Class size affects preservice teachers' physiological and psychological stress reactions: An experiment in a virtual reality classroom

Article *in* Computers & Education · March 2022

DOI: 10.1016/j.compedu.2022.104503

4 authors:

Yizhen Huang
Universität Potsdam
6 PUBLICATIONS 34 CITATIONS
SEE PROFILE

Eric Richter
Universität Potsdam
23 PUBLICATIONS 109 CITATIONS
SEE PROFILE

Thilo Kleickmann
Christian-Albrechts-Universität zu Kiel
56 PUBLICATIONS 1,862 CITATIONS
SEE PROFILE

Dirk Richter
Universität Potsdam
108 PUBLICATIONS 3,463 CITATIONS
SEE PROFILE

Computers & Education 184 (2022) 104503

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Class size affects preservice teachers' physiological and psychological stress reactions: An experiment in a virtual reality classroom

Yizhen Huang [a,1,*], Eric Richter [a], Thilo Kleickmann [b], Dirk Richter [a]

[a] Department of Education, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476, Potsdam, Germany
[b] Department of Research on Teaching and Teacher Education, Kiel University, Olshausenstraße 75, 24098, Kiel, Germany

ARTICLE INFO

Keywords:
Augmented and virtual reality
Simulations
Improving classroom teaching
Media in education
Pedagogical issues

ABSTRACT

Teachers frequently express stress associated with teaching in large classrooms. Despite the time-honored tradition in teacher stress research of treating class size as a job-related stressor, the underlying premise that class size directly impacts teachers' stress reactions remains untested. In this randomized controlled experiment targeted at preservice teachers, we utilized a standardized virtual reality (VR) classroom to examine whether class size (number of student avatars) directly affected physiological (heart rate) or psychological (subjective rating) stress reactions among 65 preservice teachers. Results from linear mixed-effects modeling (LMM) showed that class size significantly predicted both their physiological and psychological stress reactions in the simulated environment: Average heart rate and subjective stress ratings were both significantly higher in the large class size condition. Further investigations into the causes of this association has been proposed. These findings may contribute to a better understanding of the effects of classroom features on preservice teachers' emotional experiences and well-being.

Author credit

Yizhen Huang: Conceptualization, Formal analysis, Investigation, Data Curation, Writing-Original Draft. Eric Richter: Conceptualization, Investigation, Data Curation, Writing-Review & Editing. Thilo Kleickmann: Conceptualization, Methodology, Resources, Data Curation, Writing-Review & Editing. Dirk Richter: Conceptualization, Project administration, Resources, Writing-Review & Editing.

1. Introduction

Teaching is a highly stressful profession, even compared to ambulance and social service workers (Johnson et al., 2005). Each and every day, teachers juggle a great number of job demands—"physical, psychological, social, or organizational aspects of the job"

Abbreviations: virtual reality, VR; linear mixed-effects model, LMM; heart rate, HR; beats per minute, BPM.

* Corresponding author.
E-mail addresses: yihuang@uni-potsdam.de (Y. Huang), eric.richter@uni-potsdam.de (E. Richter), kleickmann@paedagogik.uni-kiel.de (T. Kleickmann), dirk.richter@uni-potsdam.de (D. Richter).
[1] Department of Research on Teaching and Teacher Education, Kiel University, Olshausenstraße 75, 24098, Kiel, Germany.

https://doi.org/10.1016/j.compedu.2022.104503
Received 23 July 2021; Received in revised form 14 March 2022; Accepted 17 March 2022
Available online 29 March 2022

(Bakker & Demerouti, 2007, p. 312) that require both their physical and psychological efforts. These job demands typically include classroom management, relationships with colleagues, time pressure in grading and course preparation, and our current focus—teaching a large class size (Aldrup et al., 2018; Kyriacou, 2001; Taris et al., 2017). Class size is commonly defined as the student-to-teacher ratio in a classroom (e.g., Blatchford & Russell, 2019b; Harfitt, 2012; Solheim & Opheim, 2019), and it equals the number of students in a classroom if only one teacher is present.

Facing concurrent and constant job demands, teachers often report high levels of job-related stress, or work stress: the experience of "unpleasant, negative emotions, such as anger, anxiety, tension, frustration or depression" (Kyriacou, 2001, p. 28) due to aspects of the job (e.g., Klassen & Chiu, 2010; Klusmann et al., 2016). Job demands that lead to stress reactions are also referred to as (job) stressors (Kyriacou & Sutcliffe, 1978), and individuals' acute or chronic reactions to stressors are called stress reactions or strains (Geurts & Sonnentag, 2006; Harmsen et al., 2019).

Catering to each individual student's unique needs is arguably one of the most crucial job demands of teachers (Klem & Connell, 2004). However, as the number of students (class size) grows, so do the quantity and simultaneity of classroom events (Doyle, 2006), making it difficult for teachers to attend to each student (Blatchford & Russell, 2019a) and manage the classroom (Harfitt, 2012). Teachers frequently comment about how challenging and stressful it is to teach in overcrowded classrooms, i.e., with large class size (Finn et al., 2003; French, 1993). This issue is particularly relevant for preservice teachers, as they are more susceptible to job stress due to a lack of teaching experience (Dicke et al., 2014; Harmsen et al., 2019).

Class size is therefore often regarded, rather intuitively, as a stressor for teachers (Saloviita & Pakarinen, 2021). Despite the time-honored practice of including class size as a class-level factor in teacher stress studies (e.g., Geving, 2007; Lambert et al., 2009), the underlying assumption that class size directly affects teachers' stress reactions remains untested. The difficulties of randomly assigning teachers to classes of varying sizes while controlling for context- and process-variables contributes to the lack of investigation into the causal relationship between class size and stress reactions. Therefore, a classroom environment that enables experimental control but also allows teachers to behave realistically would be ideal for investigating the effect of class size. Furthermore, the field heavily relies on retrospective self-reports for measuring stress reactions (Donker et al., 2018), which disregards the equally significant physiological stress reactions. As proposed by Francis et al. (2017), including both "subjective self-report of internal states and emotions" and "objective measurement of physiology" (p. 447) will greatly improve the research methodology in this area.

The current study aimed to empirically investigate whether class size directly affects preservice teachers' stress reactions. We addressed the existing research gaps by randomly assigning preservice teachers to different class sizes while measuring both their physiological and psychological stress reactions. Our classroom environment was constructed in virtual reality (VR) to ensure both realism and controllability (see 2.2 for details). Specifically, we were interested in the influence of class size on preservice teachers' physiological stress reactions (average heart rate) and their psychological stress reactions, which were their self-reported stress levels after teaching in the VR classroom.

1.1. Teacher stress: outcomes and causes

Teacher stress as a negative emotion is accompanied by other response correlates, called stress reactions or strains (Griffin & Clarke, 2011; Kyriacou & Sutcliffe, 1978). Stress reactions can be classified as physiological (e.g., increased adrenaline/cortisol levels and heart rate), psychological (e.g., fatigue and job dissatisfaction), and behavioral (e.g., absenteeism and friction with colleagues) responses to the job demands (Sonnentag & Fritz, 2015; van Dick & Wagner, 2001). Recurrent and extended stress reactions can result in negative physical, psychological, and vocational consequences for teachers, such as burnout (Guglielmi & Tatrow, 1998), low self-efficacy (Skaalvik & Skaalvik, 2017), job dissatisfaction (Ouellette et al., 2017), absenteeism (Troman & Woods, 2001), and even attrition (Harmsen et al., 2018). Aside from the negative impact on teachers' own psycho-physical well-being, teacher stress may also damage students' learning motivation (Ouellette et al., 2017) and academic achievement (Klusmann et al., 2016).

Novice teachers, either in training (student/trainee/preservice teachers) or already beginning the job (beginning teachers), have been found to be particularly vulnerable to job stress due to a lack of experience in areas such as classroom management, acquiring support, and handling heavy workload (e.g., Chaplain, 2008; Fimian & Blanton, 1987; Murray-Harvey et al., 2000). According to a recent survey in the United States, one quarter of beginning teachers are at risk of stress during their first year (Fitchett et al., 2018). Harmsen and others (2018) reported that the level of emotional stress caused by stressors such as student misbehavior and poor relationships with students was an important factor causing attrition among the 545 beginning teachers surveyed. The most powerful predictor of retention among trainee teachers, on the other hand, was how much enjoyment—the opposite of stress—they expected to get from their work (Wilhelm et al., 2000). To support positive job experiences of novice teachers, it is therefore paramount to investigate not only the outcomes, but also the causes of their stress (Brok et al., 2017).

Stress causes, or stressors, are conceived as antecedents of teacher stress according to the classic model of teacher stress conceptualized by Lazarus and colleagues (1966) and later adopted by Kyriacou and Sutcliffe (1978). According to this model, stressors are described as primarily psychological, physical, or a combination of the two (Kyriacou & Sutcliffe, 1978, p. 4). Stressors that have been identified repeatedly in the literature include psychological stressors such as conflicts with colleagues and school leaders (Troman, 2000), student misbehavior (Punch & Tuettemann, 1990), and a lack of autonomy (Chen & Miller, 1997; Schmidt et al., 2017); as well as a combination of psychological and physiological stressors such as heavy workload (Borg & Falzon, 1990; Timperley & Robinson, 2000).

Class size has often been mentioned alongside stressors like workload (e.g., Klassen & Chiu, 2010) and classroom management (Finn, 2019) based on the logic that more pupils in a classroom means greater workload and more opportunities for misbehavior. Yet class size has rarely been examined as the focal point in the context of teacher stress. Given that teaching in large classrooms may be a prominent source of everyday stress (also called daily hassles)—"irritating, frustrating, distressing" emotions that "characterize everyday transactions with the environment" (Kanner et al., 1981, p. 3; Klusmann et al., 2021)—and that its associated classroom management issues have been considered as a major source of concerns for preservice teachers (Evertson & Weinstein, 2006), in this study we regard class size alone as a critical stressor for preservice teachers.

1.2. Class size as a stressor

Researchers, teachers' unions, and policymakers have long debated over the impact of class size on students' academic outcomes (Blatchford & Russell, 2019b). For instance, Achilles and Finn (2000) argued that small class size should be the "cornerstone of education policy" (p. 1) based on the findings of Tennessee's Project STAR that smaller classes lead to higher achievement in the elementary grades (Finn & Achilles, 1999). Others, however, have contended that lowering class size is not an efficacious way to improve academic performance due to the high human and financial costs involved (e.g., Hanushek, 1999). The long-running debate about class size has been primarily focused on the inconclusive results regarding its associations with student academic achievement (Hattie, 2005). However, the effect of class size on actual classroom processes is often overlooked, especially from the standpoint of teachers (Blatchford et al., 2012).

The association between class size and teachers' emotional experiences, either stressful or pleasant, has often been indicated in teachers' retrospective self-reports. For example, in a New York-based initiative to reduce class size to an average of under 20, ninety-six percent of 213 participating teachers said the reduced class was more enjoyable and less stressful to teach (Finn et al., 2001, 2003). Class size is also an important decision factor when choosing workplaces: 73% of Arizona teachers said they would choose schools based on class size (Mulholland, 1999). Aside from the general preference for smaller classes, teachers also reported high levels of stress in large classes. In an early survey of 223 Colorado elementary teachers, French (1993) found that teachers who taught in schools with large student-teacher ratios reported that class size was a prominent source of stress for them. Eighty-three percent of the responding teachers in a more recent survey conducted in the United Kingdom reported that large class sizes adversely affected their stress levels (Association of Teachers & Lecturers, 2009). Likewise, 97 Israeli teachers reported more stress related to larger class sizes in a cross-cultural study (Pines, 2002). These findings based on teachers' retrospective self-reports revealed a widespread view among in-service teachers of disfavoring large class size and treating it as a source of stress. Nevertheless, the causal relationships between class size and preservice teachers' stress reactions remain unestablished due to a lack of experimental investigations targeting this particular population.

