Received: 13 May 2019 | Revised: 1 August 2019 | Accepted: 19 September 2019

DOI: 10.1111/psyp.13500

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR WILEY

Controlling the uncontrolled: Are there incidental experimenter effects on physiologic responding?

Katherine R. Thorson[1] | Wendy Berry Mendes[2] | Tessa V. West[2]

[1]Department of Psychology, New York University, New York, New York

[2]Department of Psychiatry, University of California San Francisco, San Francisco, California

Correspondence
Katherine R. Thorson, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003.
Email: katherine.thorson@nyu.edu

Funding information
Division of Research on Learning in Formal and Informal Settings, Grant/Award Number: DRL1535414

Abstract

The degree to which experimenters shape participant behavior has long been of interest in experimental social science research. Here, we extend this question to the domain of peripheral psychophysiology, where experimenters often have direct, physical contact with participants, yet researchers do not consistently test for their influence. We describe analytic tools for examining experimenter effects in peripheral physiology. Using these tools, we investigate nine data sets totaling 1,341 participants and 160 experimenters across different roles (e.g., lead research assistants, evaluators, confederates) to demonstrate how researchers can test for experimenter effects in participant autonomic nervous system activity during baseline recordings and reactivity to study tasks. Our results showed (a) little to no significant variance in participants' physiological reactivity due to their experimenters, and (b) little to no evidence that three characteristics of experimenters that are well known to shape interpersonal interactions—status (using five studies with 682 total participants), gender (using two studies with 359 total participants), and race (in two studies with 554 total participants)—influenced participants' physiology. We highlight several reasons that experimenter effects in physiological data are still cause for concern, including the fact that experimenters in these studies were already restricted on a number of characteristics (e.g., age, education). We present recommendations for examining and reducing experimenter effects in physiological data and discuss implications for replication.

KEYWORDS

autonomic nervous system, experimenter effects, multilevel modeling, psychophysiology, replication

1 | INTRODUCTION

In 1904, *The New York Times* reported a story about a German horse who dazzled the public by performing feats thought only capable of humans—he could do complex arithmetic and read in German. Several years later, psychologist Oskar Pfungst revealed that the horse could not actually complete these tasks but was instead responding to unintended, nonverbal cues from his trainer that allowed him to answer questions correctly (Samhita & Gross, 2013). The Clever Hans effect is a canonical example of how experimenters can unknowingly shape the behaviors of research participants.

The potentially hidden ways in which experimenters shape participant behavior has received newfound attention, leading some researchers to conclude that "experimenters can be a more powerful stimulus than many researchers, ourselves included, might care to imagine" (Gilder & Heerey, 2018, p. 12). Experimenters have been identified as potential "hidden moderators" in the replication debate (Mitchell, 2014), and the failure to consider individual differences of experimenters, such as gender and race (e.g., Chapman, Benedict, & Shioth, 2018), and difficult-to-detect variations in experimenter behavior (Baumeister, 2016) have been implicated in replication failures (e.g., Gilder & Heerey, 2018).

The use of physiological measures in psychological research is becoming more common (Wilson, 2010), and physiology studies almost always require close contact with an experimenter and, often, an evaluator or confederate (e.g., Cundiff, Smith, Baron, & Uchino, 2016; Lepore, Allen, & Evans, 1993; Mendes, Major, McCoy, & Blascovich, 2008). Although there is some work examining how much experimenters influence participants' physiology within social science research (e.g., Hicks, 1970; Rankin & Campbell, 1955), researchers do not consistently test for experimenter effects in studies of physiology. Here, we demonstrate a set of analytic procedures that takes into account ways in which experimenters can affect physiological responses across different types of research studies. We demonstrate how experimenters can be treated as random effects in multilevel models as a way to understand whether experimenters account for meaningful variation in participants' physiological responding. We also show how researchers can test whether specific theoretically informed experimenter-level characteristics (such as experimenters' status within the lab and whether experimenters are the same race as participants; Aslaksen, Myrbakk, Høifødt, & Flaten, 2007; Carter et al., 2002) account for variance in participants' physiological activity. We show how to apply these approaches to nine data sets that differ from each other in important conceptual and methodological ways (e.g., in the degree to which participants experience physiological reactivity and in the types of tasks that participants complete). Our goal is to describe a set of procedures that can be used flexibly by researchers and that are not dependent upon one type of experimental design or analytic approach.

There are two primary reasons to expect that experimenter effects could occur in studies of peripheral physiology. First, during social interactions, people can exhibit different physiological responses depending on certain visible characteristics of their interaction partners, which may vary across experimenters. For example, the gender, race, facial appearance, and clothing of a person can all influence the physiological responses of their interaction partners (Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell, 2001; Cundiff et al., 2016; Hoggard, Hill, Gray, & Sellers, 2015; Kraus & Mendes, 2014). To the extent that these characteristics vary across experimenters, there may be experimenter effects on participant physiology.

Second, experimenters may engage in different behaviors that research has shown can affect the physiological responses of people's interaction partners. For example, experimenters might behave in a dominant manner while giving instructions to participants, they might engage in direct physical contact when applying physiological sensors, or they might provide support while participants complete a stressful task—all of which might shape participants' physiological responses (Cundiff et al., 2016; Lepore et al., 1993; Waters, West, Karnilowicz, & Mendes, 2018). Borrowing from theoretical models of personality judgment and person perception, experimenters are likely to vary in how expressive their behaviors are, even when they are instructed to behave in certain ways (e.g., as an evaluator or a confederate; Brunswik, 1955; Funder, 1995). When such behaviors differ across experimenters in ways that are not perfectly uniform across all interactions, there may be variability in how participants physiologically respond to those behaviors.

Understanding experimenter effects on physiology is important because researchers often use physiological measures when self-report and overt behaviors are not ideal for answering their theoretical questions of interest (e.g., Cundiff et al., 2016; Lepore et al., 1993; Mendes et al., 2008). Even if researchers attempt to control the behaviors of experimenters and find that experimenters do not affect self-report or behavioral responses, it is still possible that experimenters affect physiological responses. Indeed, physiological responses can differ across participants even when behaviors and self-reports do not differ (e.g., Scheepers, Ellemers, & Sintemaartensdijk, 2009), and they can also be inconsistent with behavior and self-report (e.g., positive behavior exhibited alongside physiological threat; Mendes & Koslov, 2013).

1.1 | Current work

We describe a set of analytic tools for examining experimenter effects in peripheral physiology. Using these tools, we investigate nine data sets totaling 1,341 participants and 160 experimenters across different roles (e.g., lead research assistants, evaluators, confederates) to demonstrate how researchers can test for experimenter effects in participant autonomic nervous system (ANS) activity. We examine whether there is evidence that experimenters yield variability in participants' physiology during baseline recordings and in their physiological reactivity to study tasks. We also test for several factors that might cause this variability.

We look at several measures of ANS activity: pre-ejection period, cardiac interbeat interval, and heart rate variability (described in detail below). These measures reflect different levels of activity in the two branches of the ANS—the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS mobilizes the body for action, and increases in SNS activity are often used as a measure of general arousal or affective intensity (Mendes, 2016). The PNS can coregulate SNS responses and supports homeostasis; increases in PNS activity can be observed during relaxing and positive situations, and decreases in PNS activity can be seen during threatening situations (Grossman & Taylor, 2007). We examine resting baseline responses, which are recorded after physiological sensors have been applied but before any experimental manipulations or study tasks have occurred (given that these measures can differ across people based on their comfort with the research setting, for example; Soto et al., 2012) and reactivity responses from baseline to study tasks (see Method and Table 1 for more study task details).

We outline a conceptual and analytic approach for examining experimenter effects in physiological research. This approach includes testing whether experimenters account for significant variance in participants' physiological activity (i.e., do the participants of some experimenters have, on average, higher physiological activity whereas participants with other experimenters have lower physiological activity?) using random effects in multilevel models. This approach also includes testing for specific experimenter characteristics that could predict physiological responses in systematic ways (i.e., do experimenter characteristics—such as race—predict participants' physiological activity?) using fixed effects in multilevel models.

