Test Anxiety and Physiological Arousal: A Systematic Review and Meta-Analysis

Anna-Lena Roos

University of Applied Sciences and Arts Northwestern Switzerland

Thomas Goetz and Martin Voracek

University of Vienna

Maike Krannich

University of Zurich

Madeleine Bieg

Center for Psychiatry Reichenau

Amanda Jarrell

McGill University

Reinhard Pekrun

University of Essex and Australian Catholic University

Author Note

Anna-Lena Roos*, Institute for Research and Development of Collaborative Processes, School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Olten, Switzerland, Postal address: FHNW School of Applied Psychology, Riggenbachstrasse 16, CH-4600 Olten, Switzerland; Thomas Goetz, Department of Developmental and Educational Psychology, University of Vienna, Austria; Martin Voracek, Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Austria; Maike Krannich, Teaching and Educational Technology,

Institute of Education, University of Zurich, Zurich, Switzerland; Madeleine Bieg, Research Unit, Center for Psychiatry Reichenau, Reichenau, Germany; Amanda Jarrell, Department of Educational and Counselling Psychology, McGill University, Montreal, Canada; Reinhard Pekrun, Department of Psychology, University of Essex, Essex, UK, and Institute for Positive Psychology and Education, Australian Catholic University, Sydney, Australia.

*Corresponding author: Anna-Lena Roos, annalena.roos@fhnw.ch, +41 62 957 29 09

Test Anxiety and Physiological Arousal: A Systematic Review and Meta-Analysis

Resubmitted on: 28 April 2020

Abstract

Test anxiety is a widespread and mostly detrimental emotion in learning and achievement settings. Thus, it is a construct of high interest for researchers and its measurement is an important issue. So far, test anxiety has typically been assessed using self-report measures. However, physiological measures (e.g., heart rate or skin conductance level) have gained increasing attention in educational research, as they allow for an objective and often continuous assessment of students' physiological arousal (i.e., the physiological component of test anxiety) in real-life situations, such as a test. Although theoretically one would assume self-report measures of test anxiety and objective physiological measures would converge, empirical evidence is scarce and findings have been mixed. To achieve a more coherent picture of the relationship between these measures, this systematic review and meta-analysis investigated whether higher self-reported test anxiety is associated with expected increases in objectively measured physiological arousal. A systematic literature search yielded an initial 231 articles, and a structured selection process identified 29 eligible articles, comprising 31 studies, which met the specified inclusion criteria and provided sufficient information about the relationship under investigation. In line with theoretical models, in 21 out of the 31included studies there was a significant positive relationship between self-reported test anxiety and physiological arousal. The strengths of these correlations were of medium size. Moderators influencing the relation between these two measures are discussed, along with implications for the assessment of physiological data in future classroom-based research on test anxiety.

Keywords: test anxiety, physiological measures, self-report, systematic review, meta-analysis

Test Anxiety and Physiological Arousal: A Systematic Review and Meta-Analysis

In educational research there is a growing interest in using physiological measures of affect (e.g., cardiovascular measures, such as heart rate measurement, electrodermal measures, or saliva samples) to complement traditional self-report measures of emotion (D'Mello and Calvo 2013). However, the reliability of these measures, which assess physiological arousal as a proxy for emotional processes, is still questioned, and interpreting findings obtained from these measures in relation to existing educational research and theory remains a challenge in educational psychology (Mauss et al. 2005; Mauss and Robinson 2009; Kreibig 2010). Therefore, the purpose of the present review is to address these issues by systematically reviewing and meta-analyzing research findings on the relationship between physiological and self-report measures which traditionally have been considered as benchmarks for assessing emotions (Larsen and Prizmic-Larsen 2006). We focus in our study on test anxiety – one of the most prominent and intensively investigated emotions in the academic context.

Test anxiety has been extensively investigated since the early 1950s (Mandler and Sarason 1952). To date, there are more than 2,000 published studies on this often detrimental emotion (e.g., Hembree 1988; McDonald 2001; Zeidner 2014). A primary indicator of test anxiety is heightened levels of physiological arousal. It can be unobtrusively assessed through physiological measures, such as cardiovascular measures (e.g., heart rate) or electrodermal measures (e.g., skin conductance response). Physiological responses associated with anxiety are also difficult to suppress or control. Therefore, physiological measures represent an objective way of assessing students' arousal in testing situations (Harley 2015; Houtveen and de Geus 2009; Wilhelm and Grossman 2010).

Nevertheless, efforts to measure test anxiety typically have relied on self-report measures, which can be influenced by factors such as social desirability or subjective beliefs (Goetz et al. 2013; Kaplan et al. 2013; Pekrun and Bühner 2014; Robinson and Clore 2002).

Therefore, it is possible that self-report assessments of test anxiety might not accurately reflect the actual ("true") level of test anxiety.

Several theories of emotion support the assumption that self-reported emotion experiences and physiological responses would be related, for test anxiety and other emotions. associations. For example, as outlined by Barrett (2014) in her conceptual act theory, the brain interprets bodily inputs resulting in the experience of corresponding emotions. However, input from the natural world is also taken into account. Further, internally generated input is considered contributing to cognition. Thus, depending on individual attentional focus (i.e. body, world or internal), the same bodily reaction can result in more or less intense emotional episodes across different persons or across different situations. The notion that bodily information is translated to emotional episodes has been outlined in numerous studies (e.g., D'Mello et al. 2018). The relation between emotion self-report and physiology is also assumed in the functionalist perspective of emotion (e.g., Lench et al. 2011). This approach is grounded in the idea that discrete emotions initiate an evolutionary adaptive response to changes in the environment. These responses include cognitive, behavioral, experiential, and physiological reactions (see also Rosenberg and Ekman 1994; for a critic on this approach see Lench et al. 2013). From this perspective, clear relations between emotions and coordinated physiological responses can be assumed.

However, as self-reports might be impacted by subjective beliefs about emotions (Robinson and Clore, 2002), relations between objective measures of physiology and self-reports of emotions might differ as a function of individual belief systems. Further, there is an ongoing critique of the functionalist perspective with some researchers arguing that this position primarily reflects a behaviourist, stimulus-response, approach to emotion, even though emotions should be conceived as being constructed by the human mind (e.g., Lindquist et al. 2013; see on this discussion also Lench et al. 2013).

Finally, and referring to classical component models of emotions (e.g., Kleinginna and Kleinginna 1981; Scherer 2000), since heightened physiological arousal is one constituting component of test anxiety (Zeidner 1998), it is reasonable to assume that students' objectively measured levels of physiological arousal during test taking and levels of self-reported test anxiety (i.e. often including the subjective experience of all components) would converge (i.e. showing a certain degree of overlap). For instance, when test anxious students are asked about their feelings while taking a test, they should report both high levels of anxiety and perceptions of high physiological arousal (e.g., a faster heart beat). However, empirical findings investigating these relations have been mixed (e.g., Mauss et al. 2005; Mauss and Robinson 2009). One reason might be, that arousal is just one component of anxiety.

In sum, various theoretical approaches lead to the assumption that physiological measures of arousal and self-reports of test anxiety should be related. However, all approaches also lead to the assumption that there are factors which might influence these relations. Given that the nature of the relations between physiology and subjective experience is controversial, it would be important to gain knowledge about the relationship between these measures. This might contribute to a deeper understanding of the nature of test anxiety, which is an important research theme, especially in educational psychology (Campbell 1986; Van Yperen 2007; Zeidner 2014).

The aim of this review and meta-analysis is to provide a more coherent picture of the relationship between these two measures of test anxiety in school and academic contexts, and to identify possible moderating variables between them by systematically investigating whether self-reported test anxiety is associated with objectively measurable physiological arousal (e.g., heart rate, blood pressure, or cortisol level). This review might help determine how closely physiological measures of arousal reflect subjective reports of test anxiety and identify potential moderators of this relation (e.g., sample characteristics). Moderators and their functions could be considered in further studies. Further, traditionally in test anxiety

research, two components of anxiety are differentiated, namely worry and emotionality. Our study will address whether there are differences between how self-reported worry and emotionality relate to physiological arousal. In the long run, a deeper understanding might also help guiding the development and implementation of interventions counteracting test anxiety.

We start with defining test anxiety and its constituting components. We then describe the two assessment methods discussed in this review in detail and point out advantages and disadvantages of each method, followed by a description of the research questions of the current review. Next, we provide a systematic review of empirical findings on the relationship between these self-report and physiological measures of anxiety, along with a quantitative synthesis of the evidence (meta-analysis). The last section summarizes the findings and discusses possible moderator variables (i.e., in terms of study design and setting or sampling rate of physiological measure) influencing the relation between the two measures, as well as methodological and practical implications for the assessment of physiological data in future research in educational psychology.

Concept and Assessment of Test Anxiety

Test Anxiety and its Components

Test anxiety is a specific type of anxiety that can be experienced before, during, and after an examination or other test situations (Beidel and Turner 1988; Zeidner 2014). It is usually conceptualized as comprising multiple theoretically distinct, yet correlated, components (Zeidner 2014; Pekrun et al. 2004). Liebert and Morris (1967) first introduced the distinction between the “worry” (cognitive) and “emotionality” (affective) components of test anxiety. The worry component has been defined as a cognitive concern about one’s performance (e.g., thinking about the potential consequences of failure in examinations and tests, as well as doubts about one’s ability to perform adequately). Emotionality refers to negative affective and physiological experiences, such as arousal and unease (e.g., nervous

stomach, muscle tension, sweaty palms, or generally upset feelings) that may occur during the stressful testing situation (King et al. 2000; Morris and Liebert 1970). In more recent studies, researchers have moved from a dichotomous view to multi-component views of emotions and more specifically test anxiety (Pekrun et al. 2002; Shuman and Scherer 2014; Zeidner and Matthews 2005), which frequently include affective, cognitive, physiological, motivational, and behavioral components (Matthews et al. 1999; Pekrun et al. 2004; Scherer 1984).

Assessing Test Anxiety

Test anxiety can be assessed through multiple measures, including self-report, behavioral measures (e.g., detection and classification of facial expressions), measures based on neuroimaging techniques (e.g., fMRI), and measures of peripheral physiological processes (e.g., heart rate or skin conductance). This paper focuses on two types of test anxiety measures, namely, self-reports (as the traditionally used measures) and physiological measures (as emerging measures in the field of educational psychology), both of which are described in detail in the following.

Self-report measures. Self-report measures are the most common measures to assess achievement emotions, such as test anxiety (Larsen and Prizmic-Larsen 2006). These measures ask individuals to report their typical or habitual level of test anxiety (i.e., trait test anxiety) or their current level of test anxiety (i.e., state test anxiety). Trait test anxiety measures are the most popular type of self-report measures. One of the most frequently used self-report instruments is the Test Anxiety Inventory (TAI, Spielberger et al. 1980; Szafranski et al. 2012). Trait test anxiety measures typically involves paper-pencil questionnaires (e.g., a typical sample item would be: “Even when I am well prepared for a test I feel very anxious about it”; TAS, Sarason 1978). It is common for these measures to assess both the cognitive component (also referred to as worry) and the affective component of test anxiety (also referred to as emotionality; Morris et al. 1981). Worry is assessed with items about being concerned about one’s achievement or the consequences of failure, whereas the emotionality

component is assessed with items about the emotional distress experience (e.g., nervousness or tension) or self-perceived arousal in terms of reactions of one's autonomic nervous system (Spielberger et al. 1980). The advantages of trait measures of anxiety are that they offer quick and cost-effective assessments of self-observed thoughts and emotional experiences (Wilhelm and Grossman 2010). However, these measures can be biased, for instance by memory effects and by subjective beliefs (Robinson and Clore 2002; Buehler and McFarland 2001; Fredrickson and Kahneman 1993; Goetz et al. 2013).

State test anxiety measures examine students' momentarily experienced anxiety in a test situation (e.g., via ambulatory psychological assessment; a typical sample item could be: "How much anxiety are you experiencing right now?"). These state assessments of anxiety are less susceptible to bias due to closer temporal proximity to the test (i.e. real-time or online self-reports) and because they are situated in the testing situation (i.e., in real life, e.g., in the school context during a test; Wilhelm and Grossman 2010). However, during tests it can be difficult for students to report their current emotional state (e.g., because of high cognitive load during a test; Putwain 2007; Wilhelm and Grossman 2010). Moreover, there are concerns that even such online self-reports are also susceptible to bias. For example, individual differences in the awareness of and willingness to report on emotional states (i.e., response tendencies or social desirability) can influence responding (Mauss and Robinson 2009). Therefore, researchers often emphasize the need to use measures that are less susceptible to such biases (Podsakoff 2003).

Physiological measures. The above limitations of self-report measures call for applying physiological measures of emotions (Houtveen and de Geus 2009; Wilhelm and Roth 2001). Physiological channels (e.g., heart rate, cortisol sampling, or skin conductance) are difficult for individuals to mask or control, which creates the possibility of more objectively gauging individuals' arousal in testing situations (Harley 2015; Scollon et al. 2009; Houtveen and de Geus 2009). There are two primary biological systems that become

active in a coordinated effort to respond to stress and, by extension, to test anxiety. The first is the autonomic nervous system which is responsible for modulating peripheral function and consists of sympathetic and parasympathetic branches, which generally are associated with activation and relaxation, respectively (Mauss and Robinson 2009). In response to stress, the autonomic nervous system reacts by making the body ready for action, including increased heart rate and vasodilation. However, it should be noted that different emotions (e.g., anxiety and excitement) can have the same physiological signatures and, since the autonomic nervous system serves a general purpose, its activity is not exclusively a function of emotional responding, but rather encompasses a wide variety of other functions, such as homeostasis (i.e., keeping the body's internal environment in balance) and digestion (Cacioppo et al. 2000). The second system is the adrenocortical system which is activated via the hypothalamic-pituitary-adrenal (HPA) axis. This system is responsible for regulating the hormone cortisol, which often is referred to as the classical stress hormone (Lundberg 2005). Physiological measures of arousal measure the body's above-mentioned response to stress.

As the body's responses to stress are numerous, there are several physiological measures to examine test anxiety. For the present review, we selected those measures which typically and currently are used (Zeidner and Matthews 2011; Hodges 2015). These include cardiovascular measures, such as heart rate (HR), heart rate variability (HRV), and blood pressure (BP); electrodermal activity measures (i.e., often also referred to as electrodermal activity, EDA; Boucsein 2012), such as skin conductance (SC) or skin resistance (SR); and endocrine measures, such as cortisol sampling (Kantor et al. 2001; Martin 1961). Recently, measures of pH level in saliva have gained increased attention (Marques et al. 2010), as there is evidence for a negative association between saliva pH and anxiety (Morse et al. 1982). We therefore decided to additionally include this measure in our review and meta-analysis. We now describe each of these measures in more detail.

Heart rate (HR) is a cardiovascular measure (i.e., response of the blood circulatory system) and refers to the number of heart beats per time interval (e.g., beat per minute, bpm; Hugdahl 1995). The normal resting heart rate is between 60 and 100 bpm (Spodick 1993; Avram et al. 2018). Heart rate provides noninvasive information about modulation of heart rate by the autonomic nervous system (McCraty et al. 1995) and sometimes is assumed to directly reflect anxious arousal (i.e., positive correlation: the higher the heart rate, the more test anxiety; Calvo and Miguel-Tobal 1998). However, measuring heart rate also has some limitations, as it does not only increase due to emotional states, but also with increased mental work load at the beginning of a task or with increased physical activity (Myrtek et al. 1990; Saito and Nakamura 1995). Therefore, in emotion research heart-rate change scores or so-called additional heart rate is often used, which is defined as heart rate acceleration without a corresponding increase of physical activity (Turner et al. 1988).