The gap in the field could be attributed to the fact that previous qualitative investigations on class size and teachers' stress reactions were not obtained directly from the act of teaching, but from teachers' subjective and retrospective impressions (Harfitt, 2012). This approach is flawed because the negative impression of teaching in a large classroom is not equivalent to the actual stress reaction experienced while teaching in such a class. For instance, in a report regarding class size reduction reform in Hong Kong, Galton and Pell (2009) reported that even when teachers had not taught in a small class before, they claimed that they felt more "comfortable, enthusiastic and relaxed" (p. 12) in a small class. In order to parse out this type of hypothetical intuition from the actual stress reaction, it is necessary to conduct randomized controlled studies in which stress reactions are measured soon after or even simultaneously with instruction in a classroom that varies only in class size but not in other aspects.

To summarize, class size effect has been extensively researched in relation to students' academic achievement, but has rarely been explored in the sense of preservice teachers' stress reactions, especially using a randomized controlled design. While large class sizes seem to be associated with high levels of stress based on in-service teachers' retrospective self-reports, preservice teachers' in-the-moment stress reactions in response to various class sizes have never been examined until now. In the next sections, we introduce the importance of dual measurements of stress reactions, as well as the affordances of virtual reality (VR) environments for investigating class size effects.

1.3. Physiological and psychological stress reactions

Historically, physiological stress reactions have been largely overlooked in contrast to self-reported psychological stress reactions (Ganster & Rosen, 2013). This is despite the fact that the majority of job stress theories recognize the critical role of physiological processes in the workings of stress (e.g., Demerouti et al., 2001; Meurs & Perrewé, 2011). Heart rate (HR) is one of the most widely adopted indicators of physiological stress reactions (Behnke & Carlile, 1971; Bodie, 2010). Increased HR is routinely reported in highly stressful job scenarios (Galy et al., 2012). Mental effort related to cognitive processing, called cognitive load (Plass et al., 2010; Salomon, 1984), also contributes to racing heart beats (Taelman et al., 2011).

Despite the extensive use of HR as an indicator of physiological stress reactions in other professions, dual measurement of both psychological and physiological stress reactions among teachers is relatively uncommon (Francis et al., 2017). In a series of studies conducted in a natural setting, Serrano et al. (2008) observed that Spanish teachers' HR and perceived stress were stronger while they were with students, that all teachers' HR levels were higher at the end of the working day relative to the beginning and middle of the day (Serrano et al., 2008), and that teachers' HR reactions to a workday were greater at the end of a school year (Serrano et al., 2013). With a longitudinal but similar study design, Ritvanen et al. (2003) found that Finnish teachers' HRs were higher on workdays versus summer holidays. Junker et al. (2021) recently measured teachers' heart rates during real-life lessons and reported that low student engagement and motivation, as well as teacher-centered activities, significantly predicted an elevated heart rate. These studies shed light on the application of HR measurement for investigating teachers' psychophysiological stress reactions, but their descriptive or cross-sectional nature precludes them from providing cause-and-effect analyses of the relationships between stressors and stress reactions.

Similar to the criticism of Ganster and colleagues (2013) regarding the work stress literature, we would argue that teacher stress research has also been hampered by a strong dependence on subjective self-reports, as well as a dearth of randomized and controlled experiments that could establish cause-and-effect relationships between the stressor (e.g., class size) and stress reactions (both physiological and psychological). The scarcity of randomized and controlled experiments in educational science is often justified on two grounds: a) classroom characteristics such as class size are difficult to manipulate in real life (Borman, 2009), and b) the standardized laboratory setting is often limited by its poor ecological validity (Cook, 2002). By combining the complexities of authentic teaching and the meticulous control necessary for experiments, the virtual reality (VR) environment has arisen as a promising testbed for examining the causal connections between class size and stress reactions.

1.4. Virtual reality as an experimental environment

Virtual reality (VR) is a collection of technologies that create “synthetic, highly interactive three dimensional (3D) spatial environments” that may represent real or non-real situations (Mikropoulos & Natsis, 2011). VR has been widely adopted in professional trainings that demand situated practice in authentically complicated situations (for a review, see Jensen & Konradsen, 2018), such as pilots (Lele, 2013) and surgeons (Alaraj et al., 2011). According to a recent review of VR applications in teacher education (Huang et al., 2022), VR technology has also been employed for cultivating teachers’ classroom management (e.g., Ye et al., 2019) and communication skills (e.g., Spencer et al., 2019), among other competencies.

VR environments have also been considered ideal for human subject experiments in areas such as consumer behavior (Kim et al., 2020), human perception (Roberts et al., 2019), communication (Mostajeran et al., 2020; Wallach et al., 2009), and social behavior research (Kyrlitsias & Michael-Grigoriou, 2018). The realism of the VR environment allows participants to act in ways that are comparable to behaviors in real life, therefore leading to a high ecological validity and transferability of study findings (Roberts et al., 2019). Additionally, as VR environments are synthetically generated, they allow for experimental manipulations that are difficult to accomplish in real life (Tarr & Warren, 2002). For example, the occurrences and characteristics of various scenarios that are subject to chance in real classrooms can be controlled in VR classrooms (Theelen et al., 2019). In such a VR classroom, features like room layout, class size, student features, and behaviors can also be manipulated (e.g., Huang et al., 2021; Wiepke et al., 2021; Seufert et al., 2022).

Teachers’ behaviors in VR classrooms have also been investigated recently. For instance, Krämer and Zimmermann (2019) reported that student avatars’ scripted disruptions in the VR classroom affected teachers’ assessments of their academic performance. In another VR-based experiment, novice teachers were assigned to teach in VR classrooms with different complexity levels, represented by the number of disruptions, the existence of overlapping disruptions, and parallel teaching tasks (Huang et al., 2021; Wiepke et al., 2021). Teachers in the high complexity group were less likely to notice disruptions and also less likely to take their time reacting to these disruptions. Similar to these experimental studies, a lifelike and configurable VR classroom would enable researchers to investigate the class size effect on preservice teachers’ stress reactions with a random assignment design that has not been possible in the past.

1.5. The current study

As previously discussed in 1.2, while large class size has often been identified by teachers as undesirable and stressful, its causal effect on teachers’ stress reactions has never been empirically examined. The research methodology in this field could also benefit from including both physiological and psychological stress reactions to obtain a more comprehensive picture of teacher stress.

In the current study, we were interested in the potential causal relationships between class size and preservice teachers’ stress reactions among the population who’s vulnerable to work stress: Will preservice teachers demonstrate higher levels of stress reactions, both physiologically and psychologically, when teaching in a VR classroom with more students? For this research goal, we created a standardized VR classroom with precise control over the variables of interest (see 2.2 Material and Equipment for details). We randomly assigned participants to either small or large class size conditions and measured the physiological stress reaction, which was the average HR during a teaching task after controlling for each participant’s baseline HR, as well as their subjective ratings of stress level following the teaching task after controlling for their long-term stress ratings.

Additionally, we included two covariates that might have influenced the stress reactions in this study: perceived cognitive load and negative VR experience. First, cognitive load is an index of the amount of mental effort necessary to accomplish a specific cognitive task (Plass et al., 2010; Salomon, 1984). As reviewed earlier in 1.3, HR reflects changes not only in affective experience, such as stress, but also in the state of mental effort (Taelman et al., 2011). We use only perceived cognitive load as a covariate in the current analysis, since we are primarily interested in the emotional aspect of teachers’ reactions to class size rather than the cognitive aspect. Second, previous studies have also reported that participants may feel anxious and stressed in an unfamiliar VR environment (Makransky & Lilleholt, 2018; Stavroulia et al., 2019). Therefore, we treat the self-reported negative experience in the VR environment as a covariate in order to control for such potential influence.

According to previous findings that teachers often report class size as a source of stress (e.g., Finn, 2019; Klassen & Chiu, 2010) and that teachers’ HR is sensitive to stressful situations (e.g., Moya-Albiol et al., 2010; Ritvanen et al., 2006), we formulate the hypothesis as follows: Preservice teachers teaching in a VR classroom with a large class size will::

a) display stronger physiological stress reactions (higher average HR after controlling for baseline) during the teaching task;
b) report stronger psychological stress reactions (higher self-reported acute stress after controlling for long-term stress) after the teaching task.

2. Method

2.1. Participants and design

A total of 65 preservice teachers were recruited from a seminar held at a public German university. The participants' average age was 24.9 years ($SD = 4.82$), with 48% ($n = 31$) male and 52% ($n = 34$) female. All participants were native German speakers and were enrolled in the teacher preparation program. They were pursuing either a bachelor's (96.9%) or master's degree (3.1%) in various subject fields (see Table A1 in Appendix A for detailed sociodemographic characteristics). These preservice teachers were attending a weekly seminar on classroom management as part of their theoretical module with no in-field practicum component. Although the VR experiment was an integral part of the course, preservice teachers had the option of participating with or without data being collected. During the VR experiment, participants taught and oversaw student avatars in a simulated classroom that resembled the upper secondary classrooms in Germany (Wiepke et al., 2021). The class size was manipulated based on the number of student avatars in the room; all other factors, as well as the behavior of the student avatars, were identical for all participants (see 2.2 Material and Equipment: VR Classroom for more details).

Each participant was invited to attend the VR experiment twice in two weeks to ensure the reliability of measurements and to minimize the influence of missing records. Each time, they were randomly assigned to either large or small class size. The VR classroom with a large class size contained 30 student avatars, while the small class size had 10 (see Table A2 in Appendix A for the room map). The reasoning behind this setup was to go slightly above and below the current average class size (23) in German secondary schools (OECD, 2020), ensuring generalizability to real-world classrooms while also presenting a visible difference between two class sizes. This distinction is similar to prior observation studies, which on average had 33 vs. 19 (Blatchford, 2003) and 40 vs. 15 (Finn, 2019) students in different class sizes.

The condition assignment was independent of the measurement time; therefore, the same participant might or might not have been assigned to the same condition (see Fig. A1 in Appendix A for the participant flow diagram). In total, 55 observations from the small class size and 63 observations from the large class size were collected. The two class size groups did not differ significantly with regard to age, gender, semester, or prior VR experiences in each of the measurement time (see Table A1 in Appendix A for details).

2.2. Material and equipment: VR classroom

The VR classroom used in this experiment was developed by the computer science department at the university to simulate a typical upper secondary classroom in Germany.[2] The VR classroom contained five rows and three columns of desks and chairs (see Fig. 1 top). Depending on the condition, there were either 30 or 10 student avatars in this VR classroom. Student avatars had a wide range of physical characteristics, such as skin tones, hairstyles, and clothes. Students' names were visible on name tags at their desks (see Fig. 1 top). Other classroom features (clock, wall paper, blackboard, etc.) as well as directional ambient noises were intended to heighten the sense of presence in this environment and were held constant across conditions. Participants in previous studies who used this VR classroom reported it to be realistic and authentic (Wiepke et al., 2019, 2021).

Participants were instructed to behave and interact with student avatars as though they were in a real classroom, such as walking around the entire room, looking at student avatars while talking, etc. Student avatars would perform both on- and off-task behaviors, such as writing in a notebook or throwing a paper ball (see Fig. 1 bottom). Otherwise, avatars would sit naturally with different neutral postures and follow the participants around by moving their eyes or bodies. The off-task behaviors were chosen from a list of common misbehaviors that may disrupt a class (Borko, 2016; Wolff et al., 2016). The onset time, duration, location, and type of student behaviors were scripted and were consistent across conditions (see Table A3 in Appendix A for the script and all the possible behaviors).

