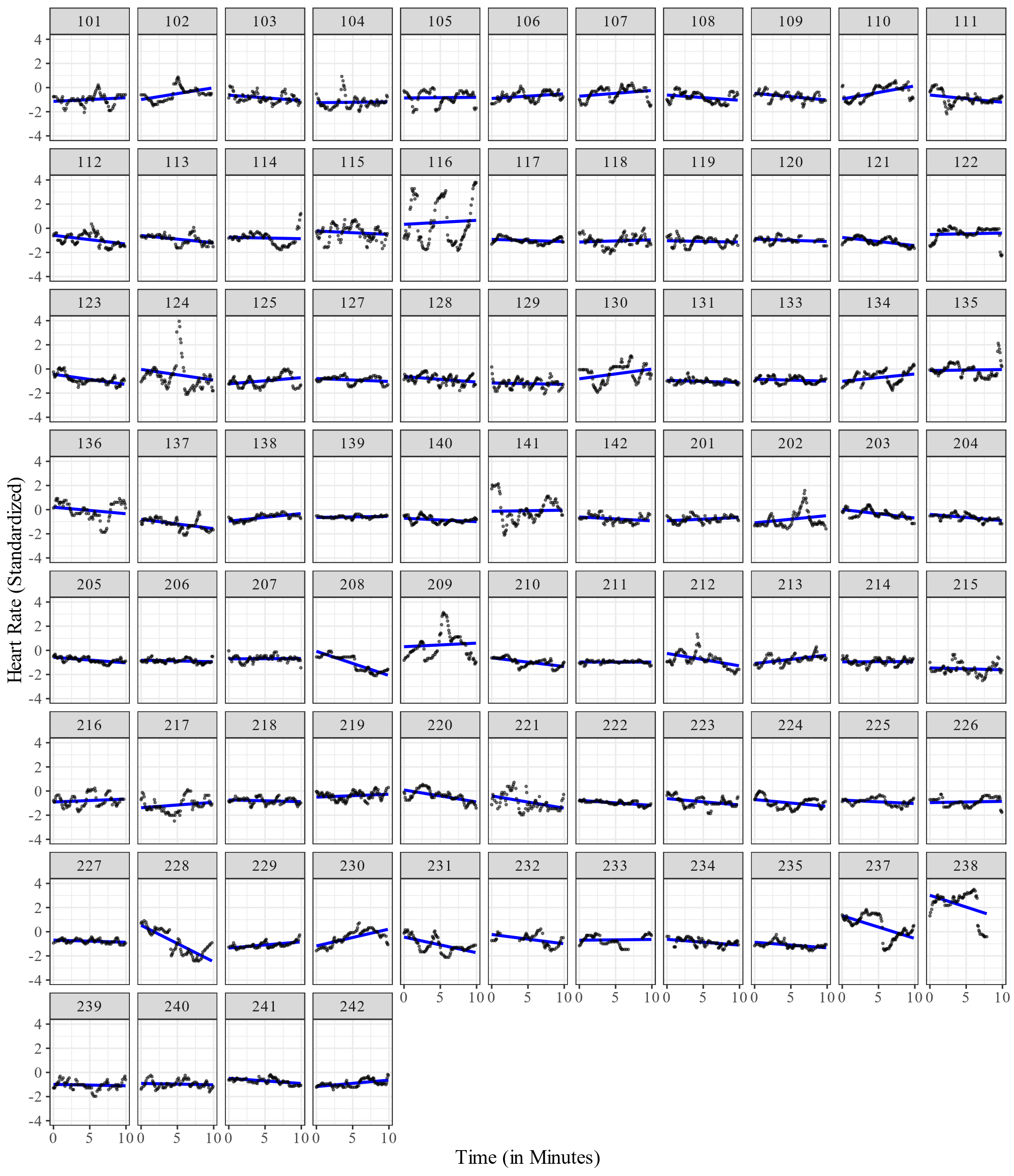
*Note.* Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

**FigureC3**

*Linear Estimation of Individual HR Changes Over Time During The* Post-Teaching Phase *For* N *= 81 Participants.*