We then apply this approach to nine studies. We focus on three experimenter characteristics that could affect participants' physiological responses: experimenter status, experimenter gender, and whether experimenters are the same race as participants. We define experimenter status as experimenters' job within the lab—either research assistants in the lab (who are undergraduates, postbaccalaureates, or terminal master's students; "junior" status) or lab managers and doctoral students ("senior" status). We examine experimenter status in five studies. Although the status distinction between those in a junior versus senior role was not made explicit to participants in our studies, people's position in social hierarchies is a well-established predictor of how they behave (Fiske, 2010), and it is possible that such behavioral differences might affect how participants react to junior versus senior experimenters. For example, people at the top of hierarchies often act more dominant and confident than those at the bottom, which could lead their interaction partners to show greater increases in SNS activity (Hall, Coats, & LeBeau, 2005; Keltner, Gruenfeld, & Anderson, 2003). Furthermore, people can react to those at the top of hierarchies with greater decreases in PNS activity compared to those at the bottom (Kraus & Mendes, 2014). Moreover, the junior versus senior distinction is a reasonable proxy for experience since the senior role requires significantly more time in the lab to achieve, and people's experience in certain roles can also affect their role-relevant interactions with others (e.g., the experience level of physicians can influence patient outcomes associated with physician-patient communication; Zolnierek & DiMatteo, 2009). Although researchers have found experimenter effects based on professional status in the past (e.g., in subjective pain reports; Kallai, Barke, & Voss, 2004), to our knowledge, experimenter effects based on professional status have not been examined in the domain of peripheral physiology.

In two studies (the only ones that had sufficient variation in experimenter and participant gender), we also examine whether the gender of experimenters affects participants' physiological activity. We test this question because research has demonstrated differences between men and women in how they communicate and interact with others

TABLE 1 Methodological details for studies analyzed

Study number	Sample size	Participant pool	Location of study	Tasks completed by participants	ANS interval length[a]
Study 1	74	Undergraduates	NYU	5-min evaluation[b] plus 29-min interaction with partner	30 s
Study 2	168	Undergraduates	NYU	5-min computer math task plus 29-min interaction with partner	30 s
Study 3	83	Undergraduates	NYU	5-min evaluation[b]	30 s
Study 4	84	Undergraduates	NYU	5-min computer math task	30 s
Study 5	119	Undergraduates	NYU	5-min speech in front of camera	60 s
Study 6	82	Undergraduates	NYU Abu Dhabi	Three 11-min interactions with different partners	30 s
Study 7	118	18- to 30-year-olds	NYU	28-min interaction with partner	45 s
Study 8	230	Undergraduates	NYU	10-min interaction with group of four other participants	30 s
Study 9	383	18- to 35-year-olds	UCSF	15-min evaluative interaction with confederate	60 s

[a] The same interval length was used across all autonomic nervous system measures per study.

[b] Participants counted backward from 2,023 in 17-step intervals in front of two evaluators and were instructed to restart each time a mistake was made (a modified version of the Trier Social Stress Test).

(e.g., in how much they talk and in how hesitant they are when speaking; Carli, 2013; Wood & Eagly, 2010) in ways that might influence participant physiology, and, indeed, people can exhibit different levels of SNS and PNS activity when interacting with those of another gender (Mendes, Reis, Seery, & Blascovich, 2003; Uno, Uchino, & Smith, 2002). In addition, there are several documented experimenter effects based on gender in other domains (e.g., subjective pain reporting; Gijsbers & Nicholson, 2005; Vigil et al., 2015).

Finally, in two studies (the only ones that had sufficient variation in experimenter race to test this question), we test whether experimenter race (relative to participants' race) affects participants' physiological responses. We test this question because the race of interaction partners can shape how interactions unfold (Toosi, Babbitt, Ambady, & Sommers, 2012), and there is a great deal of evidence that interacting with someone of a different race is more uncomfortable and anxiety provoking than interacting with someone of the same race (Blascovich et al., 2001; Richeson & Shelton, 2007). Indeed, people can exhibit different levels of SNS and PNS activity when interacting with those of another race (Mendes et al., 2008; West, Koslov, Page-Gould, Major, & Mendes, 2017), and experimenter effects based on race have been documented in physiology (e.g., Rankin & Campbell, 1955) as well as other domains (e.g., test performance; Marx & Goff, 2005).

2 | METHOD

2.1 | Methodological approach

We examine nine data sets from our labs (number of participants ranged from 74 to 383 per study) in which ANS activity was measured for 5 min of a baseline, resting period and 5–34 min of experiment time. All data sets were selected a priori, and our sample size was determined by selecting all of the data sets from our labs that met the following requirements: (a) baseline ANS activity was measured, (b) measures of ANS activity were obtained during study tasks, and (c) the identity of experimenters were documented. Table 1 provides an overview of the methodological details of each study. All studies received research ethics committee approval, and all participants gave informed consent before completing study procedures.

2.2 | Experimenters

We identified four experimenter roles across the nine studies: lead research assistant, physiology research assistant, evaluator, and confederate. We define each of these below, and we outline whether participants interacted with each of these types of experimenters and other participants in Table 2. We indicate the number of experimenters per study in Table 3. Experimenters were lab personnel at the undergraduate, postbaccalaureate, or graduate level who were trained by lab managers, graduate students, postdoctoral fellows, and/or faculty members (see online supporting information, Appendix S1, for training details).

In all studies, participants interacted with a lead research assistant (or lead RAs) who greeted them, obtained informed consent, and guided them through the study procedures. Except for Studies 1 and 3, the lead research assistant also applied physiological sensors on participants. In Studies 1 and 3, a different person from the lead research assistant applied the physiological sensors onto participants. These research assistants were never present in the room while participants' ANS responses were being measured (see Table 2) nor did they have any contact with participants after applying the sensors (we refer to these people as physiology research assistants or physiology RAs). In Studies 1 and 3, participants were evaluated by two evaluators in a modified version of the Trier Social Stress Test (Kirschbaum, Pirke, & Hellhammer, 1993). In Study 9, participants interacted with a confederate (both over video camera and in the same room) who was pretending to be another research participant and followed a scripted protocol.

2.3 | Measures

We collected four measures of ANS activity across the nine studies: pre-ejection period (PEP; the amount of time during a cardiac cycle from the electrical impulse that initiates ventricular contraction and the aortic valve opening), cardiac interbeat interval (IBI; the amount of time in milliseconds between heartbeats), and two measures of heart rate variability (respiratory sinus arrhythmia [RSA]: the fluctuation in heart rate that co-occurs with inhalation and exhalation; and the root mean square of successive differences [RMSSD] between heartbeats).

The relationship between the SNS and PNS is complex, and activity in the two branches can be reciprocal, coactivated, or de-coupled (see Berntson, Cacioppo, & Quigley, 1991, 1993, for more information). We examine PEP because it is considered to be one of the purest measures of sympathetic activity (no concurrent influence of the PNS; Cacioppo et al., 1994). Given that PEP measurements (measured here with impedance cardiography) can be more difficult and costly to obtain, we do not obtain them for every study. Thus, we decided to also examine IBI—a measure that is easier to obtain and which we have for every study—even though it is dually innervated by both the SNS and PNS (Brownley, Hurwitz, & Schneiderman, 2000). We present RSA and

TABLE 2 People with whom participants interacted

Study number	Lead research assistant	Physiology research assistant	Evaluator	Confederate	Other participants
Study 1	Yes	Yes (at least 6 min between contact with participant and speech; at least 16 min between contact with participant and partner interaction)	Yes, two (during evaluation)	No	Yes, one
Study 2	Yes	No	No	No	Yes, one
Study 3	Yes	Yes (at least 6 min between contact with participant and speech)	Yes, two (during evaluation)	No	No
Study 4	Yes	No	No	No	No
Study 5	Yes	No	No	No	No
Study 6	Yes	No	No	No	Yes, three
Study 7	Yes	No	No	No	Yes, one
Study 8	Yes	No	No	No	Yes, four
Study 9	Yes (during confederate interaction)	No	No	Yes (during confederate interaction)	No

Note: Where applicable, in parentheses, we have specified when experimenters interacted with participants in relation to the tasks when physiological reactivity was measured. In all studies except Study 9, lead research assistants interacted with participants just prior to reactivity being measured. They delivered the instructions for the tasks when reactivity was measured to participants and then left the room so that participants could begin.