The VR classroom was presented with the HTC VIVE head-mounted device (VR headset), which features a high-resolution display (2880 × 1600 pixels), motion sensors, and an immersive soundscape that ensure a highly-interactive and authentic experience (Wiepke et al., 2021). In essence, the participant could move around in reality while simultaneously experiencing the corresponding multisensory feedback within the VR classroom. During the experiment, participating teachers would teach as if they were in a real-life classroom.(see 2.4 Procedure: VR Experiment for details).

2.3. Measures and instrumentation

Participants' physiological stress reactions were operationalized as the average HR during the teaching task after controlling for the baseline HR for each participant. HR is the speed of heart contractions measured by the beats per minute (BPM) (Hugdahl, 1981). It is a form of cardiovascular response that represents the autonomic nervous system's functioning during times of stress (Andreassi, 2006). We continuously measured participants' HR at 0.3s intervals during the experiment using Polar OH1—an armband optical HR sensor that has been validated with electrocardiography (Hettiarachchi et al., 2019). To control for individual differences in HR caused by factors unrelated to the experimental manipulation (Schubert et al., 2009), we measured each participant's baseline HR (average HR during the last 3 min of a habituation phase) and then centered the average HR during the teaching task on the baseline HR to obtain the physiological stress reaction (see Table A4 in Appendix A for descriptive statistics). For instance, the physiological stress reaction

[2] Shared under GNU Affero General Public License. Objects and source codes are available at https://gitup.uni-potsdam.de/mm_vr/vr-klassenzimmer.

Fig. 1. View in the VR Classroom. *Note.* All views were displayed from the participant's visual perspective. Top: front view from the teacher's desk; bottom left: student avatar performs off-task behavior; bottom right: on-task behavior.

was 0 if the HR during the task was the same as the baseline HR.

Participants' psychological stress reactions to the teaching task were operationalized as their self-reported acute stress immediately after the VR experiment compared to their self-reported long-term stress. We adopted measurements of stress levels that have been frequently used in studies investigating stress reactions to work conditions (Cohen et al., 1983; Rissén et al., 2000). Acute stress was assessed with a single-item measure ("How are you feeling with regard to your emotional state at this moment?") on a 10-point Likert scale ($1 =$ *calm, relaxed, and composed* to $10 =$ *stressed*) (Delaney & Brodie, 2000). Another single-item measure ("In the last month, how often have you felt nervous and stressed?") was used to measure long-term stress on a 10-point Likert scale ($1 =$ *never* to $10 =$ *very often*) (Cohen et al., 1983). As the psychological stress reactions were defined as the acute stress compared to long-term stress, the range of this variable ended up being 0–18, with 0 indicating that the acute stress was the same as the participants' long-term stress.

A negative experience in the VR environment, as well as a high cognitive load while performing a task, have been found to be associated with a higher level of stress reactions (Stavroulia et al., 2019; Taelman et al., 2011). Therefore, we included the perceived negative VR experience and cognitive load as covariates in the current study. We measured participants' negative experience in the VR classroom with a five-item scale we developed ($\alpha = 0.83$; $\omega = 0.84$). Each item was a statement about negative experiences in the VR, such as "I found the time in the virtual classroom exhausting" (see Table A5 in Appendix A for all items). Participants reported their agreement with the statements on a 4-point Likert scale ($1 =$ *completely disagree* to $4 =$ *completely agree*). To measure the perceived cognitive load involved in teaching in the VR classroom, the widely-used mental effort rating scale developed by Paas (1992, p. 430) was used. Participants rated their "invested mental effort during the task" on a 9-point Likert scale (1 = *very low mental effort* to 9 = *very high mental effort*).

2.4. Procedure: VR experiment

The participants were randomly assigned to either large or small class size conditions, with all other aspects of the procedure held constant across conditions and measurement times (see Supplemental Materials for the experiment protocol). As participants arrived at the VR lab, they were instructed to sit quietly, avoid moving, and keep their hands resting on the chair and feet flat on the floor for 8 min. This standardized habituation phase was intended to stabilize participants' HR and ensure an accurate measurement of baseline HR (Goodie et al., 2000). After the habituation, participants put on the VR headset and entered the VR classroom with either 30 or 10 student avatars. First, they followed a 5-min set of instructions in the VR classroom learning how to interact with the environment. Then, the participants were instructed to give a short lecture on a given topic (2020 United States presidential election) which they had prepared before the VR session. Participants were asked to introduce the topic and get students ready for group work for 5 min. The student avatars would occasionally carry out off-task behaviors that required the participants to intervene. These behaviors were pre-programmed and remained constant regardless of the condition. Finally, after the VR experiment, participants completed an online questionnaire that included questions about demographics, self-ratings of negative VR experience, cognitive load, and short-term and long-term stress (see 2.3 Measures and Instrumentation for details). The duration of the entire experiment was approximately 40 min.

2.5. Statistical analyses

We used linear mixed-effects modeling (LMM, see Appendix B for details) to evaluate the effect of class size on stress reactions while accounting for covariate effects that could be generalized to other samples of participants and measurement times (Baayen et al., 2008). Our study design treated class size (condition) as a fixed factor, with participant and measurement time as random factors. Specifically, participants were nested in the condition in each measurement time, and the measurement time was crossed with the condition. Participants and measurement time in this design were "random" in the sense that they were assumed to be a random sample of the participants and measurement times that might have been included in the study (Judd et al., 2016). By modeling each participant and measurement time as a level of a random factor along with the fixed factor (condition), we can obtain an estimate of the mean condition difference, as well as an estimate of the variances surrounding that difference, due to the individual features of each participant or measurement time (Westfall et al., 2014).

We constructed and tested two sets of linear mixed-effects random intercept models, with and without the covariates, for each stress reaction. In the baseline models (Model 1), physiological and psychological stress reactions were modeled as functions of fixed class size effects alone, and with participant and measurement time as random intercepts. Base models served as reference models and were then compared with the full models (Model 2), which included two covariates: negative VR experience and perceived cognitive load. To clarify whether the introduction of further random components into the model would significantly improve model fit, we conducted the likelihood ratio (LR) test to compare baseline and full models (Bolker et al., 2009).

We fit the models with restricted maximum likelihood (REML) using the lmer function from the lme4 package (Bates et al., 2015) in R 4.0.4 (R Core Team, 2021). The scale-standardized regression coefficients (standardized $\widehat{\beta}$) were calculated using outputs from lme4, and effect sizes were calculated based on Brysbaert and Stevens's (2018) guidelines on conducting power analysis in mixed-effects models (see Appendix B for details). In terms of goodness of fit (see Appendix B for details), we included indices commonly used in the field (e.g., Goldhammer et al., 2014; Kliegl et al., 2011): Akaike Information Criterion (AIC; decreases with goodness of fit), the Bayesian Information Criterion (BIC; decreases with goodness of fit), and the χ^2-distributed likelihood ratio and its p-value for model comparisons.

We also conducted missing value analyses to identify potential patterns in missing data that might bias the analysis by performing Little's (1988) missing completely at random (MCAR) test with IBM SPSS Statistics 26. In the case of a non-significant Little's MCAR test, observed data is considered to be a random sample from all data, allowing for a complete-case analysis with no bias imposed

Table 1
Descriptive statistics and intercorrelations for included variables.

Variables	*n*	Missing (%)	*M* (*SD*)	Range/Category	1	2	3
1. Physiological Stress Reaction [a]	89	29 (25%)	58.50 (23.10)	7.86–108.00	–		
2. Psychological Stress Reaction [b]	117	1 (1%)	6.19 (2.66)	0–18	0.27**	–	
3. Perceived Cognitive Load	117	1 (1%)	4.25 (1.91)	1–9	0.05	0.47***	–
4. Negative VR Experience	117	1 (1%)	2.27 (0.71)	1–4	0.07	0.50***	0.74***

Note. $N = 118$ observations from 65 participants. *$p < .05$. **$p < .01$, ***$p < .001$.

[a] Physiological stress reaction was the HR in BPM during the experiment compared to baseline HR. 0 indicates that the HR during the experiment was the same as the baseline HR.

[b] Psychological stress reaction was the acute stress compared to long-term stress. 0 indicates that the acute stress was not different from the long-term stress.

(Ibrahim & Molenberghs, 2009). Due to one participant's early withdrawal from the experiment, one observation in the small class size condition was omitted. We obtained incomplete HR recordings in 29 observations due to technological failure (24.58%) and these were excluded from the current analysis. Little's MCAR test yielded non-significant results for all variables included in our models, therefore the data were considered to be missing completely at random ($\chi^2 = 37.75$, $df = 29$, $p = .13$). Additionally, the missing data is not likely to bias the estimation due to the robustness of LMM against missing data (Baayen et al., 2008).

3. Results

3.1. Descriptive statistics

As shown in the summary of descriptive statistics (Table 1), participants' mean HR during the teaching task was 58.5 BPM higher than the baseline HR. Their average acute psychological stress reactions immediately after the teaching task were higher than the baseline long-term stress by 6.19 on a 0–18 scale. Physiological and psychological stress reactions were significantly and positively correlated ($r = 0.27$, $p < .001$). The included covariates, perceived cognitive load ($r = 0.47$, $p < .001$), and negative VR experience ($r = 0.50$, $p < .001$) were both significantly and positively correlated with psychological stress reactions, but not with the physiological ones.

3.2. Class size effect on physiological stress reactions

The first question we intended to investigate was whether a large class size causes a stronger physiological stress reaction, namely a higher average HR during the teaching task, after controlling for baseline HR. To account for effects from covariates, we first compared the goodness of fit between the baseline model (Model 1) and the full model (Model 2). The change in log-likelihood was not significant ($\Delta\chi^2$ (2) = 4.22, $p = .12$), indicating that the goodness of fit did not improve with the inclusion of two covariates. This infers that the covariates were not influential in affecting participants' physiological stress reactions.

As shown in the summary of Model 2 in Table 2 (see Table C1 in Appendix C for the full model summary), class size significantly predicted the average HR (compared to each participant's baseline HR) during the teaching task ($\widehat{\beta} = 17.51$, $t\,(80) = 13.05$, $p < .001$) with a large effect size ($d = 1.17$). When participants were teaching in a large VR classroom, they had significantly stronger physiological stress reactions ($M = 75.4$ BPM, $SD = 13.2$ BPM) compared to when teaching in a small VR classroom ($M = 40.3$ BPM, $SD = 16.9$ BPM). In other words, the average HR of teaching in a large class was 1.87 times faster than in a small class after taking into account the baseline HR. In terms of the covariates, the summary of Model 2 also demonstrates that participants' negative VR experiences were not significantly associated with the physiological stress reaction, while the effect of perceived cognitive load was significant ($\widehat{\beta} = 2.39$, $t\,(80) = 1.96$, $p = .049$) with a small effect size ($d = 0.16$). In sum, these results supported our first hypothesis that preservice teachers who taught in a VR classroom with a large class size would have a higher average HR during the teaching task.

3.3. Class size effect on psychological stress reactions

For our second hypothesis, we wanted to examine whether a large class size also caused higher psychological stress reactions, i.e., higher self-reported acute stress ratings (compared to each participant's long-term stress). The comparison between Model 1 and Model 2 showed that the change in log-likelihood was significant ($\Delta\chi^2$ (2) = 42.99, $p < .001$). The goodness of fit increased from Model 1 to Model 2, as both AIC and BIC statistics decreased (AIC: 551.55, 512.56; BIC: 570.88, 537.42). This implies that the covariates also influenced psychological stress reactions of the participants.