Heart rate variability (HRV) is commonly described by the normally occurring variation of the time interval between heartbeats and is expressed by the interbeat interval (IBI; Shaffer et al. 2014). IBI are fluctuations in heart rate which result from complex interactions among a number of different physiological systems. Interest in the relationship between heartrate variability and negative emotional reactions has increased after findings suggesting that heart-rate variability is associated with perceived stress (Sloan et al. 1994). An optimal level of heart rate variability intraindividually is presumed to reflect healthy function, adaptability, and resilience (McCraty and Zayas 2014). While very high heart-rate variability (i.e., too much instability, such as cardiac or ventricular arrhythmias) can be detrimental to efficient physiological functioning, low heart rate variability (i.e., too little variation) may be a risk factor for cardiac events (e.g., heart attack, stroke) and can predict the development of hypertension (i.e., high blood pressure), because it reflects reduced regulatory capacity to respond to physiological challenges, such as stress and exercise (Berntson et al. 2008;

Matthews et al. 2003; Bigger et al. 1992; Singh et al. 1998). Thus, one would expect a negative correlation between HRV and test anxiety.

Blood pressure (BP) is measured in millimeters of mercury (mmHg) and is multiply determined: by the amount of blood pumped by the heart and by how easily blood flows through arteries (Obrist 1976). Systolic blood pressure (SBP) refers to the amount of pressure exerted against arterial walls when the heart is contracting (i.e., maximum during one heart beat). Diastolic blood pressure (DBP) describes the amount of pressure exerted against arterial walls between heart beats (i.e., minimum between two heart beats; Hugdahl 1995). Temporary increases in blood pressure and heart rate provide an indication of the activation of the sympathetic nervous system, which helps to supply the body with resources to fight an immediate stressor. Therefore, blood pressure and anxiety levels are assumed to be positively correlated. With advanced age, blood pressure reactivity to stressful events typically increases (Uchino et al. 2006; Uchino et al. 2005). Therefore, it can be important to consider age as a factor that potentially influences the association between anxiety and physiological response.

Skin conductance (SC) and skin resistance (SR) are electrodermal parameters (i.e., sweat gland responses). They are often also referred to as electrodermal activity (EDA; Boucsein 2012). Electrodermal activity varies with the state of sweat glands in the skin. If the sympathetic branch of the autonomic nervous system is highly aroused, then sweat gland activity also increases, which in turn increases skin conductance. In this way, increases in skin conductance (measured in microsiemens, μS) have been found to be correlated with emotional states (i.e. positive correlation with anxiety; Carlson 2013). It is useful to differentiate between tonic skin conductance level (SCL) and short-duration (i.e., phasic) skin conductance responses (SCRs). Skin resistance (SR) is an inverse index of skin conductance (i.e., for test anxiety, one would expect that the higher the SR, the less test anxiety: negative correlation). Electrodermal activity can be most reliably and validly recorded on the palm of the non-dominant hand. Temperature and humidity however can affect EDA measurements,

which can lead to inconsistent results (Boucsein 2012). In comparison to heart rate responses, electrodermal responses are relatively slow. They appear after a stimulus with a delay of approximately one to three seconds. This shows the complexity of determining a relationship between electrodermal activity and other physiological measures (Martin 1961).

Cortisol can be sampled from the blood, the urine, and the saliva and is expressed in micrograms per deciliter (i.e., mcg/dL). Additionally, in the case of chronic stress it can be determined in the hair (Meyer & Novak, 2012). Salivary cortisol levels are highly correlated with free serum cortisol levels (Dorn et al. 2007). Levels of cortisol follow a diurnal rhythm, which means they usually are high in the morning and then decrease throughout the day. An increased release of cortisol activated via the hypothalamus pituitary adrenal axis often is associated with the experience of academic stress during examinations (i.e. we would expect a positive correlation between cortisol level and anxiety; Dickerson and Kemeny 2004). In addition, caffeine consumption, eating patterns, physical activity, or sleep patterns can also affect the release of cortisol (Bouma et al. 2009; Liu et al. 2017; Pollard 1995; Schwartz et al. 1998).

PH level in saliva has gained increased interest and recently has been suggested as a possible useful biomarker indicating psychological stress levels (Marques et al. 2010). The advantage of this measure is that it is inexpensive, noninvasive, and relatively easy to collect. The regulation of saliva volume and its composition is regulated by the sympathetic and parasympathetic nervous system. Under stress, this may lead to a lower rate of secretion of the salivary glands in the mouth (often expressed by dry mouth) in reaction to stress, which in turn leads to increased acidity and a decrease in oral pH (Humphrey and Williamson 2001). This means that we can expect a negative correlation between saliva pH and test anxiety.

Research Questions

As noted at the outset, from a theoretical perspective, one would assume some convergence between self-reported test anxiety (i.e., including the subjective experience of all

components) and the above-described, more specific physiological measures, since heightened physiological arousal is one of the components of test anxiety (Zeidner and Matthews 2005). Furthermore, prominent models of stress (Selye 1976a) also suggest a relationship between emotional self-reports and physiological reactions to exams. To our knowledge, no review has yet systematically examined or quantified, this relation. Therefore, the current review systematically integrated the retrievable empirical research evidence to address this gap in the literature and test this assumed relation. Furthermore, we wanted to identify aspects potentially influencing the relationship between self-report measures of test anxiety and physiological measures (e.g., study design and setting, sample characteristics, or the assessment and analysis of physiological data). Finally, some researchers (e.g., Morris et al. 1981) have stated that physiological indices provide information concerning the emotionality component of anxiety, but little information about the worry component. This view holds that the inconsistent findings on the relations between physiological and self-report measures may be partly due to combining emotionality and worry in one and the same self-report measures. Therefore, we additionally addressed the question whether the relationship between physiological arousal and the emotionality component of self-reported test anxiety is different from the relationship between physiological arousal and the worry component.

Taken together, we addressed the following research questions:

1. Is self-reported test anxiety positively related to objectively measurable physiological arousal?
2. Are there study-specific variables that moderate this relation (i.e. study design and setting, sample characteristics, assessment and analysis of physiological and self-report data)?

3. Are there differential relations between the worry and emotionality components of self-reported test anxiety on the one hand, and physiological arousal on the other hand?

Methods

Literature Search Strategy

We conducted a comprehensive literature search in PsycINFO, PubMed, and ERIC databases up to 5 March 2020 to identify studies addressing the relationship between trait or state self-reported test anxiety and measures of physiological arousal. The physiological measures included cardiovascular measures, electrodermal measures, cortisol sampling, and saliva pH. No date restrictions were placed on any searches.

Study Inclusion Criteria, Study Screening, and Coding

Eligible studies were those that were published in peer-reviewed journals (i.e., the grey literature was not considered) and reported findings on the relationship between test anxiety in school or academic contexts (e.g., elementary school, high school, university) and the above mentioned physiological measures. In the first phase (out of three) of the literature search and screening of studies, publications were preselected using the search terms ("test anxiety" OR "exam anxiety" OR "examination anxiety") AND (specific physiological measure, e.g., "heart rate"). The search terms had to appear either in the title, the keywords, or the abstract of the publication. The exact search terms are listed in Table 1. Next, the results from the initial search results were screened for peer-reviewed empirical studies (e.g., no theoretical work) published in English (see Figure 1).

In the second phase, the full texts of the remaining articles were retrieved. Then, the abstracts and method sections of these studies were screened, and only studies that used at least one self-report measure of test anxiety, in tandem with one kind of physiological measure (as defined by the search terms), were included for a detailed review. To be eligible for inclusion in the systematic review and meta-analysis, the self-report measure needed to be

either a test anxiety questionnaire (i.e., mostly trait test anxiety questionnaires, like the Test Anxiety Scale; TAS, Sarason 1978; or the Test Anxiety Inventory; TAI, Spielberger et al. 1980), or state anxiety items that asked about test anxiety immediately in the context of an exam (i.e., exam-related anxiety before, during, or after an exam) or other testing situation, for example in a lab setting (e.g., intelligence test or problem/anagram solving test). Social-evaluative situations (e.g., public speaking) that did not involve a test instruction were not considered. The same applied for trait measures of general anxiety that were not administered within the context of an exam or testing situation.

In the final step of the second phase, after detailed review of the results, only studies were retained that fulfilled the following criteria: (1) they reported about nonsignificant or significant relationships between self-reported test anxiety and physiological measures, (2) or explicitly reported correlation coefficients, (3) or provided sufficient statistical information enabling to calculate these correlations. Intervention studies were only included when they reported results for the pretreatment period or for the control group. Post-treatment findings for treatment groups were not included, because intervention effects may have confounded any relations. Similarly, in studies that used experimental designs (e.g., designs involving classical conditioning to induce increased arousal) where the relationship between self-reported test anxiety and physiological measures could not be disentangled from the effects of the experimental paradigm, were not included. Studies including participants with a clinical diagnosis, for example an anxiety disorder such as social phobia, were excluded from the current review as well, because test anxiety could not be disentangled from the clinical diagnosis.

In the third phase, reference lists of identified publications were screened for further studies meeting the specified criteria. Procedural details of the three-phase literature search and screening of studies are displayed in Figure 1.

Data Coding and Synthesis

Studies included for review were coded for the study variables identified a priori (for a full description of the variables, see Tables 2 and 3). Any inconsistencies in the study selection process were resolved by discussion between the first author and the coauthors. We decided to qualitatively summarize the studies in the first place and additionally provide descriptive and inferential statistical quantifications of the various relationships between self-reported test anxiety and the set of physiological measures.

Specifically, in order to quantitatively synthesize the available research evidence, we fitted a series of fixed-effect meta-analytic models (FEM) to the study-level data (rather than calculating random-effects meta-analytic models [REM], which would aim to generalize beyond the available research evidence). Fisher's *r*-to-r_z transformation is frequently employed for meta-analyses using the *r* metric; however, effects of this transformation are only noteworthy with substantial *r* coefficients (Lipsey and Wilson 2001), whereas the literature we meta-analyzed is characterized by small-to-medium *r* values. We therefore report meta-analytic results based on untransformed *r* coefficients. Additional analyses, using the REM method instead of the FEM method, or using Fisher's r_z instead of *r*, yielded essentially the same findings than those reported below and, more importantly, left all conclusions unchanged. Thus, for the sake of convenience, we report the FEM results, based on untransformed *r* coefficients.

For each of the set of meta-analytic models, we report: *k* (the number of independent studies included in the respective meta-analysis); *N* (the associated total sample size of the *k* studies included); *r* (the resulting meta-analytic summary effect, along with its 95% confidence interval); *z* (the test statistic evaluating whether the meta-analytic summary *r* differs significantly from zero); Cochran's *Q* test for between-study effect heterogeneity (with $df = k - 1$), testing the null hypothesis of no cross-study effect variability beyond sampling variability; and I^2, the associated relative effect-size metric for quantifying cross-study effect variability, with $I^2 = 0\%$ indicating that the total variance of study effects in a meta-analysis is

as expected (or even less than that) by mere sampling variability, whereas $I^2 > 75\%$ suggesting substantial additional effect heterogeneity across studies beyond the amount due to sampling variability.

We interpreted the magnitude of the meta-analytic average correlations according to the revised guidelines (Lipsey and Wilson 2001) for the benchmarks originally proposed by Cohen (1988), whereby *r* values around .10 are considered as small, around .25 as medium, and around of .40 as large effects. An additional point of consideration is the fact that, as is often observed in meta-analyses, several of the primary studies eligible for inclusion did not report the exact *r* values when these nominally were not significant (i.e., $p > .05$). Omitting the information represented by such studies (i.e., which did not report their non-significant outcomes numerically) from a meta-analysis would upwardly bias the obtained summary effects and, in turn, would lead to exaggerated conclusions regarding the magnitude of effects. For these reasons, this type of studies was accounted for in the meta-analytic models by conservatively assuming – in the absence of any further, more specific information – that all numerically unreported nonsignificant outcomes were zero, that is, inserting $r = 0$ for these cases (Lipsey and Wilson 2001). For each meta-analytic model, we state how many studies with $r = 0$ insertions were included. All meta-analytic calculations were performed in the Comprehensive Meta-Analysis (CMA) software, version 3 (Borenstein et al. 2013). Tests for publication bias are known to be underpowered with small meta-analytic datasets (Rothstein et al. 2005). Since this was the case for the current series of meta-analyses (comprising mostly less than 10 studies, and in one case merely 2 studies), and because we had to apply outcome substitutions in some cases with reporting deficiencies in the primary studies, we spared tests for publication bias, as well as any exploratory subgroup analyses (Borenstein 2019) in this set of meta-analyses.

Results

Results from the Systematic Literature Search

The initial literature search was concluded on March 05, 2020 and resulted in 231 studies with publication dates ranging from 1957 to 2020 (most recent study: Strohmaier et al. 2020). The majority of studies (223) were found with the search word “test anxiety”. The synonyms “exam anxiety” and “examination anxiety” only yielded eight additional relevant studies. We then screened these studies and excluded 75 studies not meeting the eligibility requirements of the first phase (e.g., 52 studies were not published in a peer-reviewed journal; see Figure 1 for these exclusion details).

As depicted in Figure 1, we evaluated the remaining 156 studies in the second phase and excluded 133 studies for not meeting the eligibility requirements of this phase. For example, we excluded 63 studies that did not use a physiological measure and 21 studies that did not assess test anxiety, but rather other constructs related to test anxiety, such as stress during public speaking or math anxiety. There was a total of 24 studies from 23 articles, with publication dates ranging from 1957 to 2020, that met the specified criteria and were included for review.

In the third phase, we screened the reference lists of these studies which led to the inclusion of six additional articles reporting seven studies. Thus, our systematic literature search yielded 29 articles with 31 studies. The selected studies and the study characteristics are listed in Tables 2 and 3.

Characteristics of the Selected Studies

Samples. Participants in the selected studies were primarily college or university students (25 out of 31 studies) and within these, most were psychology undergraduates (17 studies), who were for the most part from the United States. Only six studies included younger age groups (i.e., under 16 years old, e.g., elementary school or high school students). The mean age of participants across studies was $M = 21.3$ years ($SD = 6.3$; $Mdn = 21.9$; range: 7-43 years). However, in more than half of the studies (52 %), the exact age of the participants was not reported. The average sample size amounted to $N = 60.4$ ($SD = 35.4$, *Mdn*

= 56). However, it should be noted that there was broad range between $N = 6$ and $N = 171$ in terms of sample size across studies.

Study setting and design. Approximately half of the studies (i.e., 16 out of 31 studies) were laboratory studies and used an experimental test anxiety elicitation paradigm. These experimental test situations involved tasks like anagrams or other problem solving tasks (six studies), general aptitude tests (six studies), vocabulary tests (two studies), mathematics tests (one study), or psychology tests (one study). Most of the real-life studies (11 out of 15 studies) were conducted within the context of a psychology course exam for undergraduate students. Two studies were conducted in the context of a medical exam. Other real-life settings were a mock French exam and a nursing exam. None of the naturalistic studies assessed self-reported test anxiety during the test, instead it was assessed before or after the test or with a trait test anxiety inventory. However, one study used a video-based interview to retrospectively assess self-reported test anxiety that participants experienced during the test (Spangler et al. 2002; study 2). For more details about the study setting and design, see Tables S1 and S2 in the supplementary material. Most of the reviewed studies focused on test anxiety and performance outcomes (16 studies); other common variables of interest were coping styles (four studies) or general anxiety (five studies). One study looked at treatment effects (McGlynn et al. 1981).