RMSSD because they are both thought to reflect PNS activity (Thayer, Hansen, & Johnsen, 2010). We do not have RSA and RMSSD for every study (because we have some studies in which the data were analyzed in less than 60-s intervals and the current frequency bands that are used for calculating RSA are designed for data binned in 60-s intervals or more). However, RSA and RMSSD are highly correlated with one another (Goedhart, van der Sluis, Houtveen, Willemsen, & de Geus, 2007), and we have at least one of the measures for each study.

2.3.1 | IBIs

In Studies 5 and 7, we used two snap electrodes in a modified Lead II configuration (near the right clavicle, below the rib cage on the left side of the torso) to record electrocardiography (ECG) responses with an integrated system (Biopac MP150 and ECG100C, Biopac Systems, Goleta, CA; see supporting information, Appendix S1). We processed the data using MindWare's heart rate variability software (HRV 3.0.25, MindWare Technologies, Gahanna, OH), which identified the R point of each heartbeat on the ECG waveform. Trained researchers inspected the data for any recording artifacts and cleaned the data as needed. We then obtained a mean IBI for each interval (see Table 1) and computed reactivity scores by subtracting baseline IBI responses (the last interval of baseline) from IBI responses throughout the rest of the studies.

In Study 8, participants wore Polar H7 Bluetooth heart rate sensors on their torso at heart height, which recorded IBI with the Elite HRV smartphone application. Each participant's physiological data were processed by two of three trained researchers. If the first two researchers disagreed on how to process a file, then the third researcher resolved the discrepancy. In Step 1, we used an Excel macro to divide each participant's baseline and group task recordings into 30-s segments (with 12 s of data added on each end for the filter used in Step 3). During this step, the Excel macro also identified potential artifacts and missing signals in each 30-s segment according to a set of specifications listed at https://osf.io/yw4m9/ (e.g., any instance of an IBI 30% greater than the prior IBI). In addition, the Excel macro created line graphs of each 30-s segment of IBIs so that the researchers could visually inspect the data for artifacts and missing signals. In Step 2, we applied corrections to any potential issues or artifacts in the data according to a set of guidelines listed at https://osf.io/yw4m9/ (e.g., if there was an IBI twice as long as the others in a 30-s segment, we split that IBI in half). If there was more than one issue in one 30-s segment, we marked that segment as missing. Overall, we took a conservative approach in Steps 1 and 2 to eliminate any potential artifacts or extreme responses. In Step 3, we obtained a mean IBI for each 30-s

TABLE 3 Number of experimenters

Study number	Source of variance	*n*	*M* participants	*SE*
Study 1	Physiology RA	9	7.78	3.04
	Lead RA	10	7.40	1.91
	Evaluators	13	20.57	9.77
	Combinations of evaluators	17	4.24	0.87
Study 2	Lead RA	15	11.20	2.31
Study 3	Physiology RA	5	16.6	8.73
	Lead RA	5	16.6	2.04
	Evaluators	8	20.75	7.62
	Combinations of evaluators	17	4.88	1.99
Study 4	Lead RA	4	21.00	9.08
Study 5	Lead RA	10	11.90	3.34
Study 6	Lead RA	12	6.75	0.76
Study 7	Lead RA	5	23.4	3.87
Study 8	Lead RA	6	37.5	12.08
Study 9	Lead RA	66	5.55	0.64
	Confederate	60	6.21	0.60

Note: n = the number of different people who played each role per study. For Studies 1 and 3, two evaluators were present at the same time for each study session; we list both the number of individual evaluators as well as the number of combinations of evaluators. *M* participants = average number of participants that each experimenter interacted with. *SE* = standard error.

segment using CMetX Cardiac Metric Software, available from John J. B. Allen at www.psychofizz.org and described more fully in Allen, Chambers, and Towers (2007). We computed reactivity scores by subtracting baseline IBI responses (the last interval of baseline) from IBI responses throughout the rest of the studies.

2.3.2 | PEP

We used ECG and impedance cardiography (ICG) to obtain measurements of PEP. We used band electrodes in a standard tetrapolar configuration for the recording of ICG responses and two snap electrodes in a standard or modified Lead II configuration (held constant within a study) for the recording of ECG responses. A current was passed through the outer band electrodes, and Z0 and its first derivative, Δz/Δt, were recorded from the inner bands. We recorded ICG and ECG responses using either an integrated system (Biopac MP150, Biopac Systems) with amplifiers for ECG (ECG100C) and ICG (NICO100C) or a Bio-Impedance Technology HIC-2500 impedance cardiograph (Chapel Hill, NC; see supporting information). We processed the data using MindWare's impedance cardiography software (IMP 3.0.25, MindWare Technologies, Gahanna, OH), and PEP measurements were calculated as the average amount of time between the Q point on the ECG wave (when the left ventricle contracts) and the B point on the Δz/Δt wave (when the aortic valve opens) per interval (see Table 1). We visually inspected all intervals and manually selected the Q and B points when they were incorrectly identified by the software. We selected the B point as the notch at the beginning of the longest upstroke before the Z point (Lozano et al., 2007). We computed reactivity scores by subtracting baseline PEP responses (the last interval of baseline) from PEP responses throughout the rest of the studies.

2.3.3 | HRV

We used the IBI series (described above) to calculate RSA and/or RMSSD with MindWare's heart rate variability software (HRV 3.0.25; for Studies 1, 2, 7, 9) or CMetX Cardiac Metric Software, described more fully in Allen et al. (2007; for Studies 3, 4, 5, 6, 8), for the interval lengths specified in Table 1. We computed reactivity scores by subtracting baseline responses (the last interval of baseline) from responses throughout the rest of the studies. For simplicity, we present and analyze raw RMSSD reactivity values in the main text. However, given that RMSSD is positively skewed, we applied a natural-log transformation to the data and present the results of analyses with the transformed data in Appendix S1 (the results do not differ meaningfully from those reported in the main text).

2.4 | Analytic approach

We use two approaches for examining experimenter effects for psychophysiological data: (a) estimating random effects to examine whether there is between-experimenter variability in participants' physiological responses, and (b) estimating fixed effects to examine potential sources of between-experimenter variability that predict physiological responses in a systematic way. In the main text, we estimate two-level models in which the dependent variable is either the average physiological response for each participant across the 5-min baseline recording or an average physiological reactivity value for each participant from baseline to the study tasks. We also examined three-level models in which reactivity was not averaged over time for each participant (see Appendix S1 for the results, which are consistent with those reported in the main text). In these models, the dependent variable was the physiological response at each individual time point across the study tasks listed (e.g., in Study 1, we look at baseline, the evaluation, and the partner interaction separately). Time points were nested within participant, and participants were nested within experimenter. We provide syntax for the approaches mentioned here (as well as others) in Appendix S1. For the purposes of this article in which we are interested in experimenter effects and for cross-study consistency in our analyses, we ignore potential nonindependence due to dyad or group. For all analyses, we use an alpha of .05 to determine statistical significance.

2.4.1 | Random effects

By examining random effects within mixed models, we test whether the participants of some experimenters have, on average, higher physiological responses, while participants with other experimenters have lower physiological responses. Random effects are useful because they allow researchers to test whether there is variability in participants' physiological activity between experimenters without knowing exactly what caused that variability. They are also useful if researchers have many experimenters who vary across multiple dimensions (e.g., gender, age, experience) and when experimenters do not fit cleanly into categories that could be estimated as fixed effects. In the analyses here, participants are nested within experimenter, and we specified a random intercept to examine whether intercepts (i.e., average physiological values) vary from experimenter to experimenter (see chapter 4 of Snijders & Bosker, 2012). For studies with multiple types of experimenters (1, 2, 9), we include separate random statements for each role.