As shown in the summary of Model 2 (Table 3; see Table C2 in Appendix C for the full model summary), we found that class size significantly predicted the self-reported acute stress after controlling for long-term stress ($\widehat{\beta} = 0.91$, $t\,(108) = 5.06$, $p < .001$) with a medium effect size ($d = 0.52$). Participants who taught in the large class size condition reported significantly higher psychological stress reactions ($M = 6.93$, $SD = 2.63$) than participants teaching in small classes ($M = 5.33$, $SD = 2.44$). These results supported our second hypothesis that teaching in a VR classroom with large class size would lead to higher self-reported stress immediately after the teaching task.

Table 2
Summary of the model regarding physiological stress reaction.

Terms	Model 1 [a]				Model 2 [b]			
	$\widehat{\beta}$	$SE\ (\widehat{\beta})$	Standardized $\widehat{\beta}$	t	$\widehat{\beta}$	$SE\ (\widehat{\beta})$	Standardized $\widehat{\beta}$	t
Class size (large-small)	17.36	1.36	0.75	12.75***	17.51	1.34	0.76	13.05***
Perceived cognitive load					2.39	1.21	0.20	1.96*
Negative VR experience					−3.84	3.24	−0.12	−1.19

Note. $n = 89$ observations. *$p < .05$. **$p < .01$, ***$p < .001$.
[a] AIC = 736.59; BIC = 754.01.
[b] AIC = 736.37; BIC = 758.77.

Table 3
Summary of the model regarding psychological stress reaction.

Terms	Model 1 [a]				Model 2 [b]			
	$\widehat{\beta}$	$SE(\widehat{\beta})$	Standardized $\widehat{\beta}$	t	$\widehat{\beta}$	$SE(\widehat{\beta})$	Standardized $\widehat{\beta}$	t
Class size (large-small)	0.86	0.21	0.33	4.03***	0.91	0.18	0.34	5.06***
Perceived cognitive load					0.38	0.14	0.27	2.61**
Negative VR experience					1.29	0.40	0.35	3.22***

Note. $n = 117$. $^*p < .05$. $^{**}p < .01$, $^{***}p < .001$.
[a] AIC = 551.55; BIC = 570.88
[b] AIC = 512.56; BIC = 537.42

The covariates' effects were also shown in Model 2 (Table 3); participants' ratings of cognitive load and negative VR experience were both significantly associated with their psychological stress reaction. The psychological stress reaction was found to be positively associated with the ratings of cognitive load ($\widehat{\beta} = 0.38$, $t(108) = 2.61$, $p = .009$, $d = 0.17$) and negative with VR experience ($\widehat{\beta} = 1.29$, $t(108) = 3.22$, $p = .001$, $d = 0.59$). As presented in the plot of the estimated marginal means of psychological stress reactions in Model 2 (Fig. 2), the slopes of the fitted lines between class size and psychological stress reaction were similar across different levels of cognitive load ratings and negative VR experience, but differed in their intercepts. This finding shows that a higher level of cognitive load and a negative VR experience elicited greater psychological stress reactions.

4. Discussion

In the present study, we addressed the much overlooked question regarding the effect of class size on teacher stress. We measured HR as a physiological indicator of cardiac reactivity to the stressor and collected self-reports of stress after randomly assigning

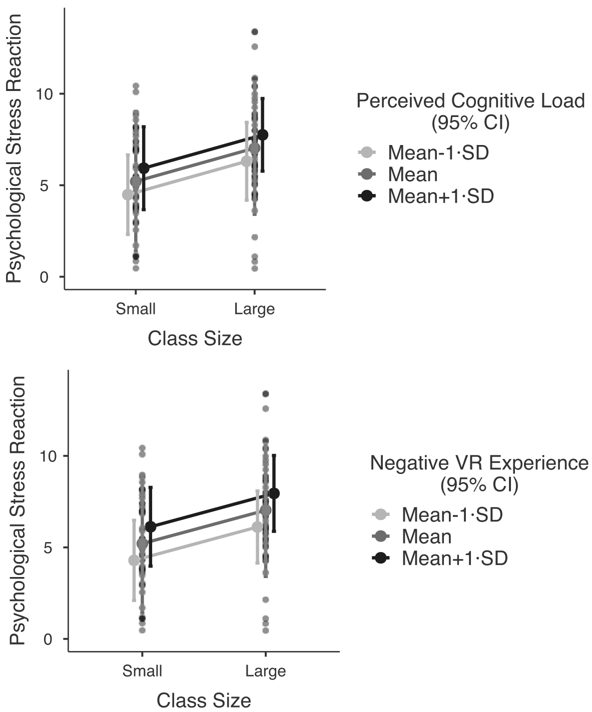

Fig. 2. Estimated Marginal Means of Psychological Stress Reactions Predicted by Class Size, Cognitive Load, and Negative VR Experience
Note. 95% confidence intervals of marginal means are presented for perceived cognitive load and negative VR experience. Individual observations are also plotted.

preservice teachers to large or small class sizes in a standardized VR classroom. We found that class size significantly predicted both physiological and psychological stress reactions: average HR and subjective stress ratings were both significantly higher in the large class size condition.

Additionally, we found that although both the physiological and psychological stress reactions were strongly affected by the class size, they were not completely identical in relation to cognitive load and negative VR experience—the two covariates were only significantly associated with psychological stress reactions. The associations between large class sizes and teachers' negative emotional experiences have often been treated as factual based on the abundant self-reports (French, 1993). In a review of class size effects, Finn (2019) even stated that it is near impossible to "find a teacher who would say they enjoy teaching a large class more than teaching a small one" (p. 130). One explanation has been that class size may influence student behaviors, which in turn affects teachers' experiences (e.g., Finn et al., 2003). The basis of this claim is that students' discipline problems will interfere with instructions. Teacher stress is reduced in small classes because student disruptions are less probable, and students are more likely to be engaged in learning in such an environment (Blatchford et al., 2007).

The present findings do not entirely support this view, since we completely standardized student off-task behaviors in the VR classroom in both large and small class size conditions. Because student avatars misbehaved identically in both conditions, we could conclude that, at least for the brief teaching task employed in the current experiment, students' problematic behaviors were not the source of teacher stress in large versus small VR classrooms.

Nevertheless, our findings do not exclude the possibility that the sheer number of classroom events, even the appropriate ones (on-task behaviors), might be influencing preservice teachers' stress reactions. For instance, they still need to actively monitor student actions (raising arms, writing and listening) even when these actions were not misbehaviors. According to Kounin (1970) and later Doyle (1985, 2006), the classroom environment is intrinsically complex because many events are occurring at the same time (i.e., simultaneity). Past studies have already identified certain associations between the complexity of the classroom environment and teachers' perceptions. For example, preservice teachers have been found to focus on only one part of the classroom (Doyle, 1977), and to report high levels of distraction in fully-packed classrooms (Ahrentzen & Evans, 1984; Phillips & Downer, 2017). More broadly, preservice teachers' restricted ability to mentally tackle the complexity of the classroom environment has also been indicated in studies investigating teachers' professional vision (Sherin et al., 2010) and withitness (Kounin, 1970). Novice teachers and preservice teachers often struggle to notice critical classroom events when watching recordings of a lesson (e.g., Calandra et al., 2008; Stockero et al., 2017), observing real classroom situations (Gegenfurtner et al., 2019; Sabers et al., 1991), and during their own lessons (e.g., , Huang et al., 2021; Cortina et al., 2015; McIntyre & Foulsham, 2018). Since a larger class size inherently embeds more classroom events with increased complexity, it is reasonable to speculate that preservice teachers may find it more challenging to teach in such an environment, resulting in heightened psychophysiological stress reactions. Future experiments that fully control for the number of classroom events could be beneficial for discerning this compounded question.

Eventually, our findings confirmed previous observational evidence about the effect of class size on teacher stress: larger class sizes did, in fact, result in higher levels of physiological and psychological stress reactions. As we looked for explanations elsewhere, we discovered that communications studies provide potentially relevant ideas. In the studies of public speaking anxiety (PSA), researchers are often interested in the influence of audience characteristics. Multiple studies have found that speakers exhibit more cognitive, physiological, and behavioral anxiety in front of larger audiences (for a review, see Bodie, 2010). One possible explanation is that with a larger group size, people invade each other's personal space and become forced to interact face to face, which may lead to increased stress reactions (Kanaga & Flynn, 1981). This hypothesis might be tested in a VR classroom with alternative room layouts that modified the extent to which teachers had to face the students directly and the relative distance between them. Another influential explanation is social impact theory (Latané, 1981). It states that an individual may exhibit physiological, psychological, and behavioral changes because of others' presences. According to this theory, as the number of people increases, so does the impact felt by the individual. To further confirm the applicability of social impact theory in the field of teacher stress, additional experiments with teachers' ratings of felt social impact from students should be conducted in the future. In sum, it is yet unclear why preservice teachers experienced higher level of stress reactions in larger classrooms. Investigations exploring the processes behind the relationships between class size and teachers' stress reactions, whether in the direction of classroom complexity or social impact, would be valuable for furthering our understanding of teacher stress.

4.1. Limitations and future directions

The current study could be further improved in the following aspects. First, among the three categories of stress reactions (physiological, psychological, and behavioral), we are still lacking an examination of teachers' behaviors in large versus small class size conditions. As observed in previous studies, teachers in large classes spent more time on "nonacademic management" of the class (Finn et al., 2003, p. 341). This observation should be examined with a similar experimental design in the future to investigate the differences in classroom management behavior and overall instructional quality in classes of different sizes.

Second, although our participants reported high levels of immersion and a strong sense of presence and rated the experience as highly authentic overall (Wiepke et al., 2019), we also acknowledge the artificiality and specificity of this VR classroom and the resulting generalizability problems. First, the virtual students' communications were not fully interactive, which may threaten teachers' willing suspension of disbelief in this virtual environment. Furthermore, the experiment's very specific classroom arrangement (secondary classroom with a grid layout) did not guarantee direct generalization to other grade levels and classroom layouts. In order to compare teacher's behavior in and out of the VR classroom, replication of the current results should ideally be carried out with other teacher populations (e.g., in-service teachers), in a different grade level, and in a real-life classroom with federated actors as students.

Additionally, our experimental manipulation created a large distinction in the large versus small class size (30 vs. 10), which is hardly viable in real life due to practical restraints. Our findings are thus not generalizable to class size reductions of smaller changes, such as in the Tennessee Project STAR (large class of 22–28 vs. small class of 13–18) (Finn & Achilles, 1999). Therefore, replication studies with smaller differences in class size would be favorable to consolidate the current findings about class size effect.

In the present study, we are focusing on teachers' acute psychophysiological stress reactions after exposure to class size. However, chronic stress reactions, which are long-term physiological and biochemical changes accompanied by psychosomatic symptoms like coronary heart diseases (van Dick & Wagner, 2001), is also a crucial research topic that should be studied in relation to class size and other classroom features. Acute stress reactions, such as increased HR, may persist after the stressor ceases to exist, such as after leaving the school and returning home. If stressors occur for an extended period of time, stress reactions can become recurrent, resulting in physical and psychological health impairments (e.g., cardiovascular disease and burnout) (Geurts & Sonnentag, 2006). Therefore, a large class size is not necessarily negative, but the acute stress reactions resulting from teaching in a large class may turn into chronic reactions if teachers have no chance to adequately recover from the stressful situations (Bakker & Demerouti, 2007; Meijman & Mulder, 1998). Sonnentag and Fritz (2015) recently suggested that avoiding work-related activities and thoughts during non-work hours may be an effective recovery strategy.