Assessment of self-reported test anxiety. Most of the reviewed studies used trait self-report measures of test anxiety (21 out of 31 studies). Among these measures, the Test Anxiety Scale (TAS; Sarason 1978) or the child version of this scale (TASC; Sarason et al. 1958) were the most common (11 out of 31 studies). The second-most frequently employed measure was the Test Anxiety Questionnaire (TAQ; Mandler and Sarason 1952), which was used in five studies, whereas the Test Anxiety Inventory (TAI; Spielberger et al. 1980) was used in three studies. Two studies used other trait test anxiety questionnaires. From the studies that applied trait measures there was one study that used a combination of two trait measures

to assess test anxiety, and there were four studies in which additionally to the trait test anxiety measure the researchers applied a state measure. The remaining studies used state anxiety items before or after a test or exam (ten studies). For details, see Table S3 in the supplementary materials.

Assessment of physiological parameters. Multiple physiological measures were used in 14 out of 31 studies, but only 10 of these studies applied a combination of two measures that were considered for the current review. The other four studies used a combination of additional measures (not considered for the review), as described below. Cardiovascular measures were the most commonly used measures: 21 studies assessed heart rate or heart-rate variability and four studies assessed blood pressure. Eight studies applied electrodermal measures, six studies used cortisol sampling, and two studies assessed saliva pH. Additional measures employed, but not considered for the current review (e.g., as they are not regarded as a typical indicator of test anxiety or because they are no longer in use or were not successfully assessed in these studies), included respiration rate (three studies), FPV (finger-pulse volume), testosterone, prolactin, LH, skin and urinary pH, and sIgA (secretory immunoglobulin; one study each). The majority of studies used well-established devices to assess physiological arousal; however, in four of the reviewed studies participants counted their own heart beats.

Strategies for comparing self-report with physiological measures. In 14 studies, samples were dichotomously split into high versus low test anxious groups, based on a general index of self-reported test anxiety, as indicated by their sum score on a test anxiety scale. Thus, these studies only compared physiological arousal between high and low test anxiety groups. Thirteen studies used a continuous assessment of test anxiety and correlated the scores on the test anxiety scales with physiological arousal, without categorization of participants. Four studies used both categorical and continuous measures.

Twelve studies differentiated between the worry and emotionality components of test anxiety, but only nine of these studies reported separate results on the components' correlation with physiological measures. The other three studies only compared total scores of test anxiety with physiological measures.

Meta-Analysis of the Correlations between Self-Report Measures and Physiological Measures

Results revealed that 21 studies found significant positive correlations between self-reported test anxiety and at least one measure of physiological arousal. Nine of the reviewed studies did not find significant correlations, and in one study low anxious students were found to be more aroused than high anxious students. We noted that the different strengths of correlations can be explained by different aspects of the reviewed studies like the study setting or the sample size. We will address these aspects and their proposed effect on the convergence between the two measures in the discussion section. In the next paragraphs we will report the meta-analytic results of the correlation coefficients for the different physiological measures which is followed by the descriptive and meta-analytic results for the worry and emotionality components of test anxiety.

Meta-analysis of the correlation coefficients for different physiological measures. Heart rate and heart rate variability (k = 20 studies, of which 6 studies did not report nominally nonsignificant associations, for which zero correlations were therefore inserted; total N = 1199) was significantly positively associated (z = 8.47, p < .001) with test anxiety, yielding a medium-sized correlation (r = .246, 95% CI = .191-.300). Effect heterogeneity across studies was low ($Q(df = 19) = 28.9$, p = .07, I^2 = 34.3%).

Skin conductance response (k = 8, with zero correlations inserted for 3 studies, N = 418) was significantly positively associated (z = 3.95, p < .001) with test anxiety, yielding a medium-sized correlation (r = .196, 95% CI = .100-.289), along with a medium-sized effect heterogeneity across studies ($Q(7) = 13.4$, p = .06, I^2 = 47.1%).

Salivary pH level ($k = 2$, $N = 100$) was significantly positively associated ($z = 3.75$, $p < .001$) with test anxiety, yielding a large effect ($r = .468$, 95% $CI = .182$-$.529$). Cross-study effect heterogeneity was negligible ($Q(1) = .40$, $p = .53$, $I^2 = 0\%$).

Systolic blood pressure ($k = 3$, zero correlation inserted for one study; $N = 225$) was significantly positively associated ($z = 2.89$, $p = .004$) with test anxiety, representing a medium-sized effect ($r = .194$, 95% $CI = .063$-$.318$). Cross-study effect heterogeneity was of medium size, but not nominally significant ($Q(2) = 4.27$, $p = .12$, $I^2 = 53.1\%$).

Diastolic blood pressure ($k = 2$, with a zero correlation inserted for one study, $N = 126$) was positively associated with test anxiety, yielding a medium-sized summary effect ($r = .242$, 95% $CI = .067$-$.401$). Effect heterogeneity for this set of 2 studies was nominally significant and large ($Q(1) = 7.0$, $p = .008$, $I^2 = 85.8\%$).

Cortisol sampling ($k = 6$, $N = 324$) was significantly positively associated ($z = 2.00$, $p = .045$) with test anxiety, with the summary effect being small ($r = .114$, 95% $CI = .002$-$.223$). Effect heterogeneity was nominally not significant and medium-sized ($Q(5) = 10.4$, $p = .06$), $I^2 = 51.1\%$).

Meta-analysis of the correlation coefficients for worry and emotionality. Regarding the correlation coefficients for the different test anxiety components, we found that in total, nine of the 31 reviewed studies differentiated between the worry and emotionality components of test anxiety and reported differential results for the two components. Only one study (Herbert et al. 1986) did not find significant results between either worry and serum cortisol or emotionality and serum cortisol. From the remaining eight studies (study 1 and 2 by Conley and Lehman 2012; Deffenbacher 1986; Deffenbacher and Hazaleus 1985; study 1 and 2 by Morris and Liebert 1970; Cohen and Khalaila 2014; Endler and Magnusson 1977; Kantor et al. 2001), six studies found higher correlations between the emotionality component of test anxiety and physiological measures. Only two studies found higher correlations between the worry component and the physiological measure. However, in most of these

studies it was not reported if the correlations for worry and emotionality were significantly different from each other. The exact values of the correlations can be seen in Table 4.

Results from the meta-analysis revealed that, when summarizing across the physiological measures (k = 9, zero correlation inserted for one study; N = 476), the emotionality component of test anxiety was significantly positively associated with these (z = 6.08, p < .001), representing a medium-sized effect (r = .220, 95% CI = .151-.288). Cross-study effect heterogeneity was significant and medium-sized ($Q(8)$ = 16.1, p = .04, I^2 = 50.2%).

Similarly, summarizing across the various physiological measures (k = 9, 2 zero correlations inserted; N = 476), the worry component of test anxiety was significantly positively associated with these (z = 5.15, p < .001), likewise representing a medium-sized effect (r = .187, 95% CI = .117-.256). Once again, cross-study effect heterogeneity was significant and of medium size ($Q(8)$ = 20.4, p = .009, I^2 = 60-7%).

Direct comparison of the confidence intervals accompanying these two latter meta-analytic summary effects (95% CI: .151-.288 vs. 117-.256 for the emotionality vs. worry components of test anxiety, respectively) showed an almost complete overlap between these, thus suggesting that both components of test anxiety were correlated with the various physiological measures to the same extent.

Discussion

Summary of the Systematic Review and Meta-Analysis

The present investigation intended to synthesize past research findings on the relationship between self-report measures of test anxiety and physiological measures, in order to advance physiological methods in educational research and to provide a deeper understanding of this widespread emotion. In the following sections, we discuss the results of our systematic review, along with the meta-analytic quantifications, and provide implications

for future research (see also Figure A1 in the supplementary materials for a summary of the implications).

Relationships between Self-Reported Test Anxiety and Objectively Measurable Physiological Arousal

As expected, the majority of the reviewed studies (i.e., 21 studies, 68 %) found significant positive correlations between self-reported test anxiety and at least one measure of physiological arousal. Thus, the results of the present review are generally consistent with theoretical concepts and support the assumption that higher self-reported test anxiety is associated with higher physiological arousal. The size and range of the observed significant correlations indicate that the two types of measures do not completely overlap. This is in line with theoretical assumptions, as physiological measures are assumed merely to tap into the physiological component of test anxiety, whereas self-report assessments usually represent all test-anxiety components, as these are subjectively experienced (i.e., cognitive, affective, motivational, and behavioral components as well; Pekrun et al. 2011; Wilhelm and Roth 2001). From the nominally significant, but generally medium-sized, relationships observed between self-report and physiological measures, one could further conclude that using physiological in addition to self-report measures is highly relevant, because this might add information about the physiological component of test anxiety not accessible with self-report alone (Wilhelm and Roth 2001). Thus, this result also provides support for the view that in order to get the most comprehensive assessment of test anxiety, different kinds of test anxiety measures should be used (e.g., Kantor et al. 2001; Zeidner and Matthews 2011; Allen et al. 1980).

Aspects That Are Potential Moderators of the Relationship

As our results suggest, there is a substantial relationship between self-reported test anxiety and physiological arousal in the majority of the reviewed studies, we now turn to factors that might influence this relationship. We begin by discussing factors from the studies

with nonsignificant or contradictory results that might explain this apparent divergence. We then suggest implications and considerations for future research. This is followed by a detailed discussion of the studies with significant results.

Studies with nonsignificant or contradictory results. Nine of the reviewed studies (29%) did not find significant correlations between self-reported test anxiety and objectively measured physiological arousal. In addition, in one study, low anxious students (as classified by self-report measures) were found to be more physiologically aroused than high-anxious students. Such contradictory or nonsignificant findings among the reviewed studies provide insights into factors that might impact expected relations, as elaborated on in the following.

Study setting. We noted that more than one half of the studies with nonsignificant or contradictory results were conducted within a lab setting (six out of ten studies). These laboratory experimental approaches have distinct advantages. For example, because of the constrained lab environment, physiological measurements contain relatively few artifacts from participants' movement, or technical problems (Houtveen and de Geus 2009). Moreover, conducting a study in a lab avoids difficulties associated with the lack of portable, sophisticated equipment and additionally allows for precise control over the task (Wilhelm and Grossman 2010). However, a major disadvantage is that laboratory studies are often conducted under artificial conditions, lack ecological validity, and therefore often also lack personal relevance for the participants (Houtveen and de Geus 2009). Furthermore, it might be the case that individuals simply are excited about the lab or the experiment in general, whereas not necessarily test anxious. In this case, increased physiological arousal in laboratory settings could also be interpreted as an indicator of activation due to positive interest (i.e., approach motivation) in the task, or as a positive interest or excitement cued off by the complex experimental setup that participants face in a lab (Schlosberg 1954).

Thus, elevated physiological arousal may merely be an artifact of the laboratory situation and not necessarily an indicator of anxiety. This is also supported by a review of

Johnston et al. (1990), who compared cardiovascular responses in the laboratory to those under natural conditions and found that approximately half of the reviewed studies reported contradictory findings. Zanstra and Johnston (2011) published a review on the extent to which responses to laboratory and natural realistic stressors are comparable. These researchers state that responses obtained in real life often are larger than those obtained in the laboratory and that subjective ratings of stress, emotion, and cognitive determinants in real life also relate to real- life cardiovascular responses. This might also have been the case in the experimental laboratory study conducted in the context of a low stakes test (Strohmaier et al. 2020). In this study, participants reported low levels of test anxiety which might account for the low convergence with electrodermal activity ($r = 0.06$) and calls for determining the relation between self-reported test anxiety and physiological arousal under *more naturalistic conditions* (i.e., real-life assessment).

Sample characteristics. The four naturalistic studies with nonsignificant results also showed methodological problems that may well partly account for the nonsignificant findings. Two of the four studies had limited sample size. One study used a mock exam and had a relatively *small* ($N = 23$) *sample size* and a mostly female sample. The authors (Daly et al. 2011) stated that their work can be seen more like a pilot study, that the results need to be tested in a larger sample under more naturalistic conditions, and that their *sample (high performers) might have been generally less anxious* about taking a test. The second real-life study (Herbert et al. 1986) looked at the cortisol response of male medicine students before a major medicine exam. This is also a very *special sample*, as medical students from the UK are accustomed to taking such exams and are likely high-achieving students. Moreover, because participants *were all male*, the nonsignificant findings of this study do not necessarily generalize to other samples and female participants since there are various studies suggesting that there are sex differences in the cortisol response (Kudielka et al. 2000; Kajantie and Phillips 2006).

The lab study of Hollandsworth et al. (1979) had the smallest sample size ($N = 6$). This study found the contradictory result of low test anxious students being more physiologically aroused (i.e., higher HR) than high anxious students. The very small size of the sample reduces the generalizability of these results.

Assessment and analysis of physiological data. The third naturalistic study with nonsignificant results (Huwe et al. 1998) had a larger sample size ($N = 58$), but also showed major methodological problems, as heart rate was counted manually by the experimenter (i.e., not with a *dedicated device*). However, this study found that heart rate and cortisol response at least tended to be higher in high test anxious students than in low test anxious students, although this difference did not reach nominal significance.

The fourth naturalistic study (Ringeisen et al. 2018) reported correlations between the intercept of the cortisol response and the intercept of the anxiety response. The authors mentioned several methodological aspects regarding the assessment and analysis of physiological data that might have affected the association patterns such as the *intervals of assessment of cortisol and self-report data* and the *level of aggregation across different measurement points*.

Another methodological issue of one of the studies with nonsignificant results (Glazeski et al. 1986) was the use of a *single arousal index*, calculated from three different measures (heart rate, skin conductance and respiration) and averaged across the whole test situation. Specifically, Glazeski et al. (1986) used a single score of arousal for each participant which they computed from mean scores on three different measures administered during test taking. Later, based on these single scores, they additionally subdivided (more precisely, trichotomized) their sample into groups with high vs. moderate vs. low arousal. They did not find any differences between high and low test anxiety participants in arousal group membership ($N = 15$ per group). By averaging and merging the different physiological measures, the authors might have simply overseen the relations between different

physiological parameters and test anxiety. As mentioned, different physiological measures have *different temporal resolutions,* and there might be systematic differences between parameters representing different physiological systems. Therefore, it is problematic to combine them at face value (Martin 1961). This shows that the selection of physiological measures should be substantiated and justified, and that different sampling rates (i.e. temporal resolutions) and links to different physiological systems should be taken into account when utilizing more than one measure.

Studies with significant results. From the studies with significant results we also found some interesting aspects that might have had an influence on the correlations and that might explain the relatively high variance of the correlations coefficients and thereby help to guide the development of future studies. These aspects concern mostly the *assessment and analysis of physiological data*.