2.4.2 | Fixed effects

By examining fixed effects within mixed models, we test for specific sources of variability in participants' physiological responses between experimenters. Fixed effects are useful because they allow researchers to test whether certain factors associated with experimenters explain variability in participants' physiological responding. We incorporate fixed effects into the models above to test whether theoretically meaningful characteristics of experimenters affect the degree of participant responding. We consider three characteristics: status, gender, and race.

Experimenter status

We examined whether the status of the experimenter within the lab (junior status: undergraduate, postbaccalaureate, or terminal master's-level research assistant; senior status: lab manager or doctoral student) affected participants' physiology in Studies 2, 4, 6, 7, and 8 (see Table 4 for *N*s). In these studies, all experimenters were in the role of lead research assistant. We could not examine status of lead research assistants as a categorical fixed effect in our other studies because all lead research assistants were junior status (Studies 1, 3, 5, 9). Statistical power to detect an effect of experimenter status of medium size ranged across the studies from 60.9% to 96.5% (see Table 4). Previous work on experimenter effects in other domains has found medium to large effects based on professional status (e.g., Kallai et al., 2004); so, conservatively, we predicted a medium effect size.

TABLE 4 Experimenter status

	Experimenters at junior status		Experimenters at senior status		
Study number	*N*	Percent of participants run (%)	*N*	Percent of participants run (%)	Statistical power to detect a medium effect (%)
Study 2	13	72.0	2	28.0	89.6
Study 4	2	32.1	2	67.9	62.0
Study 6	11	90.5	1	9.5	60.9
Study 7	4	89.8	1	10.2	76.8
Study 8	5	66.5	1	33.5	96.5

Experimenter gender

We examined whether the gender of the experimenter affected participants' physiology in two studies. In all other studies, the gender of the experimenter was either held constant across the study or was matched between participant and experimenter across the study (making it impossible to differentiate between effects of experimenter vs. participant gender). We examined only male participants in Study 6 because all female participants had a female experimenter; male participants had either a female (n = 8) or a male experimenter (n = 3). In Study 9, we examined the influence of only lead research assistants (female: n = 40; male: n = 15) because all confederates were matched on gender with participants. Previous work on experimenter effects has found medium to large effects based on gender (e.g., Gijsbers & Nicholson, 2005; Vigil et al., 2015), and statistical power to detect an effect of experimenter gender of medium size was 32.3% in Study 6 and 93.8% in Study 9.

Experimenter race

To examine whether having a same-race experimenter (vs. a cross-race experimenter) affected participants' physiology, we examined the responses of all participants in Studies 8 and 9 who identified as one race (i.e., not multiracial; Study 8: 41.3% Asian, 25.7% White, 13.0% Hispanic, 6.1% Black, 0.4% Hawaiian or other Pacific Islander, 0.4% other; Study 9: 51.7% White, 40.2% Black, 0.3% Middle Eastern, 0.3% other). In Study 8, we examined the race of the lead research assistant relative to the participant (as we had only lead research assistants in that study; Asian experimenters [n = 2] and White experimenters [n = 4]). In Study 9, we examined the race of the confederate (as we did not document the race of the lead research assistants; Black experimenters [n = 16] and White experimenters [n = 30]). We could not examine race as a categorical fixed effect in Studies 1, 2, 3, 4, 5, and 7 because we had only White experimenters in these studies. This would confound participant race with whether the experimenter was the same or different race. That is, in these studies, non-White participants always had a different-race experimenter, but White participants always had a same-race experimenter, making it impossible to know whether differences in physiology are due to participant race or matching of participant race with experimenter race. We could not examine race as a categorical fixed effect in Study 6 because this study used an international sample of participants (from 55 countries) for whom racial categories are not defined in similar ways. This makes it difficult to know (without asking explicitly, which we did not do) whether participants perceived their experimenter as same or different race. Statistical power to detect an effect of experimenter race of medium size was 47.8% in Study 8 and 93.8% in Study 9. We assumed race would have a similar-sized or smaller effect as gender, and so, conservatively, we predicted a medium effect size.

3 | RESULTS

Baseline values are presented in Table 5, and reactivity values are presented in Table 6.

3.1 | Random effects

We used random effects to estimate variance in baseline responses (see Table 7) and in reactivity (see Table 8) due to experimenter in each study. These models test whether the participants of some experimenters have, on average, higher physiological responses, while participants with other experimenters have lower physiological responses. Except for one instance (RMSSD reactivity due to evaluators in Study 1 during the partner evaluation), variance due to experimenter was not significant. Given the numerous tests we conducted

TABLE 5 Autonomic nervous system activity during baseline across studies

	Pre-ejection period		Interbeat interval		Respiratory sinus arrhythmia		RMSSD	
Study number	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1	99.35	10.35	757.68	107.39			36.70	21.82
Study 2	103.35	12.00	818.98	134.98			48.11	33.98
Study 3	100.98	12.31	782.93	128.78			39.14	26.07
Study 4	100.45	13.16	784.07	133.90			41.59	28.67
Study 5			806.58	142.87	6.48	1.21	44.89	28.77
Study 6	106.39	12.25	820.42	128.53			49.69	27.15
Study 7			784.46	116.26			42.26	26.05
Study 8			788.02	116.33			43.31	25.05
Study 9	117.96	11.91	899.09	130.31	6.91	1.32		

TABLE 6 Autonomic nervous system reactivity across studies

	Pre-ejection period reactivity		Interbeat interval reactivity		Respiratory sinus arrhythmia reactivity		RMSSD Reactivity	
Study number	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Study 1 (evaluation)	−16.94	10.76	−142.82	88.05			−10.76	17.23
Study 1 (partner interaction)	−4.40	7.07	−31.04	78.35			0.18	16.92
Study 2 (computer task)	−2.87	7.31	−42.65	76.40			1.69	22.65
Study 2 (partner interaction)	−2.43	7.30	−31.11	75.42			1.64	24.39
Study 3	−15.62	13.12	−127.22	101.84			−9.49	24.04
Study 4	−3.77	7.51	−45.66	76.62			−2.28	18.67
Study 5			−95.62	82.81	0.04	1.01	−4.72	20.67
Study 6	−8.75	10.70	−76.54	95.32			−4.28	19.93
Study 7			−5.95	87.73			−2.33	25.07
Study 8			−106.13	100.31			−8.27	20.74
Study 9 (speech)	−10.41	13.33	−159.54	110.65	−0.41	1.31		
Study 9 (confederate interaction)	−8.84	10.98	−124.36	98.68	−0.35	1.20		

TABLE 7 Variance in ANS activity during baseline due to experimenters across studies

Study number	Source of variance	Physiological response	Absolute variance	*SE*	Wald *Z*	*p*
Study 1	Lead RA	IBI	533.55	1,656.10	0.32	.75
Study 1	Lead RA	RMSSD	60.93	68.01	0.90	.37
Study 2	Lead RA	IBI	315.85	868.99	0.36	.72
Study 3	Physiology RA	PEP	3.70	8.25	0.45	.65
Study 3	Physiology RA	RMSSD	69.97	110.59	0.63	.53
Study 5	Lead RA	IBI	1,278.50	1,607.85	0.80	.43
Study 5	Lead RA	RMSSD	46.61	57.52	0.81	.42
Study 6	Lead RA	RMSSD	54.91	100.22	0.55	.58
Study 7	Lead RA	RMSSD	12.43	31.28	0.40	.69
Study 9	Lead RA	RSA	0.02	0.04	0.58	.56

Note: For simplicity, we report only the parameters of the models for which there was enough variance to estimate.

(on both baseline and reactivity responses, for different experimenter roles in nine different studies with multiple physiological measures for each one), we view this one significant effect with caution. Overall, our analyses suggest no evidence of between-experimenter variability in participants' baseline physiological responses or in their physiological reactivity. We found that this was the case both across different studies and different study paradigms as well as across different experimenter roles.