Future investigations of class size as a stressor should also address the role of teacher characteristics. For instance, teachers' self-efficacy—confidence in their ability to influence students' learning (Brouwers et al., 2001)—has been identified as a protective factor against teacher burnout (Klassen et al., 2011). Other important teacher characteristics, such as teaching experience, extraversion, self-esteem, and emotional competency, which may function as moderators or mediators between class size and stress outcomes, have not yet been formally examined.

A longitudinal measurement of stress reactions, especially physiological ones, is also missing in the field. The use of a daily diary design has already shed light on the mechanism of stress recovery (e.g., Sonnentag & Fritz, 2015), but the reliance on subjective self-reports has prevented the researchers from investigating the long-term effects of physiological activation due to the stressors.

We would like to propose that VR-based professional development offerings could be valuable job resources for mitigating the negative effect of job-related stressors for preservice teachers. Preservice teachers often encounter "reality-shock" when first entering the profession (Huberman et al., 1993), partially due to the incongruence of the training and the actual work environment. For instance, preservice teachers have very few practical training opportunities during their education (Grossman, 2018), and even when they do, the training is likely to take place in a downsized classroom with fewer students than in real life. This idea has often been captured in a popular form of teacher training—microteaching (I'Anson et al., 2003). Microteaching consists of scaled-down and simplified teaching in front of federated students (peers or tutors) that are also reduced in group size (Horgan et al., 2018). Typically, the session is videotaped and afterwards utilized for reflection and feedback. A VR classroom can be used for traditional microteaching, but it is also an optimal training environment that can be configured for stepwise training. As Gray (2002) puts it, VR provides "controlled complexity"—a delicate combination between artificial controllability and naturalistic sophistication. In a VR classroom, for example, preservice teachers may begin in a less challenging but unrealistic scenario with fewer students, and as they increasingly become accustomed to enacting core teaching practices, the level of stressors (e.g., class size) can then be adjusted accordingly to eventually reach the level of complexity in real life. We believe that this kind of stepwise training, which is unique to VR classrooms, would be an instrumental tool in preparing preservice teachers for real-world job demands.

5. Conclusions

Overall, this study advances theoretical knowledge of how class size affects preservice teachers' stress reactions in a standardized VR classroom. Teaching in a larger VR classroom raised the levels of both physiological and psychological stress reactions among preservice teachers. The causes of this association require further investigations. Lastly, since prolonged stress reactions are detrimental to preservice teachers' well-being and career prospects, the current empirical finding has important implications for designing professional development offerings that could be valuable job resources for mitigating the negative impact of job-related stressors for preservice teachers.

Declaration of competing interest

None.

We declare that there are no known conflicts of interest associated with this research.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2022.104503.

Appendix A

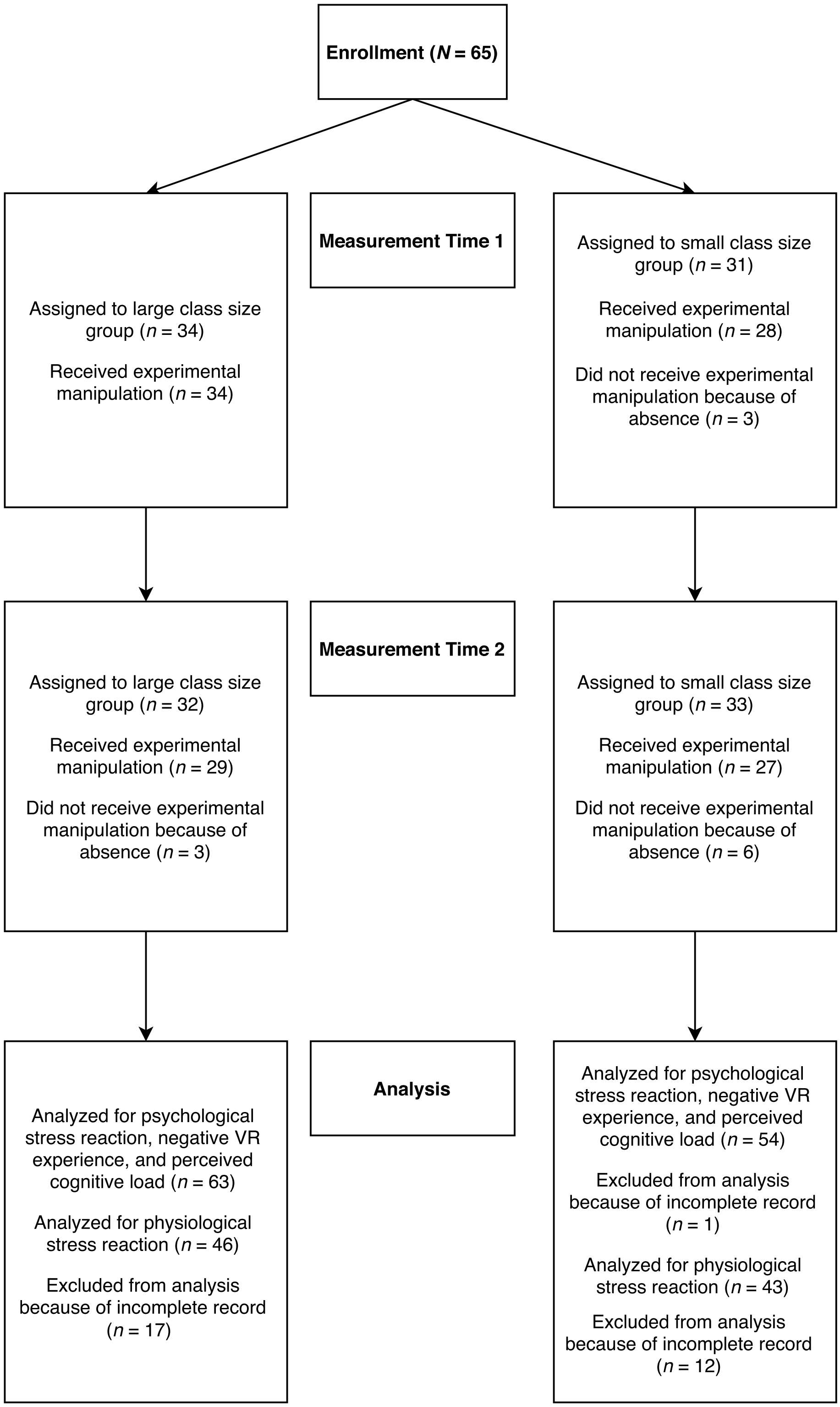

Fig. A.1. Flowchart of Participants Through Each Stage of the Experiment . *Note.* Participants were randomly assigned to conditions at each measurement time. Observations from both measurement times were used for analysis.

Table A.1
Sociodemographic Characteristics of Participants

Characteristic	Measurement time 1				Measurement time 2				Full sample
	Small class size		Large class size		Small class size		Large class size		
	n (%)	*M* (*SD*)	*n* (%)	*M* (*SD*)	*n* (%)	*M* (*SD*)	*n* (%)	*M* (*SD*)	*n* (%)/*M* (*SD*)
Gender									
Female	14 (45.2%)		20 (58.8%)		19 (57.6%)		15 (46.9%)		34 (52.3%)
Male	17 (54.8%)		14 (41.2%)		14 (42.4%)		17 (53.1%)		31 (48.2%)
Age		26.20 (4.52)		23.71 (5.19)		24.37 (3.68)		25.45 (5.40)	24.9 (4.82)
Semester		5.07 (1.91)		4.71 (1.19)		4.35 (1.09)		5.43 (1.67)	4.91 (1.56)
Degree program									
Bachelor	30 (96.8%)		33 (97.1%)		31 (93.9%)		32 (100%)		63 (96.9%)
Master	1 (3.2%)		1 (2.9%)		2 (6.1%)		0 (0.0%)		2 (3.1%)
Has teaching experience	22 (71.0%)		26 (76.5%)		27 (81.8%)		21 (65.6%)		48 (73.8%)
Has VR experience	8 (25.8%)		7 (20.6%)		11 (33.3%)		4 (12.5%)		15 (23.1%)

Note. $N = 65$ participants. In time 1, $n_{small} = 31$, $n_{large} = 34$; in time 2, $n_{small} = 33$, $n_{large} = 32$ (also see Fig. A1 in Appendix A). *$p < .05$. **$p < .01$, ***$p < .001$.

Table A.2
VR Classroom Room Map.

Plants and decorations							
	13L	**13R**	**14L**	**14R**	15L	**15R**	
	10L	10R	**11L**	11R	12L	**12R**	
Wall and posters	07L	**07R**	08L	08R	09L	09R	*Windows*
	04L	04R	**05L**	05R	06L	06R	
	01L	**01R**	02L	02R	03L	03R	
Door		△			*Teacher desk*		
	Black board						

Note. Preservice teachers could move freely in the light grey area. *Objects* were inanimate and **student avatars** behaved according to a predetermined script. △ represents the entry point when participants entered the VR classroom. Large class size had 30 (15 pairs) student avatars, small class size had 10 student avatars. The grey seats were empty in the small class size condition.

Table A.3
Behavior Script of Student Avatars

Event ID	Time in Seconds	Avatar Location	Behavior	Category
1	330	02L	Play with a pen	off-task
2	345	02L	Write on the notebook	on-task
3	350	11R	Chat with the neighbor	off-task
4	365	11L	Write on the notebook	on-task
5	365	11R	Idle	
6	365	13R	Eat an apple	off-task
7	380	13R	Idle	
8	380	03R	Chat with the neighbor	off-task
9	395	03L	Idle	
10	395	03R	Idle	
11	380	14R	Stare outside the window	off-task
12	395	14R	Idle	
13	380	11L	Stare outside the window	off-task
14	395	11L	Idle	
15	400	01R	Hit the neighbor	off-task
16	415	01R	Idle	
17	415	01L	Idle	
18	420	05L	Throw paper balls	off-task
19	435	05L	Idle	
20	440	10R	Raise the arm	on-task
21	442	10R	Ask a question	on-task
22	448	10R	Idle	
23	470	12L	Eat an apple	off-task
24	485	12L	Idle	
25	480	11R	Chat with the neighbor	off-task
26	495	11R	Write on the notebook	on-task
27	495	11L	Idle	
28	500	02L	Play with a pen	off-task
29	515	02L	Idle	
30	515	13L	Hit the neighbor	off-task
31	530	13R	Idle	
32	530	13L	Write on the notebook	on-task
33	525	03L	Chat with the neighbor	off-task
34	540	03R	Idle	
35	540	03L	Write on the notebook	on-task
36	540	10R	Raise the arm	on-task
37	542	10R	Ask a question	on-task
38	548	10R	Idle	
39	560	07R	Stare outside the window	off-task
40	575	07R	Idle	
41	575	10R	Stare outside the window	off-task
42	590	10R	Idle	
43	575	11L	Stare outside the window	off-task
44	590	11L	Idle	
45	580	15R	Eat an apple	off-task
46	595	15R	Idle	
47	580	12L	Eat an apple	off-task
48	595	12L	Idle	
49	600	05L	Throw paper balls	off-task
50	615	05L	Idle	

Note. Idle is the default behavior pattern when not specified. When idle, avatars would sit naturally in various neutral poses and move their eyes or body to follow the users around.