Sophisticated devices. The manner in which physiological arousal was assessed seemed to play an important role for the magnitude of convergence in the studies with significant findings as well. Correlations were lower or, as mentioned above, negligible, like in the study of Huwe et al. (1998), when physiological assessment was poor (e.g., when heart rate was measured manually). For example, in four studies (Deffenbacher & Hazaleus 1985; Deffenbacher 1986; Morris & Liebert 1970, Studies 1 and 2) participants had to count their own heart beats, which yielded lower correlations (HR mean $r = 0.24$; range: 0.19 to 0.30), as compared with the results from more accurate devices like heart rate monitors (HR mean $r = 0.39$; range: 0.26 to 0.51). As also noted by the authors of one of these studies (Morris and Liebert 1970), a main reason is that participants or even experimenters often fail to obtain a satisfactory count, which leads to missing or unreliable data. This implies that it is highly important to use technologically adequate devices, rather than manual counting.

Detailed assessment of physiological data. Another explanation for the high variance across correlation coefficients might be that some studies included two assessments, one

during baseline (i.e., as a control condition) and a second during stress conditions (e.g., Deffenbacher and Hazaleus 1985; Deffenbacher 1986; Morris and Liebert 1970), or reported a single mean score to reflect the level of arousal over the entire session. Other studies included *multiple assessments* (i.e., went more into detail) and investigated parameters under baseline, anticipation, stress, and recovery, or even looked at the variability of changes in physiological arousal within these phases. Correlations were generally larger with these more fine-grained assessments (averaged across all physiological parameters: $r = .22$ vs $r = .35$). For example, Montgomery (1977) focused on the anticipatory cardiac response and took a detailed look at the waveforms. He found relatively high correlations between heart rate and test anxiety, as well as a greater heart rate slowing in low test anxious subjects. Raphelson (1957) first found that low test anxious subjects initially (in their null period, i.e., baseline) had higher skin conductance levels than high test anxious, perhaps indicating that factors other than anxiety might have played a role. However, when looking at the variation during the actual task, the skin conductance of high test anxious participants increased and of low test anxious participants decreased during the task. Beidel (1988) looked at such variations during a test, using heart rate, and Spangler et al. (2002) used cortisol sampling, and in these studies similar response patterns were observed.

Harleston et al. (1965) also found that the low test anxious group showed a slight drop in heart rate during a test, whilst the high test anxious group sustained the highest increase in heart rate during the test. Furthermore, these researchers found that the relation between anxiety and arousal was significant only during the test when working on the task. If these researchers had solely looked at group differences of absolute values at a single point in time (e.g., at baseline) without inspecting change over time, they would have missed important information and may not have found any substantial association between self-report and physiological measures. Significant correlations seemed to emerge more often when participants interacted with the task (i.e., during the test). Furthermore, it seems to be

important to consider within-person associations between self-report and physiological measures across time (i.e., cross-temporally, such as in Spangler et al. 2002), rather than exclusively considering between-person associations at a single time point (Mauss and Robinson 2009).

Influencing factors that should be controlled. Additionally, the associations can be affected by variables like previous food intake or drug use (such as alcohol or smoking). Conley and Lehman (2012) included several covariates (i.e., smoking, activity level, food, alcohol and caffeine) in their analysis. It was found that these factors influenced results in different ways. Specifically, cigarette use and more strenuous physical activity during ambulatory assessment led to elevated heart rate, diastolic blood pressure, and systolic blood pressure; food consumption led to elevated systolic blood pressure and heart rate; alcohol consumption led to elevated heart rate; whereas caffeine had no influence. To control or adjust for these variables is especially relevant for studies that use cortisol or other saliva samples, as the results of these measures can be strongly biased by such factors (Hansen et al. 2008; Pollard 1995). This can be seen clearly in the study of Cohen and Khalaila (2014), which only found associations in the predicted directions for emotionality and pH, when these were controlled for the degree of physical activity per week and habitual smoking.

Specific characteristics of different physiological measures. We also found differences in the relations between reported test anxiety and the different physiological measures. Saliva pH (r = .468), cardiovascular (HR and HRV: r = .246; SBP: r = .194; DBP: r = .242), and electrodermal measures (r = .196) generally yielded stronger correlations than cortisol sampling (r = .114). This can be explained by the fact that some of these measures may be more specific to emotional arousal, like electrodermal activity, which is known to be solely controlled by the sympathetic nervous system (e.g., Setz et al. 2010). On the other hand, some measures do not only increase during sympathetic activation. Especially cortisol follows a diurnal rhythm and shows increased values after awakening which is hard to

disentangle from the anxiety reaction, furthermore it cannot be sampled continuously. When measuring heart rate it is also important to keep in mind that heart rate can also increase when cognitive load is high or during general activation or physical activity (Myrtek et al. 1990; Saito and Nakamura 1995). For example, McGlynn et al. (1981) found a correlation between test anxiety and skin conductance, but no correlation with heart rate. This shows that it is important to keep in mind that the different measures might provide different indices of arousal and are not necessarily correlated with each other. Moreover, they may also vary regarding the strength of their associations with emotions (Kreibig 2010).

Appropriate data analysis. A final important point is the way in which the data are analyzed. As mentioned earlier, there might be intraindividual (i.e., within-person) relationships between physiological and self-report measures that cannot be accounted for when using a between-persons approach and only considering interindividual differences. Conley and Lehman (2012) accounted for these intraindividual processes by using hierarchical linear regression modelling, based on multiple blood pressure readings for each participant (i.e., utilizing a high sampling rate), along with self-reports at each blood pressure reading. Future research could benefit from applying more regularly such intraindividual approaches.

Worry and Emotionality

The third goal of this systematic review and meta-analysis was to identify whether the relationship between the emotionality component of test anxiety and physiological arousal differs from the corresponding relationship for the worry component of test anxiety. Although valid instruments (such as the TAI) that include items for both components were used in most of the studies reviewed,, there were only nine studies that differentiated between the two components and reported results for their relations with physiological indicators. Only one of these studies (Herbert et al. 1986) did not find significant relations between either worry and serum cortisol, nor emotionality and serum cortisol. From the remaining eight studies, six

studies found higher correlations between the emotional component and physiological measures than for the worry component. Of the remaining two studies, one study (Conley and Lehman 2012) found that only worry predicted elevations in SBP. The other study (Endler and Magnusson 1977) also found higher correlations for worry and HR, but the correlations for the emotionality component also reached nominal significance. However, the majority of these studies did not report whether the differences between the correlations were significant.

Although at the first glance these descriptive findings suggest that the emotionality component is more strongly associated with physiological arousal than the worry component, our meta-analytic findings tell a different story. The direct comparison of the confidence intervals for the meta-analytic effects shows that there is no significant difference between the relations for worry and emotionality (95% *CI*: .151-.288 vs. 117-.256 for the emotionality vs. worry components of test anxiety, respectively). This is in line with findings from Liebert and Morris (1967) who tested the differences in correlations for significance and found that the correlations between emotionality and heart rate and worry and heart rate did not significantly differ. They explain this due to the fact that the correlations they found were much smaller than expected, which might have resulted from poor physiological assessment in their study (i.e., participants had to count their own pulse rates). Another reason for the lack of a significant difference is that due to the limited number of studies that distinguished between worry and emotionality, we could not consider the correlation coefficients from the different physiological parameters separately and rather had to calculate an average across parameters (i.e. heart rate, systolic blood pressure, and saliva pH). Since these parameters represent different physiological systems, there may be systematic differences which might have influenced our results.

Finally, the theoretical conceptualization of test anxiety components could have played a role where as well: The distinction between worry and emotionality may not be sufficiently fine grained. More recent research distinguishes between more components of test

anxiety. A common distinction of the components, which is also reflected in emotion questionnaires such as the Achievement Emotions Questionnaire (AEQ; Pekrun et al. 2011), is the differentiation between five anxiety components (Pekrun et al. 2004; Scherer 2009): cognitive, affective, motivational, physiological, and expressive components. With such a more fine-grained differentiation we might find stronger associations between the physiological component of self-reported test anxiety and physiological arousal since the items that assess this component more strongly tap into aspects of physiological arousal (i.e. they directly ask about a higher heart rate or sweating) than the items for emotionality for broadly, which typically assess the experience of both affective and physiological facets of anxiety. However, for such an analysis we would not expect more than medium-sized correlation coefficients either, as perceptions of physiological arousal (i.e., as already partly reflected in emotionality items and more strongly reflected in items that assess the physiological component) are not necessarily strongly correlated with actual autonomic reactivity assessed with objective physiological measures (Hodges 2015). By implication, physiological measures can provide additional information (especially about the physiological test anxiety component) that cannot be assessed with self-reports alone.

Another important point worth mentioning is that surprisingly, all eight studies with significant results that differentiated between the worry and emotionality components only used cardiovascular measures (six studies used heart rate, one study used blood pressure measurements) and saliva pH (one study) and we identified a lack of research on electrodermal activity and cortisol and their relation to the worry and emotionality components. Future work should close this obvious research gap.

Strengths and Limitations of the Current Systematic Review and Meta-Analysis

Due to the heterogeneity of the included studies, we chose to perform a qualitatively based systematic review in the first place, with a series of meta-analytic models supplementing this approach. The advantage such a literature review is that a broader and

more complete picture can be gained, because the review is more inclusive, better documented and more exhaustive than traditional, narrative (i.e. unsystematic) reviews are. In similar vein, a formal meta-analysis on a genuinely heterogeneous landscape of research evidence may not entail the qualitative finesse a systematic review can offer in such constellations. Pursuing this approach, we could include samples with different demographic backgrounds (e.g., different age groups) and physiological and self-report measures without a restrictive focus on only a few comparable dimensions, such as only one specific sample or physiological parameter. This was done to provide direction for generating hypotheses and future directions for applying physiological measures in the field of educational psychology.

There are several limitations of the present account which should be considered when interpreting its findings and which also can be used to derive novel directions for future inquiry along these lines. One limitation is that the research discussed and meta-analytically quantified in this review was obtained from articles published in English and in peer-reviewed journal outlets. Therefore, this review does not include findings and information provided by unpublished studies (e.g., unpublished dissertations and other types of grey literatures) or articles published in languages other than English. Future work could take this limitation into account. Another point is that the number of studies in our review was limited, because we did not include studies that investigated constructs supposedly similar to test anxiety (e.g., exam stress or social evaluative stress). This was done because stress is a broader construct that can include anxiety, but does not always do so (e.g., positive stress: Putwain 2007; Selye 1976a; Jamieson et al. 2012; Selye 1976b). Therefore, we only included studies of which we were certain that they indeed did assess test anxiety proper (i.e., with a dedicated test anxiety questionnaire).

Conclusion, Educational Significance, and Future Directions for Research

In sum, we found mostly positive associations between self-reported and physiological measures of test anxiety. The results of this review are promising and in line with theoretical

assumptions. Knowing that self-report measures can be strongly biased by subjective beliefs (Robinson and Clore 2002) and are more likely to capture the affective or cognitive components of anxiety, as they assess subjective experiences represented in the conscious mind (Pekrun and Bühner 2014), the results of our review suggest that physiological measures can provide objective information about test anxiety, particularly about the physiological component of test anxiety which is difficult to capture through self-reports. For research in educational psychology, this implies that is important to use both types of measures in order to capture the complex construct of test anxiety as objective and detailed as possible.

In most of the studies reviewed, the samples consisted of students (25 out of 31 studies) and were primarily undergraduate psychology students (in 17 studies). This is a general problem of studies in psychology and also characterizes the specific research field and topic we reviewed here. The narrow sampling frame clearly limits the conclusions and inferences that can be drawn for educational research in general (Gordon et al. 1986; Shen et al. 2011). This calls for applying these measures in classroom-based research, in order to obtain results that are not only specific to a psychology undergraduate pool and to have high ecological validity in addition. By continuously assessing students' physiological arousal (for example, during exams), we might be able to expand our knowledge of the mechanisms underlying test anxiety in ways not possible using self-report measures alone. This more ecologically valid, more objective, and detailed assessment might in the long run also help to improve interventions targeting test anxiety (e.g., cognitive test anxiety treatment and relaxation techniques).

Moreover, as there are studies that question the debilitative aspects of physiological arousal in testing situations (Glazeski et al. 1986), by using both types of measures it would be possible to identify whether it is the interpretation of physiological arousal (i.e., self-reported physiological test anxiety component) or the arousal, as assessed by objective physiological measures, that influences student performance, and further, to investigate which

direction this influence takes (Mandler and Kremen 1958). As this review also shows that psychophysiological results seem to be influenced by different aspects of the methodology used, future research could statistically consider these influencing variables as moderators in their research designs. Furthermore, when applying physiological measures, the study design, setting, sample characteristics, data assessment, and data analysis should be carefully selected.

Finally, as our results suggest that each of the two measures (i.e., self-report and physiological measures) provides non-redundant information about test anxiety of individuals and much of the information provided by each measure might be specific to a particular component of test anxiety, in future research, besides physiological measures, studies should also include additional measures of test anxiety, such as behavioral measures, to get a more comprehensive assessment and a more complete picture of the different components of this widespread and mostly detrimental emotion (Larsen and Prizmic-Larsen 2006; Zeidner and Matthews 2011).

References

Allen, G., Elias, M., & Zlotlow, S. (1980). Behavioral interventions for alleviating test anxiety: A methodological overview of current therapeutic practices. In I. G. Sarason (Ed.), *Test anxiety: Theory, research and applications* (pp. 155–185). Hillside, NJ: Erlbaum.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology, 61*(2), 207-215. doi:10.1037/h0045464

Avram, R., Kuhar, P., Vittinghoff, E., Aschbacher, K., Tison, G., Pletcher, M., et al. (2018). Redefining normal resting heart rate values using Big Data. *Circulation, 138*(Suppl_1), A15098-A15098. doi:10.1161/circ.138.suppl_1.15098

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Review, 6*(4), 292-297.

Beidel, D. C. (1988). Psychophysiological assessment of anxious emotional states in children. *Journal of Abnormal Psychology, 97*(1), 80-82. doi:10.1037/0021-843X.97.1.80

Beidel, D. C., & Turner, S. M. (1988). Comorbidity of test anxiety and other anxiety disorders in children. *Journal of Abnormal Child Psychology, 16*(3), 275-287. doi:10.1007/BF00913800

Berntson, G. G., Norman, G. J., Hawkley, L. C., & Cacioppo, J. T. (2008). Cardiac autonomic balance versus cardiac regulatory capacity. *Psychophysiology, 45*(4), 643-652. doi:10.1111/j.1469-8986.2008.00652.x

Bigger, J. T., Fleiss, J. L., Steinman, R. C., Rolnitzky, L. M., Kleiger, R. E., & Rottman, J. N. (1992). Frequency domain measures of heart period variability and mortality after myocardial infarction. *Circulation, 85*(1), 164-171, doi:10.1161/circ.85.1.1728446.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive Meta-Analysis*, version 3 [computer software]. Englewood, NJ: Biostat.

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Englewood, NJ: Biostat.

Boucsein, W. (2012). *Electrodermal activity* (2ed.). New York: Springer Science & Business Media.

Bouma, E. M., Riese, H., Ormel, J., Verhulst, F. C., & Oldehinkel, A. J. (2009). Adolescents' cortisol responses to awakening and social stress; effects of gender, menstrual phase and oral contraceptives. The TRAILS study. *Psychoneuroendocrinology, 34*(6), 884-893.

Buehler, R., & McFarland, C. (2001). Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin, 27*(11), 1480-1493. doi:10.1177/01461672012711009

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. In M. Lewis, & J. Haviland-Jones (Eds.), *The handbook of emotions* (pp. 173-191). New York: Guildford Press.