3.2 | Fixed effects

Next, we tested whether specific factors associated with experimenters affected participants' physiology.

3.2.1 | Experimenter status

For each study in Table 4, we added one fixed term to the models conducted above: experimenter status (coded as −1 for junior and 1 for senior). We found that experimenter status did not have a significant effect on ANS responses in any of the studies we analyzed for any of the measures we analyzed (see Table 9). In other words, the status of experimenters (junior vs. senior) had no effect on participants' baseline or reactivity responses.

3.2.2 | Experimenter gender

For Study 6, we added one fixed term to the model conducted above: experimenter gender (coded as −1 for female and 1

TABLE 8 Variance in ANS reactivity due to experimenters across studies

		Pre-ejection period reactivity				Interbeat interval reactivity			
Study number	**Source of variance**	**Absolute variance**	***SE***	**Wald *Z***	***p***	**Absolute variance**	***SE***	**Wald *Z***	***p***
Study 1 (evaluation)	Physiology RA	32.67	31.65	1.03	.30	*NA*	*NA*	*NA*	*NA*
	Lead RA	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
	Evaluators	3.26	11.56	0.28	.79	*NA*	*NA*	*NA*	*NA*
Study 1 (partner interaction)	Physiology RA	2.48	5.03	0.49	.62	*NA*	*NA*	*NA*	*NA*
	Lead RA	8.29	7.03	1.18	.24	*NA*	*NA*	*NA*	*NA*
	Evaluators	1.06	4.18	0.25	.80	*NA*	*NA*	*NA*	*NA*
Study 2 (computer task)	Lead RA	*NA*	*NA*	*NA*	*NA*	263.10	303.52	0.87	.39
Study 2 (partner interaction)	Lead RA	*NA*	*NA*	*NA*	*NA*	20.15	131.44	0.15	.88
Study 3	Physiology RA	5.22	15.03	0.35	.73	3.78	23.09	0.16	.87
	Lead RA	10.95	17.77	0.62	.54	*NA*	*NA*	*NA*	*NA*
	Evaluators	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Study 4	Lead RA	*NA*	*NA*	*NA*	*NA*	16.41	47.01	0.35	.73
Study 5	Lead RA					32.82	209.53	0.16	.88
Study 6	Lead RA	4.07	6.93	0.59	.56	*NA*	*NA*	*NA*	*NA*
Study 7	Lead RA					2.73	202.39	0.01	.99
Study 8	Lead RA					*NA*	*NA*	*NA*	*NA*
Study 9 (speech)	Lead RA	4.31	5.59	0.77	.44	642.92	471.78	1.36	.17
	Confederate	*NA*	*NA*	*NA*	*NA*	91.35	321.65	0.28	.78
Study 9 (confederate interaction)	Lead RA	3.04	4.10	0.74	.46	147.97	302.98	0.49	.63
	Confederate	*NA*	*NA*	*NA*	*NA*	331.67	283.01	1.17	.24
		Respiratory sinus arrhythmia reactivity				**RMSSD reactivity**			
Study number	**Source of variance**	**Absolute variance**	***SE***	**Wald *Z***	***p***	**Absolute variance**	***SE***	**Wald *Z***	***p***
Study 1 (evaluation)	Physiology RA					11.36	18.70	0.61	.54
	Lead RA					4.87	13.02	0.38	.71
	Evaluators					7.73	15.12	0.51	.61
Study 1 (partner interaction)	Physiology RA					75.94	38.76	1.20	.05
	Lead RA					4.92	5.69	1.05	.29
	Evaluators					44.97	18.51	2.43	.015
Study 2 (computer task)	Lead RA					*NA*	*NA*	*NA*	*NA*
Study 2 (partner interaction)	Lead RA					*NA*	*NA*	*NA*	*NA*

(Continues)

TABLE 8 (Continued)

Study number	Source of variance	Respiratory sinus arrhythmia reactivity				RMSSD reactivity			
		Absolute variance	*SE*	Wald Z	*p*	Absolute variance	*SE*	Wald Z	*p*
Study 3	Physiology RA					3.78	23.09	0.16	.87
	Lead RA					*NA*	*NA*	*NA*	*NA*
	Evaluators					*NA*	*NA*	*NA*	*NA*
Study 4	Lead RA					0.58	11.13	0.05	.96
Study 5	Lead RA	*NA*	*NA*	*NA*	*NA*	2.48	21.19	0.12	.91
Study 6	Lead RA					12.59	23.69	0.53	.60
Study 7	Lead RA					50.31	59.12	0.85	.40
Study 8	Lead RA					*NA*	*NA*	*NA*	*NA*
Study 9 (speech)	Lead RA	0.01	0.04	0.35	.73				
	Confederate	0.02	0.03	0.49	.63				
Study 9 (confederate interaction)	Lead RA	0.01	0.03	0.42	.67				
	Confederate	*NA*	*NA*	*NA*	*NA*				

Note: NA = covariance parameter was trimmed from the model because there was not enough variance to estimate it.

for male; as a reminder we examined the responses of only male participants because all female participants had a female experimenter). We found no fixed effects of gender in this study (see Table 10). In other words, the participants of male experimenters did not have different baseline or reactivity responses than the participants of female experimenters.

For Study 9, we added a fixed term for experimenter gender and a fixed term from participant gender (both coded as −1 for female and 1 for male), as well as an interaction term between experimenter and participant gender. We found no main effects of experimenter gender (see Table 10). We found a significant interaction between experimenter and participant gender on IBI baseline activity, but follow-up tests revealed no significant effects of experimenter gender for females or for males on IBI baseline activity. We also found a significant interaction between experimenter and participant gender on RSA baseline activity: there was no effect of experimenter gender for female participants ($b = -0.11$, $p = .32$), but male participants had higher baseline RSA with male experimenters relative to female experimenters ($b = 0.27$, $p = .02$). Given that we found no significant effects of experimenter gender and no other significant interactions between experimenter and participant gender, we view this one interaction with caution. The bulk of the analyses we conducted here suggest that the gender of experimenters and whether that gender matches with participants' gender does not influence participant physiological responding.

3.2.3 | Experimenter race

For Studies 8 and 9, we added one fixed term to the models conducted above: whether the experimenter-participant combination was different race (coded as −1) or same race (coded as 1). We found only one instance in which the match between experimenter and participant race had a significant effect on ANS responses (see Table 11): during the baseline recordings in Study 9, same-race experimenter-participant dyads had lower baseline RSA ($M = 6.76$, $SD = 1.21$) than different-race experimenter-participant dyads ($M = 7.03$, $SD = 1.15$). In general, however, this evidence suggests that having a same-race versus a cross-race experimenter does not affect participants' physiological responses.