Table A.4
Descriptive Statistics of BPM During Different Phases

	Baseline	During Audio Instruction	During Experiment	After Experiment
M	97.2	120.0	161.0	158.0
SD	22.1	10.3	15.9	12.8
Range	110.0	46.8	65.2	58.3
Minimum	46.1	94.3	126.0	132.0
Maximum	157.0	141.0	191.9	190.0

Table A.5
Summary of Measures

Measure	Prompt	Item	Scale
Acute stress	How are you feeling with regard to your emotional state at this moment?		1 = *calm, relaxed, and composed* to 10 = *stressed*
Long-term stress	In the last month, how often have you felt nervous and stressed?		1 = *never* to 10 = *very often*
Cognitive load	During the teaching task, I invested … to this task.		1 = *very, very low mental effort* to 9 = *very, very high mental effort*
Negative VR Experience	How stressful did you find your virtual classroom experience?	• I felt uncomfortable in virtual reality. • I found the time in the virtual classroom exhausting. • In virtual reality I mostly felt insecure. • During the exercise I became hectic. • In virtual reality I lost the red thread (got distracted and confused).	1 = *completely disagree* to 4 = *completely agree*

Appendix B

Linear mixed-effects model (LMM).

In a traditional linear model, the response is modeled as a linear combination of predictors with weights that are referred to as effects or coefficients (Faraway, 2006). In contrast, a linear mixed-effects model (LMM) takes into account both fixed and random effects, in other words, effects that are constant or varying for groups in a population (Gelman & Hill, 2006; Meteyard & Davies, 2020). As reviewed by Kliegl et al. (2011), LMM is superior to analyses of variance (ANOVAs) in experimental studies because of its advantages such as tolerance of an unbalanced experimental design due to missing data (Pinheiro & Bates, 2009), and researchers' capability to examine effects of both categorical factors and continuous variables (covariates), as well as their interactions, with a significant gain in statistical power (Baayen et al., 2008; Meteyard & Davies, 2020). LMM is particularly robust against missing data. In a simulation with 1000 runs of a mixed-effects model, Baayen et al. (2008) found that the main effect was almost always detected with and without 20% missing data (missing at random). Its power is only minimally reduced in the situation of missing data. Despite the fact that power is at its maximum, the Type I error rate is within the nominal range.

Maximum likelihood (ML) and restricted maximum likelihood (REML) estimation.

In the context of mixed-effects modeling, maximum likelihood (ML) estimator is considered biased as it assumes fixed effects are known without uncertainty when estimating the variance components. Restricted maximum likelihood (REML) addresses this issue by separating the estimation of fixed effects from variance components, thereby improving fixed effects standard error estimates (McNeish, 2017). REML has generally been regarded as the default estimation method in mixed-effects models (e.g., Bates et al., 2015; Luke, 2017). The comparison in the formulations can be seen in Zuur's discussion about ML versus REML (2009, p. 116).

Standardization of coefficients.

Standardized coefficients are expressed in equivalent units such as standard deviations of the mean, regardless of the original measurements, when compared to raw unstandardized coefficients. As a result, standardized coefficients can be used to compare the relative magnitude of change associated with different terms in the same model (Grace et al., 2018). Standardized coefficients (b) can be calculated by scaling the raw coefficient β by the ratio of the standard deviation of x over the standard deviation of y. This coefficient is interpreted as: for a 1 standard deviation change in x, a b unit standard deviation change in y is expected.

Goodness of fit.

A statistical model's goodness of fit describes how well it fits a set of observations. Examining indicators of the overall quality of fit of the data with the model is useful. Such indices include the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). They allow a comparison of model fit in models by taking into consideration the number of regression coefficients examined. The AIC value will be the smallest for a model that has a good fit with a small number of predictors. While BIC can be either negative or positive in value, the more negative the value of the BIC, the better the fit (Cohen & Cohen, 2003).

Effect size calculation for fixed-effects terms.

Judd et al. (2016) conceived an effect size estimate (d) that extended Cohen's d (Cohen, 1992) to mixed-effects model with one fixed factor with two levels and two random factors. This estimate might be utilized for designs where the two random factors are crossed and designs where one random factor is nested within the other. Please see Equation (5) in Judd and colleagues' original paper for the mathematical expression of this definition (2016, p. 17.9). An example of the calculation can be seen in Brysbaert and Stevens's (2018) replication study.

Appendix C

Table C.1
Full Model Summary for Physiological Stress Reaction [a]

Fixed Effects	$\hat{\beta}$	$SE(\hat{\beta})$	Standardized $\hat{\beta}$	95% CI	t
Intercept	56.25	5.01		−50.19–11.06	−1.25
Class size (large–small)	17.51	1.34	0.76	14.88–20.14	13.05***
Perceived cognitive load	2.39	1.21	0.20	0.01–4.77	1.96*
Negative VR experience	−3.84	3.24	−0.12	−10.20–2.51	−1.19
Random Effects				Variance	*SD*
Participant (Intercept)				117.31	10.83
Time | Class size (Intercept)				10.18	0.11
Model Fit					
R^2				Marginal	Conditional
				0.75	0.87

Note. $n = 89$ observations. $^{*}p < .05$. $^{**}p < .01$, $^{***}p < .001$. Marginal R^2 is a measure of the proportion of variance explained by the fixed effects, while conditional R^2 is the proportion of variance explained through both fixed and random effects (Nakagawa & Schielzeth, 2013). Table adapted from Meteyard and Davies (2020).

[a] Model specification: Physiological stress reaction ~ 1 + class size + perceived cognitive load + negative VR experience + (1 | participant) + (class size | time).

Table C.2
Full Model Summary for Psychological Stress Reaction [a]

Fixed Effects	$\hat{\beta}$	$SE(\hat{\beta})$	Standardized $\hat{\beta}$	95% CI	t
Intercept	1.58	0.81		0.01–3.16	1.97
Class size (large–small)	0.91	0.18	0.34	0.56–1.27	5.06***
Perceived cognitive load	0.38	0.14	0.27	0.09–0.66	2.61**
Negative VR experience	1.29	0.40	0.35	0.51–2.08	3.22***
Random Effects				Variance	*SD*
Participant (Intercept)				1.76	1.33
Time | Class size (Intercept)				0.28	0.53
Model Fit					
R^2				Marginal	Conditional
				0.54	0.62

Note. $n = 117$ observations. $^{*}p < .05$. $^{**}p < .01$, $^{***}p < .001$.

[a] Model specification: Psychological stress reaction ~ 1 + class size + perceived cognitive load + negative VR experience + (1 | participant) + (class size | time).

References

Achilles, C. M., & Finn, J. D. (2000). Should class size be a cornerstone for educational policy? *The National Center on Education in the Inner Cities (CEIC) Review, 9*(2), 15–23.

Ahrentzen, S., & Evans, G. W. (1984). Distraction, privacy, and classroom design. *Environment and Behavior, 16*(4), 437–454. https://doi.org/10/d8g4r6.

Alaraj, A., Lemole, M. G., Finkle, J. H., Yudkowsky, R., Wallace, A., Luciano, C., Banerjee, P. P., Rizzi, S. H., & Charbel, F. T. (2011). Virtual reality training in neurosurgery: Review of current status and future applications. *Surgical Neurology International, 2*. https://doi.org/10.4103/2152-7806.80117

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., & Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology, 110*(8), 1066–1083. https://doi.org/10.1037/edu0000256

Andreassi, J. L. (2006). *Psychophysiology: Human behavior and physiological response* (5th ed.). Psychology Press. https://doi.org/10.4324/9780203880340

Association of Teachers, & Lecturers. (2009). (School and class) size matters, finds ATL survey. In *Association of teachers and Lecturers Annual Conference*. Liverpool, United Kingdom http://www.atl.org.uk/policy-and-campaigns/conference/atl-conference-2009.asp.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10/fpb5dz.

Bakker, A. B., & Demerouti, E. (2007). The job demands-resources model: State of the art. *Journal of Managerial Psychology, 22*(3), 309–328. https://doi.org/10/cd5.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behnke, R. R., & Carlile, L. W. (1971). Heart rate as an index of speech anxiety. *Speech Monographs, 38*(1), 65–69. https://doi.org/10/fhz8x5.

Blatchford, P. (2003). A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction, 13*(6), 569–595. https://doi.org/10.1016/s0959-4752(02)00043-9

Blatchford, P., & Russell, A. (2019a). Class size, grouping practices and classroom management. *International Journal of Educational Research, 96*, 154–163. https://doi.org/10/gjsj35.

Blatchford, P., & Russell, A. (2019b). New ways of thinking about research on class size: An international perspective. Introduction to the special section. *International Journal of Educational Research, 96*, 120–124. https://doi.org/10/gjsjm7.

Blatchford, P., Russell, A., Bassett, P., Brown, P., & Martin, C. (2007). The role and effects of teaching assistants in English primary schools (years 4 to 6) 2000–2003. Results from the class size and pupil—adult ratios (CSPAR) KS2 Project. *British Educational Research Journal, 33*(1), 5–26. https://doi.org/10/cm6nzj.

Blatchford, P., Russell, A., & Webster, R. (2012). *Reassessing the impact of teaching assistants: How research challenges practice and policy* (1st ed.). Routledge.

Bodie, G. D. (2010). A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education, 59*(1), 70–105. https://doi.org/10/bgn8df.

Borg, M. G., & Falzon, J. M. (1990). Coping actions used by Maltese primary school teachers. *Educational Research, 32*(1), 50–58. https://doi.org/10/dr73fz.

Borko, H. (2016). Methodological contributions to video-based studies of classroom teaching and learning: A commentary. *ZDM, 48*(1), 213–218. https://doi.org/10/gfs2vk.

Borman, G. D. (2009). The use of randomized trials to inform education policy. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 129–138). Routledge.

Brok, P. den, Wubbels, T., & Tartwijk, J. van (2017). Exploring beginning teachers' attrition in The Netherlands. *Teachers and Teaching, 23*(8), 881–895. https://doi.org/10/gjqm5v.

Brouwers, A., Evers, W. J. G., & Tomic, W. (2001). Self-efficacy in eliciting social support and burnout among secondary-school teachers. *Journal of Applied Social Psychology, 31*(7), 1474–1491. https://doi.org/10/fp.968.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 1–20. https://doi.org/10.5334/joc.10

Calandra, B., Gurvitch, R., & Lund, J. (2008). An exploratory study of digital video editing as a tool for teacher preparation. *Journal of Technology and Teacher Education, 16*(2), 137–153.

Chaplain, R. P. (2008). Stress and psychological distress among trainee secondary teachers in England. *Educational Psychology, 28*(2), 195–209. https://doi.org/10/fsq7hq.

Chen, M., & Miller, G. (1997). *Teacher stress: A review of the international literature (ED410187)*. ERIC. http://eric.ed.gov/?id=ED410187.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10/as7.

Cohen, J., & Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). L. Erlbaum Associates.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385–396. https://doi.org/10.2307/2136404. JSTOR.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199. https://doi.org/10/d68qxk.

Cortina, K. S., Miller, K. F., McKenzie, R., & Epstein, A. (2015). Where low and high inference data converge: Validation of CLASS assessment of mathematics instruction using mobile eye tracking with expert and novice teachers. *International Journal of Science and Mathematics Education, 13*(2), 389–403. https://doi.org/10/f68m49.

Delaney, J. P. A., & Brodie, D. A. (2000). Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. *Perceptual & Motor Skills, 91*(2), 515–524. https://doi.org/10/cc5p8f.

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology, 86*(3), 499–512.

Dicke, T., Parker, P. D., Marsh, H. W., Kunter, M., Schmeck, A., & Leutner, D. (2014). Self-efficacy in classroom management, classroom disturbances, and emotional exhaustion: A moderated mediation analysis of teacher candidates. *Journal of Educational Psychology, 106*(2), 569–583. https://doi.org/10/ggdbkr.

van Dick, R. V., & Wagner, U. (2001). Stress and strain in teaching: A structural equation approach. *British Journal of Educational Psychology, 71*(2), 243–259. https://doi.org/10/czz9j6.