Calvo, M. G., & Miguel-Tobal, J. J. (1998). The anxiety response: Concordance among components. *Motivation and Emotion, 22*(3), 211-230. doi:10.1023/A:1022384022641

Campbell, S. B. (1986). Developmental issues in childhood anxiety. In R. Gittelman (Ed.), *Anxiety Disorders of Childhood* (pp. 24-57). New York: Guilford.

Carlson, N. (2013). *Physiology of behavior*. New Jersey: Pearson Education, Inc.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270-295. doi:10.1006/ceps.2001.1094

Chorot, P., & Sandin, B. (1985). The Anxiety State Bahavioral-Scale *PSIQUIS, 6*(3), 60-65.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ed.). Hillsdale, New Jersey: Erlbaum.

Cohen, M., & Khalaila, R. (2014). Saliva pH as a biomarker of exam stress and a predictor of exam performance. *Journal of Psychosomatic Research, 77*(5), 420-425. doi:10.1016/j.jpsychores.2014.07.003

Conley, K. M., & Lehman, B. J. (2012). Test anxiety and cardiovascular responses to daily academic stressors. *Stress and Health, 28*(1), 41-50. doi:10.1002/smi.1399

Curran, J. P., & Cattell, R. B. (1976). *Manual for the eight state questionnaire: 8SQ*. Champaign: Institute for Personality and Ability Testing.

D'Mello, S., & Calvo, R. A. Beyond the basic emotions: what should affective computing compute? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems, 2013* (pp. 2287-2294): ACM

D'Mello, S. K., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review, 10*(2), 174-183.

Daly, A. L., Chamberlain, S., & Spalding, V. (2011). Test anxiety, heart rate and performance in A-level French speaking mock exams: An exploratory study. *Educational Research, 53*(3), 321-330. doi:10.1080/00131881.2011.598660

Deffenbacher, J. L. (1986). Cognitive and physiological components of test anxiety in real-life exams. *Cognitive Therapy and Research, 10*(6), 635-644. doi:10.1007/BF01173751

Deffenbacher, J. L., & Hazaleus, S. L. (1985). Cognitive, emotional, and physiological components of test anxiety. *Cognitive Therapy and Research, 9*(2), 169-180. doi:10.1007/BF01204848

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin, 130*(3), 355-391. doi:10.1037/0033-2909.130.3.355

Dorn, L. D., Lucke, J. F., Loucks, T. L., & Berga, S. L. (2007). Salivary cortisol reflects serum cortisol: Analysis of circadian profiles. *Annals of Clinical Biochemistry, 44*(3), 281-284. doi:10.1258/000456307780480954

Endler, N. S., Edwards, J. M., & Vitelli, R. (1991). *Endler multidimensional anxiety scales (EMAS)*. CA, Los Angeles: Western Psychological Services

Endler, N. S., & Magnusson, D. (1977). The interaction model of anxiety: An empirical test in an examination situation. *Canadian Journal of Behavioural Science, 9*(2), 101-107. doi:10.1037/h0081612

Endler, N. S., & Okada, M. (1975). A multidimensional measure of trait anxiety: The SR Inventory of general trait anxiousness. *Journal of Consulting and Clinical Psychology, 43*(3), 319-329.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45. doi:10.1037/0022-3514.65.1.45

Glazeski, R. C., Hollandsworth, J. G., & Jones, G. E. (1986). An investigation of the role of physiological arousal in test anxiety. *Educational & Psychological Research, 6*(2).

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science, 24*(10), 2079-2087.

Gordon, M. E., Slade, L. A., & Schmitt, N. (1986). The "science of the sophomore" revisited: From conjecture to empiricism. *Academy of Management Review, 11*(1), 191-207 doi:10.5465/amr.1986.4282666

Hansen, Å. M., Garde, A. H., & Persson, R. (2008). Sources of biological and methodological variation in salivary cortisol and their impact on measurement among healthy adults: A review. *Scandinavian Journal of Clinical and Laboratory Investigation, 68*(6), 448-458. doi:10.1080/00365510701819127

Harleston, B. W. (1962). Test anxiety and performance in problem-solving situations. *Journal of Personality, 30*(4), 557-573. doi:10.1111/j.1467-6494.1962.tb01689.x

Harleston, B. W., Smith, M. G., & Arey, D. (1965). Test-anxiety level, heart rate, and anagram problem solving. *Journal of Personality and Social Psychology, 1*(6), 551. doi:10.1037/h0021991

Harley, J. M. (2015). Measuring emotions: A survey of cutting-edge methodologies used in computer-based learning environment research. In S. Tettegah, & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 89 - 114). London, UK: Academic Press, Elsevier.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77. doi:10.2307/1170348

Herbert, J., Moore, G., De La Riva, C., & Watts, F. (1986). Endocrine responses and examination anxiety. *Biological Psychology, 22*(3), 215-226. doi:10.1016/0301-0511(86)90027-X

Hodges, W. (2015). The psychophysiology of anxiety. In M. Zuckerman, & C. D. Spielberger (Eds.), *Emotions and Anxiety (PLE: Emotion): New Concepts, Methods, and Applications* (pp. 175-192): Psychology Press.

Hollandsworth, J. G., Glazeski, R. C., Kirkland, K., Jones, G. E., & Van Norman, L. R. (1979). An analysis of the nature and effects of test anxiety: Cognitive, behavioral, and physiological components. *Cognitive Therapy and Research, 3*(2), 165-180. doi:10.1007/BF01172603

Houtveen, J. H., & de Geus, E. J. (2009). Noninvasive psychophysiological ambulatory recordings: Study design and data analysis strategies. *European Psychologist, 14*(2), 132-141. doi:10.1027/1016-9040.14.2.132

Hugdahl, K. (1995). *Psychophysiology: The mind-body perspective*: Cambridge, MA: Harvard University Press.

Humphrey, S. P., & Williamson, R. T. (2001). A review of saliva: Normal composition, flow, and function. *The Journal of Prosthetic Dentistry, 85*(2), 162-169. doi:10.1067/mpr.2001.113778

Huwe, S., Hennig, J., & Netter, P. (1998). Biological, emotional, behavioral, and coping reactions to examination stress in high and low state anxious subjects. *Anxiety, Stress & Coping, 11*(1), 47-65. doi:10.1080/10615809808249313

Jamieson, J. P., Nock, M. K., & Mendes, W. B. (2012). Mind over matter: Reappraising arousal improves cardiovascular and cognitive responses to stress. *Journal of Experimental Psychology. General, 141*(3), 417-422. doi:10.1037/a0025719

Janke, W., Debus, G., & (1978). *Die Eigenschaftswörterliste: EWL; Eine mehrdimensionale Methode zur Beschreibung von Aspekten des Befindens.* Goettingen: Hogrefe.

Johnston, D. W., Anastasiades, P., & Wood, C. (1990). The relationship between cardiovascular responses in the laboratory and in the field. *Psychophysiology, 27*(1), 34-44. doi:10.1111/j.1469-8986.1990.tb02175.x

Kajantie, E., & Phillips, D. I. (2006). The effects of sex and hormonal status on the physiological response to acute psychosocial stress. *Psychoneuroendocrinology, 31*(2), 151-178.

Kantor, L., Endler, N. S., Heslegrave, R. J., & Kocovski, N. L. (2001). Validating self-report measures of state and trait anxiety against a physiological measure. *Current Psychology, 20*(3), 207-215. doi:10.1007/s12144-001-1007-2

Kaplan, S., Dalal, R. S., & Luchman, J. N. (2013). Measurement of emotions. In *Research methods in occupational health psychology: Measurement, design, and data analysis* (pp. 61-75). New York, NY: Routledge.

King, N. J., Ollendick, T. H., & Prins, P. J. (2000). Test-anxious children and adolescents: Psychopathology, cognition, and psychophysiological reactivity. *Behaviour Change, 17*(3), 134-142. doi:10.1375/bech.17.3.134

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*, 345–379.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology, 84*(3), 394-421 .:10.1016/j.biopsycho.2010.03.010

Kudielka, B. M., Hellhammer, D. H., & Kirschbaum, C. (2000). Sex differences in human stress response. In *Stress Consequences: Mental, Neuropsychological and Socioeconomic* (Vol. 3). San Diego, CA: Academic Press.

Larsen, R. J., & Prizmic-Larsen, Z. (2006). Measuring emotions: Implications of a multimethod perspective. In M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 337-351). Washington, DC: American Psychological Association.

Lench, H. C., Bench, S. W., & Flores, S. A. (2013). Searching for evidence, not a war: Reply to Lindquist, Siegel, Quigley, and Barrett (2013). *Psychological Bulletin, 113*(1), 264-268.

Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin, 137*, 834–855. doi:10.1037/a0024244

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*(3), 975-978. doi:10.2466/pr0.1967.20.3.975

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The Hundred-Year Emotion War: Are Emotions Natural Kinds or Psychological Constructions? Comment on Lench, Flores, and Bench (2011). *Psychological Bulletin, 139*(1), 264-268.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Liu, J. J. W., Ein, N., Peck, K., Huang, V., Pruessner, J. C., & Vickers, K. (2017). Sex differences in salivary cortisol reactivity to the Trier Social Stress Test (TSST): A

meta-analysis. *Psychoneuroendocrinology, 82*, 26-37. doi:https://doi.org/10.1016/j.psyneuen.2017.04.007

Lundberg, U. (2005). Stress hormones in health and illness: The roles of work and gender. *Psychoneuroendocrinology, 30*(10), 1017-1021. doi:10.1016/j.psyneuen.2005.03.014

Mandler, G., & Kremen, I. (1958). Autonomic feedback: A correlational study. *Journal of Personality, 26*(3), 388-399. doi:10.1111/j.1467-6494.1958.tb01594.x

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *The Journal of Abnormal and Social Psychology, 47*(2), 166-173.

Marques, A. H., Silverman, M. N., & Sternberg, E. M. (2010). Evaluation of stress systems by applying noninvasive methodologies: Measurements of neuroimmune biomarkers in the sweat, heart rate variability and salivary cortisol. *Neuroimmunomodulation, 17*(3), 205-208. doi:10.1159/000258725

Martin, B. (1961). The assessment of anxiety by physiological behavioral measures. *Psychological Bulletin, 58*(3), 234. doi:10.1037/h0045492

Matthews, G., Hillyard, E. J., & Campbell, S. E. (1999). Metacognition and maladaptive coping as components of test anxiety. *Clinical Psychology & Psychotherapy, 6*(2), 111-125.

Matthews, K. A., Salomon, K., Brady, S. S., & Allen, M. T. (2003). Cardiovascular reactivity to stress predicts future blood pressure in adolescence. *Psychosomatic Medicine, 65*(3), 410-415. doi:10.1097/01.PSY.0000057612.94797.5F

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion, 5*(2), 175. doi:10.1037/1528-3542.5.2.175

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion, 23*(2), 209-237. doi:10.1080/02699930802204677

McCraty, R., Atkinson, M., Tiller, W. A., Rein, G., & Watkins, A. D. (1995). The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American journal of cardiology, 76*(14), 1089-1093. doi:10.1016/S0002-9149(99)80309-9

McCraty, R., & Zayas, M. A. (2014). Cardiac coherence, self-regulation, autonomic stability, and psychosocial well-being. *Frontiers in Psychology, 5*, 1090. doi:10.3389/fpsyg.2014.01090

McDonald, A. S. (2001). The Prevalence and Effects of Test Anxiety in School Children. *Educational Psychology, 21*(1), 89-101. doi:10.1080/01443410020019867

McGlynn, F. D., Bichajian, C., Giesen, J. M., Rullan, C. M., & Pulver, L. (1981). Factorial study of component procedures in desensitization treatment of test anxiety among college students. *Psychological Reports, 49*(2), 351-362. doi:10.2466/pr0.1981.49.2.351

Montgomery, G. K. (1977). Effects of performance evaluation and anxiety on cardiac response in anticipation of difficult problem solving. *Psychophysiology, 14*(3), 251-257. doi:10.1111/j.1469-8986.1977.tb01170.x

Morris, L. W., Davis, M. A., & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry–emotionality scale. *Journal of Educational psychology, 73*(4), 541 - 555. doi:10.1037/0022-0663.73.4.541

Morris, L. W., & Liebert, R. M. (1970). Relationship of cognitive and emotional components of test anxiety to physiological arousal and academic performance. *Journal of Consulting and Clinical Psychology, 35*(3), 332-337. doi:10.1037/h0030132

Morse, D. R., Schacterle, G. R., Furst, M. L., Esposito, J., & Zaydenburg, M. (1982). Stress, relaxation and saliva: Relationship to dental caries and its prevention, with a literature review. *Annals of Dentistry, 42*(2), 47-54.

Myrtek, M., Dieterle, W., & Brügner, G. (1990). Psychophysiological response patterns to variations of the experimental load of a reaction time task. *Journal of Psychophysiology, 4(3)*, 209-220.

Obrist, P. A. (1976). The cardiovascular-behavioral interaction—As it appears today. *Psychophysiology, 13*(2), 95-107. doi:10.1111/j.1469-8986.1976.tb00081.x

Pekrun, R., & Bühner, M. (2014). Self-report measures of academic emotions. In R. Pekrun, & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 561-579). New York: Taylor & Francis.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36-48. doi:10.1016/j.cedpsych.2010.10.002

Pekrun, R., Goetz, T., Perry, R. P., Kramer, K., Hochstadt, M., & Molfenter, S. (2004). Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). *Anxiety, Stress & Coping, 17*(3), 287-316. doi:10.1080/10615800412331303847

Pekrun, R., Goetz, T., & Titz, W. (2002). Academic emotions in students' self regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist, 37*, 91-106. doi:10.1207/S15326985EP3702_4.

Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Pollard, T. M. (1995). Use of cortisol as a stress marker: Practical and theoretical problems. *American Journal of Human Biology, 7*(2), 265-274. doi:10.1002/ajhb.1310070217

Putwain, D. (2007). Researching academic stress and anxiety in students: Some methodological considerations. *British Educational Research Journal, 33*(2), 207-219. doi:10.1080/01411920701208258

Raphelson, A. C. (1957). The relationships among imaginative, direct verbal, and physiological measures of anxiety in an achievement situation. *The Journal of Abnormal and Social Psychology, 54*(1), 13. doi:10.1037/h0041374

Ringeisen, T., Lichtenfeld, S., Becker, S., & Minkley, N. (2019). Stress experience and performance during an oral exam: the role of self-efficacy, threat appraisals, anxiety, and cortisol. *Anxiety, Stress, & Coping, 32*(1), 50-66.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*(6), 934. doi:10.1037/0033-2909.128.6.934

Rosenberg, E., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition & Emotion, 8*(3), 201-229.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, West Sussex: Wiley.

Saito, M., & Nakamura, Y. (1995). Cardiac autonomic control and muscle sympathetic nerve activity during dynamic exercise. *The Japanese Journal of Physiology, 45*(6), 961-977. doi:10.2170/jjphysiol.45.961

Sarason, I. G. (1972). Experimental approaches to test anxiety: Attention and the uses of information. *Anxiety: Current Trends in Theory and Research, 2*, 383-403.

Sarason, I. G. (1978). The Test Anxiety Scale: Concept and research. In C. D. Spielberger, & I. G. Sarason (Eds.), *Stress and Anxiety* (pp. 193-216). Washington, DC: Hemisphere.

Sarason, S. B., Davidson, K., Lighthall, F., & Waite, R. (1958). A test anxiety scale for children. *Child Development*, 105-113. doi:10.2307/1126274

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer, & P. Ekman (Eds.), *Approaches to Emotion* (pp. 317). Hilsdale, NJ: Erlbaum.