4 | DISCUSSION

Using two distinct analytic approaches—treating experimenter as random and treating experimenter as fixed—we found little evidence of experimenter effects on participants' ANS activity during baseline recordings and reactivity to study tasks. Our results showed (a) little to no significant variance in participants' physiological responses due to their

TABLE 9 Effect of experimenter status on autonomic nervous system activity

	Pre-ejection period				Interbeat interval				RMSSD			
Study number	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***
Study 2 (baseline)	1.05 (1.16)	0.91	142	.36	28.93 (11.82)	2.45	3.58	.08	3.54 (2.76)	1.28	159	.20
Study 2 (reactivity during evaluation)	0.01 (0.64)	0.02	139	.99	−10.29 (9.47)	−1.09	7.19	.31	−3.08 (1.70)	−1.81	157	.07
Study 2 (reactivity during partner interaction)	−0.22 (0.61)	−0.36	138	.72	−0.78 (6.84)	−0.12	3.79	.92	−1.66 (1.77)	−0.94	151	.35
Study 4 (baseline)	−0.14 (1.53)	−0.09	81	.93	1.01 (21.32)	0.05	1.29	.97	3.98 (3.16)	1.26	81	.21
Study 4 (reactivity)	−0.27 (0.98)	−0.28	1.23	.82	−0.48 (8.32)	−0.06	81	.95	−2.27 (1.95)	−1.16	81	.25
Study 6 (baseline)	0.94 (2.26)	0.42	78	.68	21.21 (22.99)	0.92	78	.36	−2.31 (8.46)	−0.27	11.87	.79
Study 6 (reactivity)	1.09 (1.83)	0.60	1.83	.60	−18.88 (14.48)	−1.30	74	.20	−1.89 (3.74)	−0.50	7.57	.63
Study 7 (baseline)					−1.07 (17.25)	−0.06	106	.95	−4.07 (4.23)	−0.96	5.22	.38
Study 7 (reactivity)					−2.16 (13.06)	−0.17	7.63	.87	−3.13 (5.86)	−0.53	3.68	.62
Study 8 (baseline)					8.47 (8.60)	0.99	184	.33	0.53 (2.30)	0.23	0.93	.86
Study 8 (reactivity)					−11.56 (7.28)	−1.59	161	.11	−2.36 (1.54)	−1.53	161	.13

experimenters, and (b) little to no evidence that three characteristics of experimenters that are well known to shape interpersonal interactions—status, gender, and race—influenced participants' physiological responses. Although there has been a renewed focus on the subtle ways in which experimenters can influence participants outside of the awareness of researchers (Chapman et al., 2018; Gilder & Heerey, 2018; Judd & Kenny, 2010), our results—of 1,341 participants and 160 experimenters—suggest that experimenters, at least in the two labs from which these data come, are not incidentally influencing participants' physiological responses during research studies.

We found three significant experimenter effects, but given the numerous tests we conducted, we are hesitant to make strong claims about them. In Study 1, we found significant variance due to evaluators in RMSSD reactivity during the partner evaluation. This study had, on average, the most physiological reactivity, and so it is possible that variability in people's responses allowed more opportunity to detect experimenter effects here. In Study 9, we found that male participants had higher baseline RSA with male experimenters relative to female experimenters, potentially because they were more comfortable around same-gender experimenters. We also found that same-race experimenter-participant dyads had lower baseline RSA than different-race experimenter-participant dyads. If anything, we would have expected the opposite pattern—that participants with a different-race experimenter might have felt more uncomfortable than those with a same-race experimenter, resulting in lower RSA during their baseline recording. This finding does not account for individual differences that influence RSA, as it is a raw response instead of a reactivity recording, so future research would need to replicate this finding to draw a stronger conclusion.

Based on our data, we suggest several dimensions on which studies (or their experimenters) can vary that might shape whether experimenter effects are likely to emerge. First, experimenters differ on characteristics that vary in visibility, and less visible characteristics might be less likely to yield experimenter effects. In the studies we examined here, for example, one of the only instances in which we found an experimenter effect was when the race of the experimenter—a highly visible quality—was mismatched with the race of the participant. On the other hand, experimenter status, for example, was not made explicitly clear to participants, and we did not find any effects of status on participants' physiology. Research has shown that, when status differences are made clear (e.g., by explicitly stating the amount of education

TABLE 10 Effects of experimenter gender on autonomic nervous system activity

	Pre-ejection period reactivity				Interbeat interval reactivity				RMSSD reactivity (Study 6)/respiratory sinus arrhythmia reactivity (Study 9)			
Study number	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***
Study 6 (baseline)												
• Experimenter gender	−1.82 (1.90)	−0.96	36	.35	14.55 (20.60)	0.71	36	.48	−2.48 (5.04)	−0.49	6.60	.64
Study 6 (reactivity)												
• Experimenter gender	0.40 (1.30)	0.31	1.88	.79	11.63 (14.95)	0.78	4.76	.47	4.39 (3.02)	1.45	4.98	.21
Study 9 (baseline)												
• Experimenter gender	−0.26 (0.72)	−0.36	311	.72	−1.26 (8.76)	−0.14	154.01	.89	0.08 (0.08)	0.97	29.28	.34
• Experimenter by participant gender	0.57 (0.72)	0.79	311	.43	20.06 (8.71)	2.30	169.69	.02	0.19 (0.08)	2.37	187.79	.02
Study 9 (reactivity during speech)												
• Experimenter gender	0.75 (0.85)	0.88	22.37	.39	7.38 (8.11)	0.91	22.96	.37	−0.12 (0.07)	−1.68	172.21	.10
• Experimenter by participant gender	−0.01 (0.78)	−0.01	234.32	.99	−9.76 (6.45)	−1.51	150.25	.13	−0.11 (0.07)	−1.60	189.55	.11
Study 9 (reactivity during confederate interaction)												
• Experimenter gender	0.81 (0.66)	1.22	23.41	.23	8.15 (5.53)	1.47	144.76	.14	−.07 (0.07)	−1.05	173.59	.29
• Experimenter by participant gender	0.32 (0.62)	0.51	239.07	.61	−5.27 (5.52)	−0.96	163.49	.34	−0.06 (0.07)	−0.91	190.36	.37

TABLE 11 Effect of experimenter race (cross-race vs. same-race) on autonomic nervous system activity

	Pre-ejection period reactivity				Interbeat interval reactivity				RMSSD reactivity (Study 8)/respiratory sinus arrhythmia reactivity (Study 9)			
Study number	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***	***b* (*SE*)**	***t***	***df***	***p***
Study 8 (baseline)					4.78 (8.86)	0.54	159	.59	3.02 (1.82)	1.67	155.93	.10
Study 8 (reactivity)					−3.50 (8.04)	−0.44	139	.66	−1.92 (1.37)	−1.41	135.11	.16
Study 9 (baseline)	0.50 (0.61)	0.80	347	.42	−12.38 (7.43)	−1.67	119.16	.10	−0.14 (0.07)	−2.11	98.87	.04
Study 9 (reactivity during speech)	1.02 (0.73)	1.39	172.29	.17	11.73 (5.89)	1.99	118.00	.50	−0.01 (0.06)	−0.10	131.43	.92
Study 9 (reactivity during confederate interaction)	−0.08 (0.58)	−0.41	148.69	.89	8.31 (5.05)	1.65	92.50	.10	0.01 (0.05)	0.11	121.98	.91

that people have), status can influence people's behavior (Kalkhoff & Barnum, 2000). If status differences of experimenters are made clear in a similar way—for example, by making it clear that an experimenter is a physician versus an undergraduate—status might also influence participant reactivity.

In addition, when there is little variability in either experimenters, participant physiological responses, or both, there are unlikely to be meaningful relationships between experimenters and participant physiology. For instance, in our data, experimenters did differ on some qualities, like their status in the lab, but their variability on other qualities was intentionally restricted to reduce potential experimenter effects (e.g., all experimenters were between the ages of 18 and 30, they were highly educated in Western school systems, and nearly all of them spoke English as their first language). Even the race and gender of our experimenters, which varied in some of our studies, were intentionally restricted in most others. Similarly, in studies for which there is little variability in physiological responses (e.g., very little reactivity), researchers may be less likely to uncover experimenter effects in physiological data. For example, one of the two instances in which we found significant variance due to experimenter was also in the study and on the measure with the most physiological reactivity. This finding points to the possibility that study paradigms that elicit greater physiological reactivity may also be more susceptible to experimenter effects. Thus, although we did not find effects of experimenters on participant physiology in any of the data sets examined here, we did find hypothesized effects of other variables (e.g., experimental manipulation, task type) on physiological reactivity in all of the studies. In other words, there was enough variability in physiological responding for relationships between physiological reactivity and other variables to emerge.