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), 162–185. https://doi.org/10.14786/flr.v6i3.372

Doyle, W. (1977). Learning the classroom environment: An ecological analysis. *Journal of Teacher Education, 28*(6), 51–55. https://doi.org/10/c6vkpw.

Doyle, W. (1985). Recent research on classroom management: Implications for teacher education. *Journal of Teacher Education, 36*(3), 31–35. https://doi.org/10/fnkm9c.

Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson, & C. S. Weinstein (Eds.), *Handbook of classroom management. Research, practice, and contemporary issues* (pp. 97–126). Lawrence Erlbaum Associates.

Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management. Research, practice, and contemporary issues*. Lawrence Erlbaum Associates.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC.

Fimian, M. J., & Blanton, L. P. (1987). Stress, burnout, and role problems among teacher trainees and first-year teachers. *Journal of Occupational Behaviour, 8*(2), 157–165. https://doi.org/10/fj44v5.

Finn, J. D. (2019). Academic and non-cognitive effects of small classes. *International Journal of Educational Research, 96*, 125–135. https://doi.org/10/gjsjg5.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97–109. https://doi.org/10/c824tk.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record, 103*(2), 145–183. https://doi.org/10/fgsz3r.

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The "why's" of class size: Student behavior in small classes. *Review of Educational Research, 73*(3), 321–368. https://doi.org/10.3102/00346543073003321

Fitchett, P. G., McCarthy, C. J., Lambert, R. G., & Boyle, L. (2018). An examination of US first-year teachers' risk for occupational stress: Associations with professional preparation and occupational health. *Teachers and Teaching, 24*(2), 99–118. https://doi.org/10/dzf2.

Francis, D. J., Barr, C. D., Benoit, J. S., & McIntyre, T. M. (2017). Issues in research methodology on educator stress. In T. M. McIntyre, S. E. McIntyre, & D. J. Francis (Eds.), *Educator stress: An Occupational health perspective* (pp. 439–470). Springer International Publishing. https://doi.org/10.1007/978-3-319-53053-6_18.

French, N. K. (1993). Elementary teacher stress and class size. *Journal of Research & Development in Education, 26*(2), 66–73.

Galton, M., & Pell, T. (2009). *Study on small class teaching in primary schools in Hong Kong: Final report* (p. 117). Education Bureau, The Government of the Hong Kong.

Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology, 83*(3), 269–275. https://doi.org/10/fgbbg6.

Ganster, D. C., & Rosen, C. C. (2013). Work stress and employee health: A multidisciplinary review. *Journal of Management, 39*(5), 1085–1122. https://doi.org/10/gc4m6z.

Gegenfurtner, A., Lehtinen, E., Helle, L., Nivala, M., Svedström, E., & Säljö, R. (2019). Learning to see like an expert: On the practices of professional vision and visual expertise. *International Journal of Educational Research, 98*, 280–291. https://doi.org/10/ggdbx7.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. http://www.SLQ.eblib.com.au/patron/FullRecord.aspx?p=288457.

Geurts, S. A., & Sonnentag, S. (2006). Recovery as an explanatory mechanism in the relation between acute stress reactions and chronic health impairment. *Scandinavian Journal of Work, Environment & Health, 32*(6), 482–492. https://doi.org/10/dbb52x.

Geving, A. M. (2007). Identifying the types of student and teacher behaviours associated with teacher stress. *Teaching and Teacher Education, 23*(5), 624–640. https://doi.org/10/dmdhn9.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

Goodie, J. L., Larkin, K. T., & Schauss, S. (2000). Validation of the Polar heart rate monitor for assessing heart rate during physical and mental stress. *Journal of Psychophysiology, 14*(3), 159–164. https://doi.org/10/d4x6xc.

Grace, J. B., Johnson, D. J., Lefcheck, J. S., & Byrnes, J. E. K. (2018). Quantifying relative importance: Computing standardized effects in models with binary outcomes. *Ecosphere, 9*(6), 1–24. https://doi.org/10.1002/ecs2.2283

Griffin, M. A., & Clarke, S. (2011). Stress and well-being at work. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization* (pp. 359–397). American Psychological Association. https://doi.org/10.1037/12171-010.

Grossman, P. (Ed.). (2018). *Teaching core practices in teacher education*. Harvard Education Press.

Guglielmi, R. S., & Tatrow, K. (1998). Occupational stress, burnout, and health in teachers: A methodological and theoretical analysis. *Review of Educational Research, 68*(1), 61–99. https://doi.org/10/cn7.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143–163. https://doi.org/10/cnh8wj.

Harfitt, G. J. (2012). An examination of teachers' perceptions and practice when teaching large and reduced-size classes: Do teachers really teach them in the same way? *Teaching and Teacher Education, 28*(1), 132–140. https://doi.org/10/btpj5q.

Harmsen, R., Helms-Lorenz, M., Maulana, R., & Veen, K. van (2018). The relationship between beginning teachers' stress causes, stress responses, teaching behaviour and attrition. *Teachers and Teaching, 24*(6), 626–643. https://doi.org/10/ggjrjb.

Harmsen, R., Helms-Lorenz, M., Maulana, R., Veen, K. van, & Veldhoven, M. van (2019). Measuring general and specific stress causes and stress responses among beginning secondary school teachers in The Netherlands. *International Journal of Research and Method in Education, 42*(1), 91–108. https://doi.org/10/gjprjn.

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research, 43*(6), 387–425. https://doi.org/10/bh4s2h.

Hettiarachchi, I. T., Hanoun, S., Nahavandi, D., & Nahavandi, S. (2019). Validation of Polar OH1 optical heart rate sensor for moderate and high intensity physical activities. *PLoS One, 14*(5), 1–13. https://doi.org/10.1371/journal.pone.0217288

Horgan, K., Howard, S., & Gardiner-Hyland, F. (2018). Pre-service teachers and stress during microteaching: An experimental investigation of the effectiveness of relaxation training with biofeedback on psychological and physiological indices of stress. *Applied Psychophysiology and Biofeedback, 43*(3), 217–225. https://doi.org/10/gd63fs.

Huang, Y., Miller, K. F., Cortina, K. S., & Richter, D. (2021). Teachers' professional vision in action: Comparing expert and novice teacher's real-life eye movements in the classroom. *Zeitschrift Für Pädagogische Psychologie*, 1–18. https://doi.org/10.1024/1010-0652/a000313

Huang, Y., Richter, E., Kleickmann, T., & Richter, D. (2022). Virtual reality in teacher education from 2010 to 2020: A review of program implementation, intended outcomes, and effectiveness measures. In K. Scheiter, & I. Gogolin (Eds.), *Bildung für eine digitale Zukunft [Education for a Digital Future]*. Springer. In press https://edarxiv.org/ye6uw/.

Huang, Y., Richter, E., Kleickmann, T., Wiepke, A., & Richter, D. (2021). Classroom complexity affects student teachers' behavior in a VR classroom. *Computers and Education, 163*, Article 104100. https://doi.org/10.1016/j.compedu.2020.104100

Huberman, A. M., Grounauer, M.-M., & Marti, J. (1993). *The lives of teachers*. Teachers College Press.

Hugdahl, K. (1981). The three-systems-model of fear and emotion—a critical examination. *Behaviour Research and Therapy, 19*(1), 75–83. https://doi.org/10/cpr7z4.

I'Anson, J., Rodrigues, S., & Wilson, G. (2003). Mirrors, reflections and refractions: The contribution of microteaching to reflective practice. *European Journal of Teacher Education, 26*(2), 189–199. https://doi.org/10/fhbjfd.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test, 18*(1), 1–43. https://doi.org/10/c3sjcz.

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies, 23*(4), 1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Johnson, S., Cooper, C., Cartwright, S., Donald, I., Taylor, P., & Millet, C. (2005). The experience of work-related stress across occupations. *Journal of Managerial Psychology, 20*(2), 178–187. https://doi.org/10/fqv766.

Judd, C. M., Westfall, J., & Kenny, D. A. (2016). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*(1), 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

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/gj37r2.

Kanaga, K. R., & Flynn, M. (1981). The relationship between invasion of personal space and stress. *Human Relations, 34*(3), 239–248. https://doi.org/10/bghkjz.

Kanner, A. D., Coyne, J. C., Schaefer, C., & Lazarus, R. S. (1981). Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine, 4*(1), 1–39. https://doi.org/10/dssg34.

Kim, M. J., Lee, C.-K., & Jung, T. (2020). Exploring consumer behavior in virtual reality tourism using an extended stimulus-organism-response model. *Journal of Travel Research, 59*(1), 69–89. https://doi.org/10.1177/0047287518818915

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology, 102*(3), 741–756. https://doi.org/10.1037/a0019237

Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? *Educational Psychology Review, 23*(1), 21–43. https://doi.org/10/ffmpwk.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*(7), 262–273. https://doi.org/10/dc7vsr.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology, 1*, 1–12. https://doi.org/10/ft4h2w.

Klusmann, U., Aldrup, K., Schmidt, J., & Lüdtke, O. (2021). Is emotional exhaustion only the result of work experiences? A diary study on daily hassles and uplifts in different life domains. *Anxiety, Stress & Coping, 34*(2), 173–190. https://doi.org/10/gj34h5.

Kounin, J. S. (1970). *Discipline and group management in classrooms*. Holt, Rinehart & Winston.

Krämer, S., & Zimmermann, F. (2019). Zum Einfluss von störendem Schülerverhalten im Unterricht auf Leistungsbeurteilungen: Explizite Einschätzungen und experimentelle Befunde [On the influence of disruptive student behavior in the classroom on performance assessments: Explicit assessments and experimental findings]. *Zeitschrift Für Pädagogische Psychologie [Journal for Educational Psychology], 34*(2), 99–115. https://doi.org/10/ggr59f.

Kyriacou, C. (2001). Teacher stress: Directions for future research. *Educational Review, 53*(1), 27–35. https://doi.org/10/fgv2b2.

Kyriacou, C., & Sutcliffe, J. (1978). A model of teacher stress. *Educational Studies, 4*(1), 1–6. https://doi.org/10/dm3c6w.

Kyrlitsias, C., & Michael-Grigoriou, D. (2018). Asch conformity experiment using immersive virtual reality. *Computer Animations and Virtual Worlds, 29*(5), 1–16. https://doi.org/10/ggq6v3.

Lambert, R. G., McCarthy, C., O'Donnell, M., & Wang, C. (2009). Measuring elementary teacher stress and coping in the classroom: Validity evidence for the classroom appraisal of resources and demands. *Psychology in the Schools, 46*(10), 973–988. https://doi.org/10/bntpw5.

Latané, B. (1981). The psychology of social impact. *American Psychologist, 36*(4), 343–356. https://doi.org/10/b7vjn5.

Lazarus, R. S. (1966). *Psychological stress and the coping process*. McGraw-Hill.

Lele, A. (2013). Virtual reality and its military utility. *Journal of Ambient Intelligence and Humanized Computing, 4*(1), 17–26. https://doi.org/10.1007/s12652-011-0052-4

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202. https://doi.org/10/gd23bv.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10/gbsd4m.

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research & Development, 66*(5), 1141–1164. https://doi.org/10/gd52wq.