Scherer, K. R. (2000). Emotions as episodes of subsystems synchronization driven by nonlinear appraisal processes. In M. D. Lewis & I. Granic (Eds.), *Emotion, development, and self-organization* (pp. 70-99). Cambridge, United Kingdom: Cambridge University Press.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81 - 88. doi:10.1037/h0054570

Schwartz, E. B., Granger, D. A., Susman, E. J., Gunnar, M. R., & Laird, B. (1998). Assessing Salivary Cortisol in Studies of Child Development. *Child Development, 69*(6), 1503-1513. doi:10.1111/j.1467-8624.1998.tb06173.x

Scollon, C. N., Prieto, C.-K., & Diener, E. (2009). Experience sampling: Promises and pitfalls, strength and weaknesses. In *Assessing well-being* (pp. 157-180). Dordrecht: Springer.

Selye, H. (1976a). The stress concept. *Canadian Medical Association Journal, 115*(8), 718.

Selye, H. (1976b). Stress without distress. In *Psychopathology of human adaptation* (pp. 137-146): Springer.

Setz, C., Arnrich, B., Schumm, J., Marca, R. L., Tr, G., #246, et al. (2010). Discriminating stress from cognitive load using a wearable EDA device %J Trans. Info. Tech. Biomed. *14*(2), 410-417. doi:10.1109/titb.2009.2036164

Shaffer, F., McCraty, R., & Zerr, C. L. (2014). A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability. *Frontiers in Psychology, 5*, 1040. doi:10.3389/fpsyg.2014.01040

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*(5), 1055. doi:10.1037/a0023322

Shuman, V., & Scherer, K. R. (2014). Concepts and structures of emotions. In R. Pekrun, & E. A. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 13-35). New York, NY: Taylor & Francis.

Singh, J. P., Larson, M. G., Tsuji, H., Evans, J. C., O'Donnell, C. J., & Levy, D. (1998). Reduced heart rate variability and new-onset hypertension. *Hypertension, 32*(2), 293-297. doi:10.1161/01.HYP.32.2.293

Sloan, R., Shapiro, P., Bagiella, E., Boni, S., Paik, M., Bigger, J., et al. (1994). Effect of mental stress throughout the day on cardiac autonomic control. *Biological Psychology, 37*(2), 89-99. doi:10.1016/0301-0511(94)90024-8

Spangler, G., Pekrun, R., Kramer, K., & Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety, Stress & Coping, 15*(4), 413-432. doi:10.1080/1061580021000056555

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Anton, W. D., Algaze, B., & Ross, G. K. (1980). *Test anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., & Vagg, P. R. (1995). *Test anxiety: A transactional process model*. Philadelphia, PA: Taylor & Francis.

Spodick, D. H. (1993). Survey of selected cardiologists for an operational definition of normal sinus heart rate. *The American journal of cardiology, 72*(5), 487-488.

Stern, R. M., & Higgins, J. D. (1969). Perceived somatic reactions to stress: Sex, age and familial occurrence. *Journal of Psychosomatic Research, 13*(1), 77-82. doi:10.1016/0022-3999(69)90022-1

Strohmaier, A. R., Schiepe-Tiska, A., & Reiss, K. M. (2020). A Comparison of Self-Reports and Electrodermal Activity as Indicators of Mathematics State Anxiety. An Application of the Control-Value Theory. *Frontline Learning Research, 8*(1), 16-32.

Szafranski, D. D., Barrera, T. L., & Norton, P. J. (2012). Test anxiety inventory: 30 years later. *Anxiety, Stress, & Coping, 25*(6), 667-677, doi:10.1080/10615806.2012.663490.

Taylor, J., & Deane, F. P. (2002). Development of a short form of the Test Anxiety Inventory (TAI). *The Journal of General Psychology, 129*(2), 127-136. doi:10.1080/00221300209603133

Turner, J. R., Carroll, D., Hanson, J., & Sims, J. (1988). A comparison of additional heart rates during active psychological challenge calculated from upper body and lower body dynamic exercise. *Psychophysiology, 25*(2), 209-216. doi:10.1111/j.1469-8986.1988.tb00990.x

Uchino, B. N., Berg, C. A., Smith, T. W., Pearce, G., & Skinner, M. (2006). Age-related differences in ambulatory blood pressure during daily stress: Evidence for greater blood pressure reactivity with age. *Psychology and Aging, 21*(2), 231 doi:10.1037/0882-7974.21.2.231

Uchino, B. N., Holt-Lunstad, J., Bloor, L. E., & Campo, R. A. (2005). Aging and cardiovascular reactivity to stress: longitudinal evidence for changes in stress reactivity. *Psychology and Aging, 20*(1), 134 - 143. doi:10.1037/0882-7974.20.1.134

Van Yperen, N. W. (2007). Performing well in an evaluative situation: The roles of perceived competence and task-irrelevant interfering thoughts. *Anxiety, Stress, and Coping, 20*(4), 409-419. doi:10.1080/10615800701628876

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology, 84*(3), 552-569. doi:10.1016/j.biopsycho.2010.01.017

Wilhelm, F. H., & Roth, W. T. (2001). The somatic symptom paradox in DSM-IV anxiety disorders: Suggestions for a clinical focus in psychophysiology. *Biological Psychology, 57*(1-3), 105-140. doi:10.1016/S0301-0511(01)00091-6

Zanstra, Y. J., & Johnston, D. W. (2011). Cardiovascular reactivity in real life settings: Measurement, mechanisms and meaning. *Biological Psychology, 86*(2), 98-105. doi:10.1016/j.biopsycho.2010.05.002

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum Press.

Zeidner, M. (2014). Anxiety in education. In R. Pekrun, & E. A. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 265 - 288). New York, NY: Taylor & Francis.

Zeidner, M., & Matthews, G. (2005). Evaluation anxiety: Current theory and research. In A. J. Elliot, & C. S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 141-163). New York, NY: Guilford Publications.

Zeidner, M., & Matthews, G. (2011). *Anxiety 101*. New York, NY: Springer.

Zung, W. W. (1976). SAS, self-rating anxiety scale. *ECDEU assessment manual for psychopharmacology, revised edition. Rockville, Maryland*, 337-340.

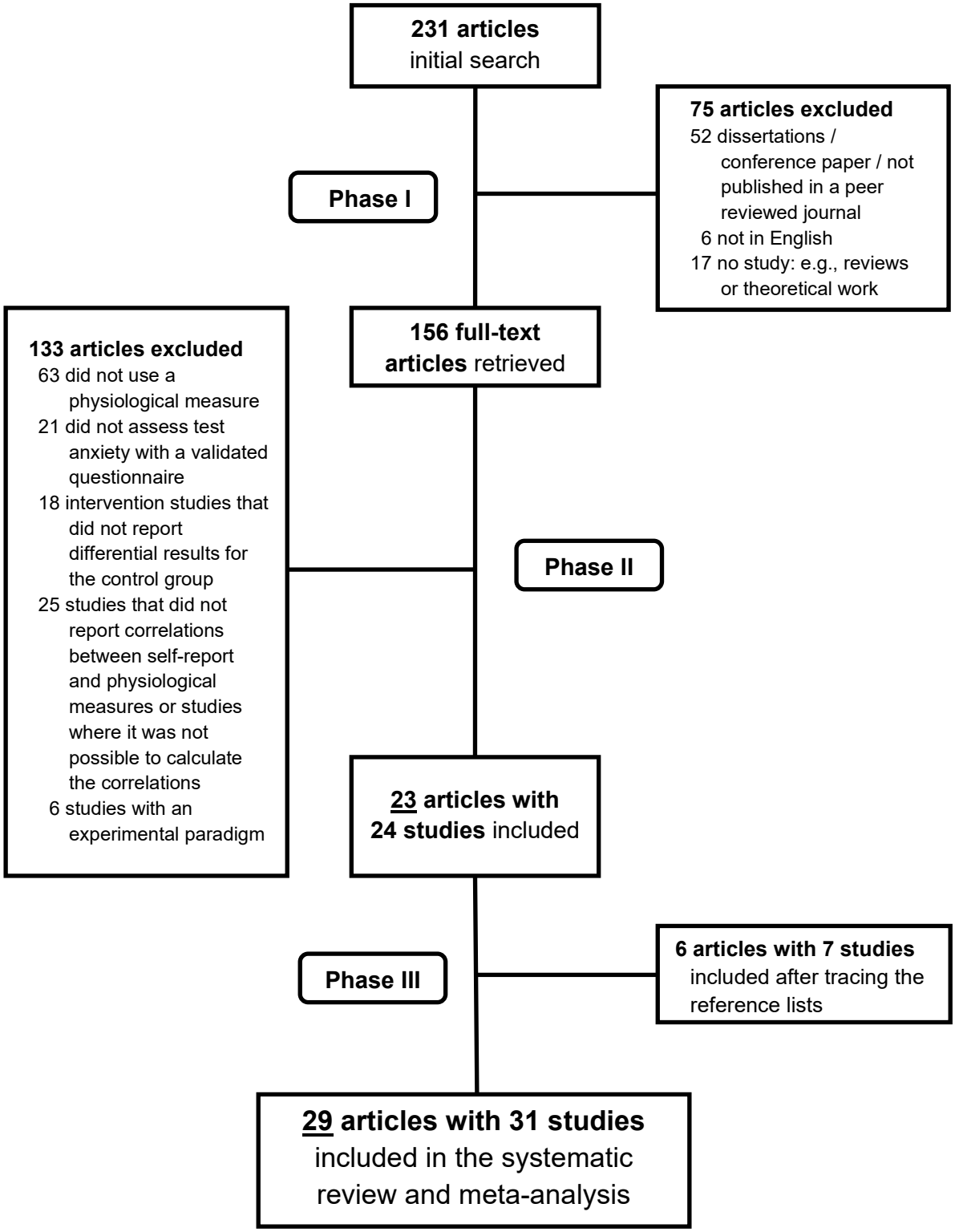

Fig. 1 Procedure and results of the literature search.

Table 1

Search Terms

		PHYSIOLOGICAL MEASURE
"TEST ANXIETY" OR "EXAM ANXIETY" OR "EXAMINATION ANXIETY"	**AND**	Physiology
		Physiological arousal
		Autonomic nervous system
		Cardiac
		Cardiovascular
		Heart rate
		Pulse
		Blood pressure
		Electrodermal
		Skin conductance
		Skin resistance
		Cortisol
		Saliva pH

Table 2

Description of study characteristics and variables coded for the systematic review

Variables	Description
Authors, Year, Country, Journal	
Sample size,	Sample size of the study (*N*),
Gender ratio (f/m),	Proportion female/male subjects (N),
Age,	Age in years: Mean (M_{age}); Standard deviation (*SD*); Range,
Characteristics,	Sample characteristics: i.e., UPSY, HS, ES, USO,
Cont. or G	Assessment: Continuous (Cont.) i.e., without grouping or Grouping (G) i.e., into HTA/LTA
Setting (L/RL),	Setting (i.e., lab vs real life): L / RL; T: study that involves a treatment. Only results from the pretreatment are reported.
Design,	Design: e.g., experimental anxiety induction (Exp.), exam, test,
Area	Area: e.g., language, problem solving
Self-report measure of test anxiety,	Inventory to assess test anxiety: Note: studies sometimes used other measures to examine other constructs as well (e.g., depression, general anxiety). Only the self-report measure for test anxiety are mentioned here,
Test anxiety components (Y/N)	Distinction between the test anxiety components (Y/N)
Physiological measure(s),	Physiological Measure(s): if in ***(i.e., + respiration)*** this measure was not considered for the current review, no results are reported
Assessment,	Assessment of Physiology: device and site,
Measurement schedule	*Physiology measurement schedule.* Anticipation: After subjects received the instructions but did not start with the test yet.
Significant measures,	Measures that yielded significant results, correlation with self-report measure and effect size.
Correlation,	When it is not further mentioned: HTA had higher arousal (as expected): For SR this means lower values for HTA. For HRV this
Effect size	means greater variability for LTA. For pH this means pH sores tend to acidity in HTA.
Measures not significant	Measures where no significant correlation was found.
Baseline (Y/N/NN)	Baseline: Was there a BL assessment (Y/N/NN)? It is also noted when BL was used to calculate change/range corrected scores,
Increase (Y/N/NN)	Increase (Y/N/NN): was there an increase from BL,
Significant (sign. / not sign. / Sign. NN)	Significant (sign. / not sign. / Sign. NN): Information about the significance of the increase from BL.
(Mean) significant Corr.,	(Mean) significant correlation: when there was more than one significant correlation a mean score was calculated,
Effect size	Effect size.

Note. All abbreviations (in alphabetical order) can be retrieved from the Note in Table 3.