Finally, another dimension on which studies vary is the length of contact that experimenters have with participants and the timing of that contact in relation to physiological recordings. It seems likely that, when experimenters have little contact with participants (perhaps simply asking them to sign a consent form), they would have less influence on participants' physiology than when they have more contact (such as applying physiological sensors). Similarly, when more time elapses between experimenter contact and the recording of participants' physiology (e.g., when participants do another task in between talking to an experimenter and then having their physiology recorded), experimenters might have less influence on participants. For this reason, it is particularly noteworthy that we found no evidence of experimenter effects even in studies for which experimenters were interacting with participants immediately prior to or while physiological responses were being measured. One possibility is that another dimension of studies—how much attention is required for the tasks that participants complete—reduces the potential for experimenter influences on physiology. Experimenters may influence participants less when participants' attention is directed more at the tasks they are completing and less at their experimenters.

Although we did not find evidence of experimenter effects in these studies, we outline several reasons why researchers should continue to be wary of experimenter effects in physiological data. First, as noted above, the variability of our experimenters on several qualities was intentionally restricted, and more varied experimenters might be more likely to yield meaningful variance in participant physiological reactivity. That being said, we argue that, at present, most psychophysiology studies are conducted by and run on people who also fit many of these restrictions, and thus our findings likely generalize to other current psychophysiology research studies conducted at academic institutions (Rad, Martingano, & Ginges, 2018).

Second, experimenters may affect participants' physiology for certain participants only; that is, some participants might be more susceptible to experimenter effects than others. In the current investigation, we considered the interaction between participant and experimenter characteristics only in our analyses examining experimenter gender and experimenter race. That is, we examined whether the matching of experimenter race or gender with participant race or gender, respectively, influenced participants' responses. It is possible that there are few "main effects" of experimenters, and, instead, there are interactions between experimenters and participants. Identifying who is likely to be susceptible to experimenter effects is an important avenue for future work in this area.

Third, we examined only three potential sources of experimenter variance—status, gender, and race—to illustrate our analytical approach, but there are many other characteristics that researchers could consider as well (for a list of potential characteristics, such as physical attractiveness, accent, or knowledge of research hypotheses, see Appendix S1). Some of these characteristics might matter more in particular study contexts or during particular parts of a study: for example, the gender of an experimenter might matter but only during parts of a study when participants are completing gender-stereotyped tasks (i.e., tasks at which one gender is stereotypically better), such as doing math. In these settings, women might have greater physiological reactivity when doing the task around male experimenters than around female experimenters (see Stone & McWhinnie, 2008).

Fourth, in our analyses examining the influence of same-race versus different-race experimenters, we had experimenters and participants identify their own races using fairly broad racial categories. Most notably, the Asian racial category aggregates across many subgroups of racial identities, such as East Asian (people with Chinese heritage) and South Asian (people with Indian heritage). Because of this, it is possible that some experimenter-participant combinations were classified as same-race when one or both parties considered themselves to be of different races. Future research examining the influence of the alignment of experimenter and participant racial identities on participant physiology should explicitly ask both experimenters and participants whether they consider themselves to have the same racial identity as each other to provide more precision to this question.

Fifth, future work might consider whether experimenters influence other physiological processes as well. Although we did not find experimenter effects for four cardiovascular measures of participants' SNS and PNS activity, it is possible that experimenter effects exist for other physiological processes instead. Our results do not rule out the potential presence of experimenter effects on other measures, particularly of other types of physiological activity.

4.1 | Recommendations

We have demonstrated two general analytic approaches that researchers can take if they are interested in understanding experimenter effects in their own studies of physiology, and there are additional ways in which these approaches can be used. In general, we recommend that researchers first start with a model they would estimate if they were not considering variance due to experimenter and build experimenter variance into it. For example, if the interest is in examining change over time in physiological responding, one could estimate a two-level growth curve model (with time points nested within participants) in which participants' intercepts and slopes (and their covariance) are included as random effects (see chapter 4 of Bolger & Laurenceau, 2013, and chapter 5 of Snijders & Bosker, 2012). This model could be turned into a three-level model, with time points (Level 1) nested within participants (Level 2) nested within experimenters (Level 3). An important point here is to make sure that one accounts for the repeated nature of the data (i.e., any nonindependence that is due to multiple physiological measurements coming from the same person over time) before looking for variance due to experimenter. We include syntax for some of these additional options in Appendix S1.

Researchers can test for, as well as control with study design (e.g., having only female experimenters or randomly assigning female vs. male experimenters), the influence of highly visible experimenter characteristics, such as gender and race. When researchers suspect that less visible qualities, like the amount of dominance that experimenters exhibit or the comfort experimenters have when directing participants to complete stressful tasks, might influence participants, more scripted protocols throughout the entire study (especially during times when researchers might tend to go "off script," such as when applying sensors), training, and/or pilot tests (during which researchers test their data for experimenter effects) might be needed.

Researchers might also consider whether experimenters have less of an influence on participant physiology in studies where participants interact with other participants or with several types of experimenters. In other words, might an individual experimenter be more influential if he or she is the only person with whom the participant interacts during the study? Studies could systematically examine this question by having all participants complete the same tasks but randomizing whether they do that task with another participant or another experimenter (a confederate). It is possible that experimenter effects are more likely when participants interact only with an individual experimenter and no one else, and understanding whether this is the case might help researchers better design their studies.

How many experimenters are ideal for a physiology study? To reduce experimenter variance, one might be tempted to use only one or two experimenters. We caution against this approach because it limits the generalizability of one's findings (Judd & Kenny, 2010). Instead, we recommend using multiple experimenters when possible (e.g., five or more) and counterbalancing experimenters across roles. When researchers are concerned that there may be specific characteristics of experimenters affecting participant outcomes (e.g., gender or race), they should conduct power analyses to ensure that they have enough experimenters and participants to detect differences based on those characteristics should they exist. In studies in which experimenters work together, we recommend using different combinations of researchers. Researchers should also consider the contextual variables that might affect participant responding and ensure that these factors are not confounded with experimenter. For example, if researching academic stress, which might vary throughout the week, experimenters should rotate throughout the week.

We suggest that researchers document as transparently as possible how their experimenters are trained and the type of interactions that participants have with experimenters. We also recommend that researchers document experimenter characteristics—such as race, gender, status, and length of time working in the lab—while studies are ongoing, as certain information is often difficult to recover later. Using some of the methods outlined in this article, researchers might also document how much variance is due to experimenter in their studies to help the field determine if there are any paradigms that are particularly sensitive to experimenter effects. For example, paradigms in which experimenters interact with participants while reactivity is being measured may be more sensitive to experimenter effects than paradigms in which they do not. Physiological effects that are less robust—both in terms of effect size and likelihood of replicating—may also be more sensitive to experimenter variance.

Finally, we recommend that researchers include several self-report items in their studies to address whether participants have different subjective experiences based on their experimenters. These items could directly assess participants' reactions to their experimenters ("How threatened did you feel by your experimenter?), or they might simply assess participants' experiences with study tasks on which experimenters might have had an influence ("How threatened did you feel during the speech?"). Although we typically think of the latter as being influenced by aspects of our study (e.g., an experimental manipulation), these might also be influenced by experimenters. Subjective experiences do not always align with people's physiological experiences (and do not align uniformly across people; Cacioppo, Tassinary, & Berntson, 2017), but such responses might help researchers pinpoint study designs or particular experimenters that might have experimenter effects.

In conclusion, we found little evidence that experimenters—across different roles and different study designs—account for meaningful variance in participants' physiology. We hope that the current work will provide researchers with options and ideas for investigating the presence of experimenter effects in their own data as well. Understanding whether and when experimenters account for variance in participants' physiology will help ensure that researchers remain aware of the true inputs to participants' physiological reactivity.

ACKNOWLEDGMENTS

The authors would like to thank the research assistants from the West Interpersonal Perception Lab at NYU and the Emotion, Health, and Psychophysiology Lab at UCSF for their dedication to this research.

DATA AVAILABILITY STATEMENT

All datasets and syntax are provided at https://osf.io/egqvk/. Additional methodological and analytic details can be found in the supplemental material (SM: https://osf.io/egqvk/). Methodological details and study materials for individual studies can be found in the following locations: Thorson, Forbes, Magerman, and West (2019) (Studies 1 through 4; https://osf.io/gpw4j/) and the Open Science Framework (https://osf.io/egqvk/ for Studies 5, 6, 7, and 9; https://osf.io/xu6ep/ for Study 8).