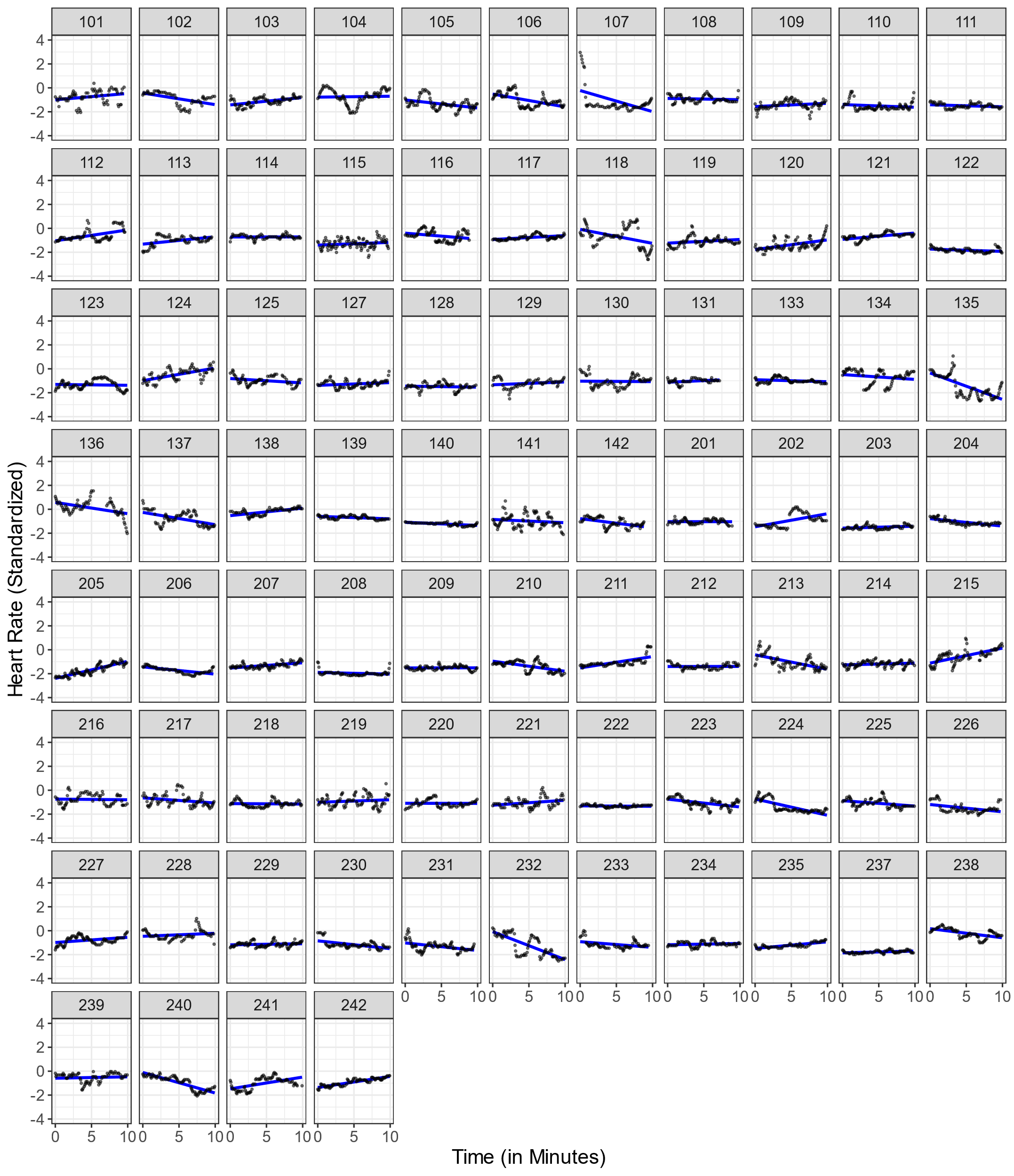
*Note*. Each plot illustrates the mean standardized HR values (y-axis) across 10 minutes (x-axis), with the black dots representing observed HR data points and the blue line showing the estimated linear trend.

**Figure C4**

*Linear Estimation of Individual HR Changes Over Time During The* Interview Phase *For* N *= 81 Participants.*

*Note.* 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 C5**

*Linear Estimation of Individual HR Changes Over Time During The* End Phase *For* N *= 81 Participants.*

*Note.* 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.

1. 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-2)
2. Although this procedure does not account for nonmonotonic progressions in individual HR, a graphical evaluation revealed that the linear estimates corresponded fairly well to the majority of the cases (see Appendix C in the supplementary material). [↑](#footnote-ref-3)