Table 3

Studies included in the systematic review and meta-analysis

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
1	Raphelson 1957 USA Journal of Abnormal and Social Psychology	*N* = 25 (0/25) M_{age} = NN, 18 – 25 **USO** **G:** 8/8 + (8 MTA) based on TAQ	**L** Exp. Test instruction: Intellectual and motor abilities	**TAQ** After the test (grouping) **N**	**SC *(+Respiration)*** SC: electrode left hand (palm and wrist) Respiration: pneumograph, no data reported *(measured during null period (i.e., BL), anticipation, test, and posttest)*	**SC** r = 0.47 d = 1.07 Difference in the direction of the trend of SC: HTA increased, LTA decreased		**Y** BL to calculate SC **NN**	**r = 0.47** **d = 1.07**
2	Kissel & Littig 1962 USA Journal of Abnormal and Social Psychology	*N* = 96 (NN) M_{age} = NN **UPSY** **G:** 48/48 based on TAQ	**L** Exp. Perceptual reasoning test: Line drawing problems (unsolvable: induced failure)	**TAQ** (grouping) specific to testing situation **N**	**SC** electrodes non-dominant hand (1. and 3. finger) *(measured during rest period, 1-min. BL, tasks)*	**SC** higher in HTA under induced failure r = 0.28 d = 0.59	-	**Y** BL used as a control variable **NN**	**r = 0.28** **d = 0.59**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
3	Harleston, Smith, & Arey 1965 USA Journal of Personality and Social Psychology	*N* = 42 (42/0) M_{age} = NN **UPSY** **G:** 13 HTA/ 14 MTA/ 15 LTA based on TAQ	**L** Exp. Anagram solving	**TAQ** modified 2 months before the exp. **N**	**HR** Photocell amplifier-recorder unit right hand (thumb) *(measured during pretest, anagram-solving and posttest)*	**HR change** higher in HTA whole task r = 0.31 d = 0.65 first 5 min of task r = 0.30 d = 0.63		**Y** Pretest as BL, BL to calculate change scores **NN**	**Mean** **r = 0.31** **d = 0.64**
4	Morris & Liebert 1970 USA Journal of Consulting and Clinical Psychology	*N* = 95 (NN) M_{age} = NN **UPSY** **Cont.**	**RL** Psychology exam	**TAQ** before the exam (i.e., after HR assessment) **Y** Worry Emotionality (from TAQ)	**HR** Participants count own pulse rate wrist *(measured during 4 times for 15 sec. before normal class and before a regular course examination)*	***HR** and emotionality* *r = 0.34* *d = 0.71* *worry* *r = 0.29* *d = 0.60* ***HR change** and **emotionality*** *r = 0.27* *d = 0.56*	***HR change** and worry*	**Y** Normal class HR as BL (to calculate HR change scores) **NN**	**Mean** **r = 0.30** **d = 0.63**
5	Morris & Liebert 1970 USA	*N* = 91 (NN) M_{age} = NN **HS** **Cont.**	**RL** Sociology-psychology final exam	**TAQ** Before the exam Groups: before HR assessment vs. after HR assessment **Y** Worry Emotionality	**HR** Participants count own pulse rate wrist *(Measured before the exam: either before TAQ administration or after TAQ*	***HR change** and **emotionality** when TAQ was administered first* *r = 0.24* *d = 0.49*	**HR** **HR change** and worry and emotionality	**Y** BL to calculate change scores **NN**	**r = 0.24** **d = 0.49**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	Journal of Consulting and Clinical Psychology			(from TAQ)	*administration; BL collected one month after the exam)*		when HR was assessed first		
6	Darley & Katz 1973 USA Child Development	N = 20 (0/20) M_{age} = NN **HS** 5th grade **Cont.**	**L** Exp. Mental task test group (N= 10) vs. game instruction group (N= 10)	**TASC** after experiments **N**	**HR** EKG electrodes Wrist and ankle *(measured cont. during BL and during and after the experimental manipulations)*	-	**HR change** across both groups did not correlate with TASC scores	**Y** Before experiment **Y** **Sign.** Increase from BL in test group	**NS** **r = NN** **d = NN**
7	Endler & Magnusson 1977 Canada Canadian Journal of Behavioural Science	*N* = 56 (NN) M_{age} = NN **UPSY** **Cont.**	**RL** Psychology exam	**State exam anxiety**: **BRQ** before exam **Y** Cognitive Physiological (from BRQ)	**HR** device not mentioned *(measured before exam, and one week later (i.e., BL))*	**HR** and **BRQ total score** r = 0.59 d = 1.44 ***BRQ Cognitive*** *r = 0.56* *d = 1.33* ***BRQ Physiological*** *r = 0.38* *d = 0.81*	-	**Y** HR one week after exam as BL **Y** **Sign.**	**Mean** **r = 0.51** **d = 1.19**
8	Montgomery	*N* = 48 (0/48)	**L** Exp.	**TAS** (grouping)	**HR *(+Respiration)***	**HR change**		**Y**	**r = 0.43** **d = 0.97**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	1977 USA Psycho-physiology	M_{age} = NN **UPSY** **G:** 24/24 based on TAS: 12/12 (evaluative stress condition) 12/12 (no stress)	(evaluative stress vs. no stress) Anagram solving	**N**	HR: silver electrodes Left and right anterolateral lower ribs Respiration: not reliably obtained *(measured during BL and anticipation)*	higher in HTA in anticipatory phase of the evaluative stress condition r = 0.43 d = 0.97		BL to calculate change scores **NN**	
9	Holroyd, Westbrook, Wolf, & Badhorn 1978 USA Journal of Abnormal Psychology	*N* =72 (72/0) M_{age} = NN **UPSY** **G:** 36/36 based on TAS	**L** Exp. Stroop task, Anagram task	**TAS** (grouping) Before exp. **State test anxiety**: **STAI-State** **N**	**HR, HRV + SC/SR** SC/SR: active electrode non-dominant hand (thenar) HR/HRV: electrodes non-dominant (forearm and opposite leg) *(measured during pretest, stroop task and anagram-solving)*	**HRV** greater variability in LTA during all assessment periods r = 0.26 d = 0.545	**HR** **SC/SR**	**Y** Pretest as BL **Y** **Sign. NN**	**r = 0.26** **d = 0.55**
10	Hollands-worth, Glazeski, Kirkland, Jones, & Van Norman 1979	*N* = 6 (6/0) M_{age} = NN **UPSY** **G:** 3/3 based on TAS scores	**L** (Exp.) Mental ability/ scholastic aptitude test (videotaped)	**TAS** (grouping) Before exp., N = 6 selected from N = 239 students based on extreme TAS and average GAS scores.	**HR + SR *(+ Respiration)*** HR: Silver electrodes, Standard Lead II placement SR: Electrodes, thenar and hypothenar, non-dominant hand	-	**HR *(and respiration)*:** LTA higher arousal during BL and test; more aroused as	**Y** **NN/ not sign.** inconsistent	**NS** **r = NN** **d = NN**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	USA Cognitive Therapy and Research			+ interview to validate TAS scores **N**	*(measured at 1 min intervals during BL, test instruction, anticipatory period and test)*		test began than HTA; **SRR:** inconsistent: HTA highly aroused but 2 of the 3 LTA as well		
11	McGlynn, Bichajian, Giesen, Rullan, & Pulver 1981 USA Psychological Reports	*N* = 47 (20/27) M_{age} = NN **UPSY** **G:** 38/9 based on TAS and palmar sweating item on the PSRSQ	**L + T** Exp. Psychology test (HTA: potential therapy offered LTA: study of test-taking behavior)	**TAS + PSRSQ** (grouping) **N**	**SC + HR** SC: electrodes non-dominant hand (index and 3. finger) HR: photoelectric pulse-pickup non-dominant hand (2. finger) *(pretreatment: measured during habituation, anticipation, and test)*	**SC** higher in HTA during pretreatment Anticipation: r = 0.49 d = 1.142 Pretreatment test-taking: r = 0.48 d = 1.086	**HR**	**Y** HR: habituation as BL SC: range corrected scores **NN**	**Mean** **r = 0.49** **d = 1.12**
12	Smith, Houston, & Zurawski 1984 USA	*N* = 62 (31/31) M_{age} = NN **USO** **Cont.**	**L** (Exp.) Instruction: Intelligence interview (stress condition) vs. listening and	**State exam anxiety** 3 times: baseline, after instruction (i.e., anticipation), after interview **N**	**HR *(+ FPV)*** Physiograph Non-dominant middle finger *(measured during BL, anticipation period and interview)*	**HR** in interview and **state exam anxiety** in anticipation: r = 0.34 d = 0.71	-	**Y** **Y** higher in stress condition **Sign**.	**r = 0.34** **d = 0.71**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	Psycho-physiology		speaking (no-stress condition)						
13	Deffenbacher & Hazaleus 1985 USA Cognitive Therapy and Research	*N* = 129 (65/64) M_{age} = NN **UPSY** **Cont**. and **G:** 66/63 based on TAS	**L** Exp. carried out in a classroom General aptitude test	**TAS** before exp. **State test anxiety**: **STAR** After exp. **Y** State worry and emotionality (from STAR)	**HR** Participants count own pulse rate wrist *(measured by participants themselves for a 15-second period before the exp. (i.e., BL) and immediately after the exp.)*	**HR** higher in HTA **Cont. (TAS)** r = 0.26 d = 0.54 ***Worry*** *r = 0.20* *d = 0.41* ***Emotionality*** *r = 0.27* *d = 0.56*	**HR** when **grouping**	**Y** Prepulse rate as BL **NN**	**Mean** **r = 0.24** **d = 0.5**
14	Sandin & Chorot 1985 Spain Psycho-physiology	*N* = 32 (32/0) M_{age} = 22.8, 20 – 25 **UPSY** **Cont.**	**RL** Psychology Comprehensive oral exam	**State anxiety**: **ECAE** **8SQ-a** 15 days before exam (BL1), immediately before exam, 15 days after exam (BL2) **Y but no results**	**Saliva pH** ***(+ Skin and urinary pH)*** Saliva samples pH meter (measured during BL1, immediately before exam and BL2)	**Saliva pH change** and **ECAE (exam)** r = 0.56 d = 1.31 **8SQ-a (exam)** r = 0.34 d = 0.70	-	**Y** Mean of BL1+2 to calculate change scores **Y** pH decrease from BL **Sign.**	**r = 0.45** **d = 1.01**
15	Tennes & Kreye 1985	*N* = 70 (32/38) M_{age} = 7.7, 7 – 9 **ES** 2nd grade **Cont. +**	**L** standardized achievement test	**TASC** adapted version (grouping) one week after achievement test	**Cortisol** urine samples, *(2hr collections at the same time each day, once a month on normal school days and*	-	**Cortisol and TASC** scores **(cont.)** r = -0.12 NS,	**Y** **Y** **Sign.**	**NS** **r = -0.12** **d = -0.24**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	USA Psycho-somatic Medicine	**G:** median split based on TASC scores	+ classroom observation on normal school days (**RL**)	**N**	*on days of achievement tests)* Controlled for age, gender, and weight.		**Cortisol and HTA/LTA groups**	increase from normal days to test day	
16	Deffenbacher 1986 USA Cognitive Therapy and Research	*N* = 171 (114/57) M_{age} = NN **UPSY** **Cont.** and **G:** 42/52 based on TAS	**RL** Psychology exam	**TAS** 5 weeks before exam **State test anxiety**: **STAR** after exam **Y** State worry and emotionality (from STAR)	**HR** Participants count own pulse rate Non-dominant wrist *(measured by participants themselves immediately upon completion of the exam for a 15-second period)*	**HR** higher in HTA **Group** r = 0.26 d = 0.54 **Cont. (TAS)** r = 0.20 d = 0.40 ***Worry*** *r = 0.14* *d = 0.28* ***Emotionality*** *r = 0.17* *d = 0.34*		**N**	**Mean** **r = 0.19** **d = 0.39**
17	Glazeski, Hollands-worth, & Jones 1986 USA Educational & Psychological Research	*N* = 30 (28/2) M_{age} = NN **UPSY + USO** (business) **G:** 15/15 based on upper and lower 10% of AAT difference score scores	**L** Exp. Problem solving: Mental ability test	**Trait exam anxiety (AAT)** (grouping) Before exp. **N**	**HR + SR + Respiration (= only a single combined arousal index was used)** HR: Silver electrodes Standard Lead II placement SR: Finger-tip electrodes (non-dominant hand) Respiration: Velcro chest band *(measured during BL and test)*	-	Single arousal index (combi-nation of HR, SR and respiration): No difference between HTA and LTA groups	**Y** **Y** **Sign.**	**NS** **r = NN** **d = NN**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
18	Herbert, Moore, De La Riva, & Watts 1986 UK Biological Psychology	N = 38 (0/38) M_{age} = NN **UMS** **Cont.**	**RL** Medicine before major medicine exam	**State anxiety**: **STAI X, worry, emotionality** 3 weeks before (= BL) and immediately before exam **Y** Worry Emotionality	**Cortisol** ***(+ Testosterone + Prolactin + LH)*** Serum from venous blood *(measured 3 weeks before the exam (BL) and immediately before exam)*		**Cortisol (serum)** and state anxiety r = -0.05 NS, ***worry*** r = -0.03 NS, ***emotionality*** r = -0.13 NS	**Y** BL to calculate change **Y** Increase from BL to exam **Sign.**	**NS** **Mean** **r = -0.07** **d = -0.14**
19	Beidel 1988 USA Journal of Abnormal Psychology	*N* = 50 (23/27) M_{age} = 9.1; 8 - 12 **ES** **G:** 25/25 based on TASC	**L** Exp. 2 tasks: Vocabulary test + oral reading session	**TASC** (grouping) Before exp. **N**	**BP + HR** BP: automatic electrosphygmomanometer HR: pulse-rate monitor Non-dominant arm *(measured during adaptation, BL, vocabulary test, oral reading)*	**HR change** higher in HTA in both tasks r = 0.36 d = 0.79	**SBP** **DBP**	**Y** BL to calculate change scores **NN**	***r* = 0.36** ***d* = 0.79**
20	Beidel, Turner, & Trager 1994 USA	*N* = 62 (NN) M_{age} = 10 **ES** **G:** 36/26 based on TASC	**L** (Exp.) vocabulary test + reading aloud before an audience	**TASC** (grouping) Before exp. to identify children with HTA and LTA (initial sample N = 195) **N**	**HR + BP** automated BP/HR monitor Non-dominant arm BP: controlled for weight. *(measured at 2min intervals during BL, test-taking, reading aloud)*	-	**HR + SBP + DBP**: no difference between HTA and LTA	**Y** **NN**	**NS** **r = NN** **d = NN**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	Journal of Anxiety Disorders								
21	Calvo & Miguel-Tobal 1998 Spain Motivation and Emotion	*N* = 56 (36/20) M_{age} = 21.0; *SD* = 2.2 **UPSY** **G:** 28/28 based on TAI	**L** Exp. 2 aptitude tasks (motor skills + linguistic ability) in front of a camera with ego-threat instructions in writing	**TAI** (grouping) Before exp. **N**	**HR + SR** HR: photoelectric finger pulse transducer (middle finger of non-dominant hand) SR: electrodes (thenar and hypothenar) *(measured during adaptation, BL, anticipation, tasks, recovery)*	**HR** higher in HTA in anticipation phase r = 0.37 d = 0.80 **SR** lower in HTA in anticipation r = 0.22 d = 0.45		**Y** **Y** **Sign.** for HR + SC	**Mean** **r = 0.30** **d = 0.62**
22	Huwe, Hennig, & Netter 1998 Germany Anxiety, Stress, & Coping	*N* = 58 (36/22) M_{age} = 24.81, SD = 3.82 **UPSY** **G:** 29/29 based on STAI	**RL** Psychology oral examination (video-taped)	**State exam anxiety: STAI – State** (grouping) immediately before the exam **N**	**HR + Salivary Cortisol *(+sIgA secretory immunoglobulin)*** HR: manually by experimenter: Palpitation of right radial artery Salivary cortisol (+sIgA): collected by salivettes *(measured immediately before and after the exam)*	-	**Cortisol** r = 0.23 NS, **HR** r = 0.23 NS**,** correlations go in the expected directions	**Y** Four weeks after exam **Y** **Sign**.	**NS** **Mean** **r = 0.23** **d = 0.47**
23	Kantor, Endler, Heslegrave, & Kokovski	*N* = 17 (13/4) *Median*$_{age}$ = 26, 24-43 **UPSY**	**RL** Class seminar presentation (students	**State exam anxiety**: **EMAS-State** prior to presenting in a seminar	**HR** Ambulatory heart rate recorder; electrodes attached to upper sternum and fifth left rib	**HR** **state exam anxiety total score** r = 0.46 d = 0.98	-	**Y** **Y** **Sign. NN**	**Mean** **r = 0.47** **d = 1.00**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	2001 Canada Current Psychology	**Cont.**	received a grade)	**Y** Worry Emotionality (from EMAS-State)	*(measured cont. during BL and seminar presentation)*	***Emotionality*** *r = 0.53* *d= 1.18* ***Worry*** *r = 0.41* *d = 0.84*			
24	Spangler, Pekrun, Kramer, & Hofmann 2002 Germany Anxiety, Stress, & Coping	*N* = 40 (24/16) M_{age} = 25.3, *SD* = 2.5, 22 – 32 **USO** **Cont.** and **G:** by median of trait exam anxiety to look at the cortisol response	**RL** Final oral exam in psychology	**Trait exam anxiety** (12 items; see **TEQ**) 4 months before the exam (grouping) **State test anxiety** (Schmitz and Skinner, 1993)) immediately before and after the exam (asked for anxiety during and after the exam) **Y in the trait measure**; **but no results reported**	**Cortisol** Saliva samples salivettes *(measured in the morning of the exam, immediately before the exam, as well as 5 and 15 min after the exam and in the morning of an exam free day (i.e., BL))*	**Cortisol (5 and 15 min.) after the exam** and trait exam anxiety r = 0.49 d = 1.1 r = 0.45 d = 0.99 **Cortisol after the exam** and state anxiety before the exam: r = 0.28 d = 0.57 during the exam: r = 0.30 d = 0.62 **Cortisol response** and **trait** exam anxiety: During: HTA increase, LTA decrease After: HTA higher	-	**Y** + controlled for circadian differences in cortisol level **Y?** anticipatory cortisol response to the exam **NN**	**Mean** **r = 0.38** **d = 0.82**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
25	Spangler, Kramer, Pekrun, & Hofmann 2002 Germany Anxiety, Stress, & Coping	*N* = 26 (15/11) M_{age} = 25.2, 22 – 31 **USO** **G:** Low, medium, high frequency of state test anxiety	**RL** Oral psychology exam videotaped	**State test anxiety during the exam:** Event sampling with a video-based semi-structured interview (grouping) **N**	**Cortisol** Saliva samples *(measured in the morning of the exam, immediately before the exam, as well as 15 min after the exam and in the morning of an exam free day)*	**Cortisol** in the morning and test anxiety: r = 0.34 d = 0.7 before exam and test anxiety: r = 0.40 d = 0.84	-	**Y** morning cortisol of an exam free day **Y?** anticipatory response in HTA **NN**	**Mean** **r = 0.37** **d = 0.77**
26	Daly, Chamberlain, & Spalding 2011 UK Educational Research	N = 23 (21/2) M_{age} = 17.31; SD = 0.47; 16 – 19 **SEC (UK college)** **Cont.**	**RL** **Pilot study** Mock French A-level exam	**TAS** and **CTA** one week before exam **Y but no results** Worry Emotionality (from CTA and TAS)	**HR** Suunto heart rate Memory Belt Chest *(measured in 10-second intervals during BL, immediately before and during the exam)*	**-**	**HR** r = 0.08 NS	**Y** **Y** **Sign.**	**NS** **r = 0.08** **d = 0.16**
27	Zhang, Su, Peng, Yang, & Cheng 2011 China	*N* = 64 (33/31) M_{age} = 20.0, *SD* = 0.1 **UMS** **G:** 20/44 based on SAS (scores over 50)	**RL** Medicine Exam (Pre-review, review, exam period)	**State anxiety**: **SAS** on the first day of the exam period **N**	**HR + BP** Auscultatory blood pressure readings, Cuff around right upper arm *(measured during pre-review (i.e., BL), review and exam period)*	**HR** IA higher in HTA r = 0.44 d = 0.97 **SBP** IA higher in HTA r = 0.36 d = 0.76	-	**Y** Pre-review period as BL **Y** Increase from BL **Sign.**	**Mean** **r = 0.42** **d = 0.91**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	Clinical and Experimental Hypertension					**DBP** IA higher in HTA r = 0.45 d = 0.99 in exam period			
28	Conley & Lehman 2012 USA Stress and Health	*N* = 99 (65/34) M_{age} = 21 **UPSY** **Cont.** and **G:** low, mean, high based on TAI	**RL** Most stressful events of the day (e.g., writing an exam) = acute academic stressor vs. anticipated academic stressors	**TAI short** **Y** Worry Emotionality (8 items each) from TAI short	**BP + HR** ambulatory blood pressure monitors *(Ambulatory assessment over 4 days both at times with and without academic stressor, reports at each time of BP reading, online survey at the end of the day)*	**SBP** higher in HTA r = 0.20 d = 0.41 during acute academic stressors ***SBP*** *and higher* ***worry****:* *r = 0.25* *d = 0.53*	**DBP** **HR**	**N** but **intra-individual analysis** and covariates (alcohol, smoking, alcohol, food, caffeine)	**mean** **r = 0.23** **d = 0.47**
29	Cohen & Khalaila 2014 Israel Journal of Psychosoma-tic Research	*N* = 68 (51/32) M_{age} = 21.9, *SD* = 3.0, 19 - 36 **USO** **Cont.**	**RL** Nursing studies exam (T1) vs non exam period (T2)	**TAI** one hour before exam (T1) and 3 months post-exam (T2) **Y** worry emotionality (from TAI)	**Salivary pH** Saliva samples (Measured at T1 and T2, i.e., BL)	*T1* ***saliva PH*** *and* ***T2 emotionality*** *r = - 0.33* *d = 0.69* (negative correlation as expected)	*T1* ***saliva PH*** *and* ***T2 worry*** **Corr. with-out control variables:** no corr. between pH and TAI	**Y** + control variables: physical activity, smoking **Y** pH decrease from BL **Sign.**	**r = 0.33** **d = 0.69**
30	Ringeisen, Lichtenfeld, Becker, & Minkley 2018	*N* = 92 (46/46) M_{age} = 24.53, SD = 3.07 **UPSY** **Cont.**	**RL** Psychology oral examination (30min)	**State exam anxiety:** Adjective based instrument (Carver and Schreier 1994; Ringeisen and Buchwald 2010):	**Salivary Cortisol** collected via a shortened straw into polypropylene micro tubes *(BL one week before exam, 30min before exam, t2 directly after exam, t3 before*	-	Intercept Anxiety (t1 to t4) and Intercept Cortisol (t1 to t4) **r = 0.10**	**Y** + controlled for age, gender, cultural background, smoking **Y**	**NS** **r = 0.10**