McIntyre, N. A., & Foulsham, T. (2018). Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instructional Science, 46*(3), 435–455. https://doi.org/10/gdmpkc.
McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the kenward-roger correction. *Multivariate Behavioral Research, 52*(5), 661–670. https://doi.org/10/ggxb5f.
Meijman, T. F., & Mulder, G. (1998). Psychological aspects of workload. In *A handbook of work and organizational psychology*. Psychology Press.
Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092
Meurs, J. A., & Perrewé, P. L. (2011). Cognitive activation theory of stress: An integrative theoretical approach to work stress. *Journal of Management, 37*(4), 1043–1068. https://doi.org/10/fp88p.4.
Mikropoulos, T. A., & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999-2009). *Computers and Education, 56*(3), 769–780. https://doi.org/10.1016/j.compedu.2010.10.020
Mostajeran, F., Balci, M. B., Steinicke, F., Kühn, S., & Gallinat, J. (2020). The effects of virtual audience size on social anxiety during public speaking. In *2020 IEEE Conference on virtual reality and 3D user Interfaces (VR)* (pp. 303–312). https://doi.org/10/gjsjhq.
Moya-Albiol, L., Serrano, M. A., & Salvador, A. (2010). Burnout as an important factor in the psychophysiological responses to a work day in teachers. *Stress and Health, 26*(5), 382–393. https://doi.org/10/bpmnvs.
Mulholland, L. A. (1999). *Arizona charter school progress evaluation*. http://eric.ed.gov/?id=ED429339.
Murray-Harvey, R., Slee, P. T., Lawson, M. J., Silins, H., Banfield, G., & Russell, A. (2000). Under Stress: The concerns and coping strategies of teacher education students. *European Journal of Teacher Education, 23*(1), 19–35. https://doi.org/10/b3n3vt.
Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x
OECD. (2020). *Education at a glance 2020: OECD indicators*. OECD. https://doi.org/10.1787/69096873-en
Ouellette, R., Frazier, S., Shernoff, E., Cappella, E., Mehta, T., Maríñez-Lora, A., Cua, G., & Atkins, M. (2017). Teacher job stress and satisfaction in urban schools: Disentangling individual, classroom, and organizational level influences. *Behavior Therapy, 49*. https://doi.org/10.1016/j.beth.2017.11.011
Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429
Phillips, K., & Downer, J. (2017). Classroom context and years of teaching experience as predictors of misalignment on ratings of preschoolers' classroom engagement. *Early Education & Development, 28*(3), 343–367. https://doi.org/10.1080/10409289.2016.1218730
Pines, A. M. (2002). Teacher burnout: A psychodynamic existential perspective. *Teachers and Teaching, 8*(2), 121–140. https://doi.org/10/c2t58d.
Pinheiro, J. C., & Bates, D. M. (2009). *Mixed-effects models in S and S-PLUS (reprinted paperback ed. of the 2000 ed)*. Springer.
Plass, J. L., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load theory*. Cambridge University Press.
Punch, K. F., & Tuettemann, E. (1990). Correlates of psychological distress among secondary school teachers. *British Educational Research Journal, 16*(4), 369–382. https://doi.org/10/chqdmw.
R Core Team. (2021). *R: A language and environment for statistical computing (4.0.4) [computer software]*. R Foundation for Statistical Computing. https://www.R-project.org/.
Rissén, D., Melin, B., Sandsjö, L., Dohns, I., & Lundberg, U. (2000). Surface EMG and psychophysiological stress reactions in women during repetitive work. *European Journal of Applied Physiology, 83*(2), 215–222. https://doi.org/10.1007/s004210000281
Ritvanen, T., Louhevaara, V., Helin, P., Halonen, T., & Hänninen, O. (2003). Psychophysiological stress in high school teachers. *International Journal of Occupational Medicine & Environmental Health, 16*(3), 255–264.
Ritvanen, T., Louhevaara, V., Helin, P., Väisänen, S., & Hänninen, O. (2006). Responses of the autonomic nervous system during periods of perceived high and low work stress in younger and older female teachers. *Applied Ergonomics, 37*(3), 311–318. https://doi.org/10/dgtbkt.
Roberts, A. C., Yeap, Y. W., Seah, H. S., Chan, E., Soh, C.-K., & Christopoulos, G. I. (2019). Assessing the suitability of virtual reality for psychological testing. *Psychological Assessment, 31*(3), 318–328. https://doi.org/10.1037/pas0000663
Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*(1), 63–88. https://doi.org/10/chqpdr.
Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology, 76*(4), 647–658. https://doi.org/10/bq9z6d.
Saloviita, T., & Pakarinen, E. (2021). Teacher burnout explained: Teacher-, student-, and organisation-level variables. *Teaching and Teacher Education, 97*, 1–14. https://doi.org/10.1016/j.tate.2020.103221
Schmidt, J., Klusmann, U., Lüdtke, O., Möller, J., & Kunter, M. (2017). What makes good and bad days for beginning teachers? A diary study on daily uplifts and hassles. *Contemporary Educational Psychology, 48*, 85–97. https://doi.org/10.1016/j.cedpsych.2016.09.004
Schubert, C., Lambertz, M., Nelesen, R. A., Bardwell, W., Choi, J.-B., & Dimsdale, J. E. (2009). Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological Psychology, 80*(3), 325–332. https://doi.org/10.1016/j.biopsycho.2008.11.005
Serrano, M. A., Moya-Albiol, L., & Salvador, A. (2008). The role of gender in teachers' perceived stress and heart rate. *Journal of Psychophysiology, 22*(1), 58–64. https://doi.org/10/b8dwzr.
Serrano, M. A., Moya-Albiol, L., & Salvador, A. (2013). Psychobiological responses at the beginning and the end of an academic year in teachers. *Anales de Psicología, 29*(2), 590–597. https://doi.org/10/gjt7sf.
Seufert, C., Oberdörfer, S., Roth, A., Grafe, S., Lugrin, J.-L., & Latoschik, M. E. (2022). Classroom management competency enhancement for student teachers using a fully immersive virtual classroom. *Computers & Education, 179*, 104410. https://doi.org/10/gnw5bk.
Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Eds.). (2010). *Mathematics teacher noticing*. Routledge.
Skaalvik, E. M., & Skaalvik, S. (2017). Teacher stress and teacher self-efficacy: Relations and consequences. In T. M. McIntyre, S. E. McIntyre, & D. J. Francis (Eds.), *Educator stress: An Occupational health perspective* (pp. 101–125). Springer International Publishing. https://doi.org/10.1007/978-3-319-53053-6_5.
Solheim, O. J., & Opheim, V. (2019). Beyond class size reduction: Towards more flexible ways of implementing a reduced pupil–teacher ratio. *International Journal of Educational Research, 96*, 146–153. https://doi.org/10.1016/j.ijer.2018.10.008
Sonnentag, S., & Fritz, C. (2015). Recovery from job stress: The stressor-detachment model as an integrative framework. *Journal of Organizational Behavior, 36*, S72–S103. https://doi.org/10/f6xv9x.
Spencer, S., Drescher, T., Sears, J., Scruggs, A. F., & Schreffler, J. (2019). Comparing the efficacy of virtual simulation to traditional classroom role-play. *Journal of Educational Computing Research, 57*(7), 1772–1785. https://doi.org/10.1177/0735633119855613
Stavroulia, K.-E., Christofi, M., Baka, E., Michael-Grigoriou, D., Magnenat-Thalmann, N., & Lanitis, A. (2019). Assessing the emotional impact of virtual reality-based teacher training. *The International Journal of Information and Learning Technology, 36*(3), 192–217. https://doi.org/10.1108/IJILT-11-2018-0127
Stockero, S. L., Rupnow, R. L., & Pascoe, A. E. (2017). Learning to notice important student mathematical thinking in complex classroom interactions. *Teaching and Teacher Education, 63*, 384–395. https://doi.org/10/f93pkn.
Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A., & Van Huffel, S. (2011). Instantaneous changes in heart rate regulation due to mental load in simulated office work. *European Journal of Applied Physiology, 111*(7), 1497–1505. https://doi.org/10.1007/s00421-010-1776-0
Taris, T. W., Leisink, P. L. M., & Schaufeli, W. B. (2017). Applying occupational health theories to educator stress: Contribution of the job demands-resources model. In T. M. McIntyre, S. E. McIntyre, & D. J. Francis (Eds.), *Educator stress: An Occupational health perspective* (pp. 237–259). Springer International Publishing. https://doi.org/10.1007/978-3-319-53053-6_11.
Tarr, M. J., & Warren, W. H. (2002). Virtual reality in behavioral neuroscience and beyond. *Nature Neuroscience, 5*(11), 1089–1092. https://doi.org/10/c6vc67.
Theelen, H., van den Beemt, A., & Brok, P. den (2019). Classroom simulations in teacher education to support preservice teachers' interpersonal competence: A systematic literature review. *Computers & Education, 129*, 14–26. https://doi.org/10.1016/j.compedu.2018.10.015

Timperley, H., & Robinson, V. (2000). Workload and the professional culture of teachers. *Educational Management & Administration, 28*(1), 47–62. https://doi.org/10/fr7792.

Troman, G. (2000). Teacher stress in the low-trust society. *British Journal of Sociology of Education, 21*(3), 331–353. https://doi.org/10/b39xsb.

Troman, G., & Woods, P. (2001). *Primary teachers' stress.* RoutledgeFalmer.

Wallach, H. S., Safir, M. P., & Bar-Zvi, M. (2009). Virtual reality cognitive behavior therapy for public speaking anxiety: A randomized clinical trial. *Behavior Modification, 33*(3), 314–338. https://doi.org/10/c6mcnq.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wiepke, A., Heinemann, B., Lucke, U., & Schroeder, U. (2021). Jenseits des eigenen Klassenzimmers: Perspektiven & Wei- terentwicklungen des VR-Klassenzimmers [Beyond the own classroom: Perspectives & further developments of the VR classroom.]. In *Delfi 2021 - Die 19. Fachtagung Bildungstechnologien [educational technologies symposium]* (pp. 331–336). http://dl.gi.de/handle/20.500.12116/37031.

Wilhelm, K., Dewhurst-Savellis, J., & Parker, G. (2000). Teacher stress? An analysis of why teachers leave and why they stay. *Teachers and Teaching, 6*(3), 291–304. https://doi.org/10/bs4fv3.

Wolff, C. E., Jarodzka, H., van den Bogert, N., & Boshuizen, H. P. A. (2016). Teacher vision: Expert and novice teachers' perception of problematic classroom management scenes. *Instructional Science, 44*(3), 243–265. https://doi.org/10/f8tcpm.

Ye, X., Liu, P.-F., Lee, X.-Z., Zhang, Y.-Q., & Chiu, C.-K. (2019). Classroom misbehaviour management: An SVVR-based training system for preservice teachers. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2019.1579235, 0(0).

Zuur, A. F. (Ed.). (2009). *Mixed effects models and extensions in ecology with R.* Springer.

Wiepke, A., Richter, E., Zender, R., & Richter, D. (2019). Einsatz von Virtual Reality zum Aufbau von Klassenmanagement-Kompetenzen im Lehramtsstudium [Using virtual reality to build classroom management skills in undergraduate teacher education]. In N. Pinkwart & J. Konert (Eds.), *DELFI 2019—Die 17. Fachtagung Bildungstechnologien [Educational technologies symposium]* (pp. 133–144). Gesellschaft für Informatik e.V. https://doi.org/10.18420/delfi2019_319.

Wiepke, A., Heinemann, B., Lucke, U., & Schroeder, U. (2021). Jenseits des eigenen Klassenzimmers: Perspektiven & Weiterentwicklungen des VR-Klassenzimmers [Beyond the own classroom: Perspectives & further developments of the VR classroom.]. *DELFI 2021 - Die 19. Fachtagung Bildungstechnologien [Educational technologies symposium]*, 331–336. http://dl.gi.de/handle/20.500.12116/37031.