No.	Authors, Year, Country, Journal	Sample size, Gender ratio (f/m), Age, Characteristics, Cont. or G	Setting (L/RL) Design, Area	Self-reported test anxiety measure, Test anxiety components (Y/N)	Physiological measure(s), Assessment, *Measurement schedule*	Measures significant, Correlation, Effect size	Measures not significant	Baseline (Y/N/NN), Increase (Y/N/NN), Significant (Sign./ not sign./Sign. NN)	(Mean) sign. Corr., Effect size
	Germany Anxiety, Stress, & Coping			anxious, fearful, worried assessed at t1 to t4 **N**	*announcement of the grade, t4 30min after announcement of the grade) Intraindividual control design: Starting time varied between 9 am and 4 pm to take into account that cortisol also changes throughout the day.*			**Sign.** increase from control day to exam day	
31	Strohmaier, Schiepe-Tiska, & Reiss 2020 Germany Frontline Learning Research	*N* = 86 (53/33) M_{age} = 23.2, SD = 4.07 **USO** **Cont.**	**L** Experiment Mathematics test	**State anxiety in a mathematics test** Before the two mathematics test asking if participants felt anxious in the previous situation (Goetz et al. 2013) **N**	**SC** Empatica E4 wristbands, frequency of nonspecific skin conductance responses per minute *(BL 5min relaxation, state anxiety assessment, 10 min mathematics test, state anxiety assessment, 10 min mathematics test, state anxiety assessment)* Controlled for gender and achievement		**r = 0.06** ***p* = .63**	**Y** **Y** **Sign.** increase from relaxation period to mathematics test	**NS** **r = 0.06**

Note. Studies are presented in chronological order. Results for the worry and emotionality components are printed in *italics.*

Abbreviations (in alphabetical order): **AAT** = Achievement Anxiety Test (Alpert and Haber 1960); **BL** =baseline; **BP** = Blood pressure; **BRQ** = Behavioral Reactions Questionnaire (Endler and Okada 1975); **CTA** = Cognitive Test Anxiety Scale (Cassady and Johnson 2002); **DBP** = diastolic blood pressure; **ECAE** = Anxiety State Behavioral Scale (Chorot and Sandin 1985); **EMAS** = Endler Multidimensional Anxiety Scales (Endler et al. 1991); **ES** = Elementary school; **EWL** = Adjective checklist on emotional states (Janke et al. 1978); **Exp.** = Experiment; **f** = female; **FPV** = finger pulse volume**; GAS** = General Anxiety Scale (I. G. Sarason 1972); **HR** = Heart rate; **HRV** = Heart rate variability**; HS** = High school; **HTA** = High test anxious; **IA** = increase amplitudes; **L** = Laboratory; **LTA** = Low test anxious; **m** = male; **MTA** = Medium test anxious; **N** = No; **NN** = ot mentioned; **PSRSQ** = Perceived Somatic Reactions to Stress Questionnaire (Stern and Higgins 1969); **RL** = real life; **SAS** = Zung Self-Rating Anxiety Scale (Zung 1976); **SBP** = systolic blood pressure; **SC** = Skin conductance; **SEC** = Secondary

school students; **SR** = skin resistance; **STAI** = State-Trait Anxiety Inventory (Spielberger et al. 1970); **STAR** = State Test Anxiety Report (Deffenbacher and Hazaleus 1985); **T** = treatment; **TAI** = Test Anxiety Inventory (Spielberger et al. 1980); **TAQ** = Test Anxiety questionnaire (Mandler and Sarason 1952) modified by Harleston (1962); **TAS** = Test Anxiety Scale (I. G. Sarason 1972); **TASC** = Test Anxiety Scale for Children (S. B. Sarason et al. 1958); **TAI short** = Test Anxiety Inventory short form (Taylor and Deane 2002); **TEQ** = Test Emotions Questionnaire (Pekrun et al. 2004); **UMS** = University Medicine students; **UPSY** = University Psychology students; **USO** = University students (other than Psychology) **Y** = Yes; **8SQ-a** = Eight State Questionnaire Anxiety (Curran and Cattell 1976).

Table 4

Correlations between Test Anxiety Components and Physiological Measures

Author	Physiological measure	Worry		Emotionality	
		r	*d*	*r*	*d*
Conley and Lehman (2012)	SBP	**0.25***	0.53	0.14	0.28
Deffenbacher (1986)	HR	0.14	0.28	**0.17***	0.34
Deffenbacher and Hazaleus (1985)	HR	0.20	0.41	**0.26***	0.54
Morris and Liebert (1970)	HR	0.29	0.60	**0.34***	0.71
	HR change[a]	0.15	0.30	**0.27***	0.56
Morris and Liebert (1970)	HR	n.s.[b]	-	**0.24***	0.49
Cohen and Khalaila (2014)	SpH	0.11	0.22	**0.33**	0.49
Endler and Magnusson (1977)[c]	HR	**0.56****	1.33	0.38**	0.81
Kantor et al. (2001)	HR	0.41*	0.84	**0.53****	1.18

Note. Abbreviations and full names: Physiological parameters: HR = Heart rate; SBP = Systolic blood pressure; SpH = Saliva pH. n.s. = in this study physiological measure did not correlate significantly with the component. **Bold** = this correlation was higher, but the difference was not tested for significance. *$p < .05$; **$p < .01$.

[a] worry and emotionality were not significantly different for HR. However, HR change scores significantly correlated with emotionality ($r = 0.27$) but not worry.

[b] correlation coefficient was not mentioned in this study.

[c] in this study the worry component was called cognitive component and the emotionality component was called physiological component.

Supplementary Materials

Table S1

Sample Characteristics

Characteristics		Number of studies	Study
College or university students:	Psychology students	17	Kissel and Littig 1962; Harleston et al. 1965; Morris and Liebert 1970/1; Endler and Magnusson 1977; Montgomery 1977; Holroyd et al. 1978; Hollandsworth et al. 1979; McGlynn et al. 1981; Deffenbacher and Hazaleus 1985; Sandin and Chorot 1985; Deffenbacher 1986; Glazeski et al. 1986; Calvo and Miguel-Tobal 1998; Huwe et al. 1998; Kantor et al. 2001; Conley and Lehman 2012; Ringeisen et al. 2018
	Medicine students	3	Smith et al. 1984; Herbert et al. 1986; Zhang et al. 2011
	Other students	5	Raphelson 1957; Spangler et al. 2002/1; Spangler et al. 2002/2; Cohen and Khalaila 2014; Strohmaier et al. 2020
Elementary school or high school students:		6	Morris and Liebert 1970/2; Darley and Katz 1973; Tennes and Kreye 1985; Beidel 1988; Beidel et al. 1994; Daly et al. 2011

Note. Studies are presented in chronological order. Details can be retrieved from Table 3.

Table S2

Study Design and Setting

Design		**Number of studies**	**Study**
Lab studies with experimental paradigms:	Problem solving task	6	Kissel and Littig 1962; Harleston et al. 1965; Darley and Katz 1973; Montgomery 1977; Holroyd et al. 1978; Glazeski et al. 1986
	General aptitude test	6	Raphelson 1957; Hollandsworth et al. 1979; Smith et al. 1984; Deffenbacher and Hazaleus 1985; Tennes and Kreye 1985; Calvo and Miguel-Tobal 1998;
	Vocabulary test	2	Beidel 1988; Beidel et al. 1994
	Mathematics test	1	Strohmaier et al. 2020
	Psychology test	1	McGlynn et al. 1981
Real-life studies:	Psychology exam	11	Morris and Liebert 1970/1; Morris and Liebert 1970/2; Endler and Magnusson 1977; Sandin and Chorot 1985; Deffenbacher 1986; Huwe et al. 1998; Kantor et al. 2001; Spangler et al. 2002/1; Spangler et al. 2002/2; Conley and Lehman 2012; Ringeisen et al. 2018
	Medicine exam	2	Herbert et al. 1986; Zhang et al. 2011
	French exam	1	Daly et al. 2011
	Nursing exam	1	Cohen and Khalaila 2014

Note. Studies are presented in chronological order. Details can be retrieved from Table 3.

Table S3

Self-Report Test Anxiety Measures

Measure	**Number of studies**		**Study**
TAS	7		Montgomery 1977; Hollandsworth et al. 1979; McGlynn et al. 1981
		+ trait measure (1)	Daly et al. 2011
		+ state measure (3)	Holroyd et al. 1978; Deffenbacher and Hazaleus 1985; Deffenbacher 1986
TASC	4		Darley and Katz 1973; Tennes and Kreye 1985; Beidel 1988; Beidel et al. 1994
TAQ	5		Raphelson 1957; Kissel and Littig 1962; Harleston et al. 1965; Morris and Liebert 1970/1; Morris and Liebert 1970/2
TAI	3		Calvo and Miguel-Tobal 1998; Conley and Lehman 2012; Cohen and Khalaila 2014
Other trait test anxiety measures	2		Glazeski et al. 1986
		+ state measure (1)	Spangler et al. 2002/1
State test anxiety	10		Endler and Magnusson 1977; Smith et al. 1984; Sandin and Chorot 1985; Herbert et al. 1986; Huwe et al. 1998; Kantor et al. 2001; Spangler et al. 2002/2; Zhang et al. 2011; Ringeisen et al. 2018; Strohmaier et al. 2020

Note. Studies are presented in chronological order. Details can be retrieved from Table 3.

Implications and Future Directions

Studies with nonsignificant results	Studies with significant results
Study setting - more realistic settings to induce a substantial amount of anxiety (e.g., real life settings)	
Sample characteristics - larger sample sizes - less selective samples (i.e., not only high achievers like medicine or psychology students) - gender balance	
Assessment and analysis of physiological data - sophisticated devices - detailed assessment and analysis: o more measurement points: e.g., baseline, anticipation, stress, and recovery o take different temporal resolutions of different physiological measures into account i.e., do not use a single arousal index across different measures	
	o look at variability of changes in different phases of the situation o look at different wave forms / response patterns of high vs. low anxious o within-individual associations across time better than only between-individual associations at a single measurement point - take influencing variables (e.g., activity, food, and smoking) into account and control for them in the data analysis - take specific characteristics of different physiological measures into account - appropriate data analysis: e.g., use intra-individual analysis (e.g., hierarchical regression)
Assessment and analysis of self-report data - use reliable questionnaires - take the test anxiety components into account - use state self-report measures and conduct intra-individual analyses	

Figure A1. Implications and future directions.