ORCID

Katherine R. Thorson https://orcid.org/0000-0003-1528-1071

REFERENCES

Allen, J. J., Chambers, A. S., & Towers, D. N. (2007). The many metrics of cardiac chronotropy: A pragmatic primer and a brief comparison of metrics. *Biological Psychology*, *74*(2), 243–262. https://doi.org/10.1016/j.biopsycho.2006.08.005

Aslaksen, P. M., Myrbakk, I. N., Høifødt, R. S., & Flaten, M. A. (2007). The effect of experimenter gender on autonomic and subjective responses to pain stimuli. *Pain*, *129*(3), 260–268. https://doi.org/10.1016/j.pain.2006.10.011

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1991). Autonomic determinism: The modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychological Review*, *98*(4), 459–487. https://doi.org/10.1037/0033-295x.98.4.459

Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1993). Cardiac psychophysiology and autonomic space in humans: Empirical perspectives and conceptual implications. *Psychological Bulletin*, *114*(2), 296–322. https://doi.org/10.1037/0033-2909.114.2.296

Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. *Journal of Personality and Social Psychology*, *80*(2), 253–267. https://doi.org/10.1037/0022-3514.80.2.253

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford.

Brownley, K. A., Hurwitz, B. E., & Schneiderman, N. (2000). Cardiovascular psychophysiology. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 224–264). New York, NY: Cambridge University Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Cacioppo, J. T., Berntson, G. G., Binkley, P. F., Quigley, K. S., Uchino, B. N., & Fieldstone, A. (1994). Auotnomic cardiac control. II. Noninvasive indices and basal response as revealed by autonomic blockades. *Pyschophysiology*, *31*, 586–598. https://doi.org/10.1111/j.1469-8986.1994.tb02352.x

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2017). Strong inference in psychophysiological science. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (4th ed.). New York, NY: Cambridge University Press.

Carli, L. L. (2013). Gendered communication and influence. In M. K. Ryan & N. R. Branscombe (Eds.), *The SAGE handbook of gender and psychology* (pp. 199–215). Washington, DC: Sage.

Carter, L. E., McNeil, D. W., Vowles, K. E., Sorrell, J. T., Turk, C. L., Ries, B. J., & Hopko, D. R. (2002). Effects of emotion on pain reports, tolerance and physiology. *Pain Research and Management*, *7*(1), 21–30. https://doi.org/10.1155/2002/426193

Chapman, C. D., Benedict, C., & Shioth, H. B. (2018). Experimenter gender and replicability in science. *Science Advances*, *4*(1), 1–7. https://doi.org/10.1126/sciadv.1701427

Cundiff, J. M., Smith, T. W., Baron, C. E., & Uchino, B. N. (2016). Hierarchy and health: Physiological effects of interpersonal experiences associated with socioeconomic position. *Health Psychology*, *35*(4), 356–365. https://doi.org/10.1037/hea0000227

Fiske, S. T. (2010). Interpersonal stratification: Status, power, and subordination. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 941–982). Hoboken, NJ: John Wiley & Sons Inc.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Gijsbers, K., & Nicholson, F. (2005). Experimental pain thresholds influenced by sex of experimenter. *Perceptual and Motor Skills*, *101*(3), 803–807. https://doi.org/10.2466/pms.101.3.803-807

Gilder, T. S. E., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science*, *29*(3), 403–417. https://doi.org/10.1177/0956797617737128

Goedhart, A. D., Van Der Sluis, S., Houtveen, J. H., Willemsen, G., & De Geus, E. J. (2007). Comparison of time and frequency domain measures of RSA in ambulatory recordings. *Psychophysiology*, *44*, 203–215. https://doi.org/10.1111/j.1469-8986.2006.00490.x

Grossman, P., & Taylor, E. W. (2007). Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions. *Biological Psychology*, *74*(2), 263–285. https://doi.org/10.1016/j.biopsycho.2005.11.014

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, *131*(6), 898–924. https://doi.org/10.1037/0033-2909.131.6.898

Hicks, R. G. (1970). Experimenter effects on the physiological experiment. *Psychophysiology*, *7*, 10–17. https://doi.org/10.1111/j.1469-8986.1970.tb02272.x

Hoggard, L. S., Hill, L. K., Gray, D. L., & Sellers, R. M. (2015). Capturing the cardiac effects of racial discrimination: Do the effects "keep going"? *International Journal of Psychophysiology*, *97*(2), 163–170. https://doi.org/10.1016/j.ijpsycho.2015.04.015

Judd, C., & Kenny, D. (2010). Data analysis in social psychology: Recent and recurring issues. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 115–139). Boston, MA: McGraw-Hill.

Kalkhoff, W., & Barnum, C. (2000). The effects of status-organizing and social identity processes on patterns of social influence. *Social Psychology Quarterly*, *63*(2), 95–115. https://doi.org/10.2307/2695886

Kallai, I., Barke, A., & Voss, U. (2004). The effects of experimenter characteristics on pain reports in women and men. *Pain*, *112*(1–2), 142–147. https://doi.org/10.1016/j.pain.2004.08.008

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, *110*(2), 265–284. https://doi.org/10.1037/0033-295X.110.2.265

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test': A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychology*, *29*(1–2), 76–81. https://doi.org/10.1159/000119004

Kraus, M. W., & Mendes, W. B. (2014). Sartorial symbols of social class elicit class-consistent behavioral and physiological responses: A dyadic approach. *Journal of Experimental Psychology*, *143*(6), 2330–2340. https://doi.org/10.1037/xge0000023

Lepore, S. J., Allen, K. A., & Evans, G. W. (1993). Social support lowers cardiovascular reactivity to an acute stressor. *Psychosomatic Medicine*, *55*(6), 518–524. https://doi.org/10.1097/00006842-199311000-00007

Lozano, D. L., Norman, G., Knox, D., Wood, B. L., Miller, B. D., Emery, C. F., & Berntson, G. B. (2007). Where to B in dZ/dt. *Psychophysiology*, *44*(1), 113–119. https://doi.org/10.1111/j.1469-8986.2006.00468

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology*, *44*(6), 645–657. https://doi.org/10.1348/014466604X17948

Mendes, W. B. (2016). Emotion and the autonomic nervous system. In L. E. Barrett, M. Lewis, & J. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 166–181). New York, NY: Guilford Press.

Mendes, W. B., & Koslov, K. (2013). Brittle smiles: Positive biases toward stigmatized and outgroup targets. *Journal of Experimental Psychology: General*, *142*(3), 923–933. https://doi.org/10.1037/a0029663

Mendes, W. B., Major, B., McCoy, S., & Blascovich, J. (2008). How attributional ambiguity shapes physiological and emotional responses to social rejection and acceptance. *Journal of Personality and Social Psychology*, *94*(2), 278–291. https://doi.org/10.1037/0022-3514.94.2.278

Mendes, W. B., Reis, H. T., Seery, M. D., & Blascovich, J. (2003). Cardiovascular correlates of emotional expression and suppression: Do content and gender context matter? *Journal of Personality and Social Psychology*, *84*(4), 771–792. https://doi.org/10.1037/0022-3514.84.4.771

Mitchell, J. (2014). *On the evidentiary emptiness of failed replications*. Retrieved from http://jasonmitchell.fas.harvard.edu/Papers/Mitchell_failed_science_2014.pdf

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academies of Sciences of the United States of America*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Rankin, R. E., & Campbell, D. T. (1955). Galvanic skin response to Negro and white experimenters. *The Journal of Abnormal and Social Psychology*, *51*(1), 30–33. https://doi.org/10.1037/h0041539

Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science*, *16*, 316–320. https://doi.org/10.1111/j.1467-8721.2007.00528.x

Samhita, L., & Gross, H. J. (2013). The "Clever Hans Phenomenon" revisited. *Communicative & Integrative Biology*, *6*(6), e27122-1–e27122-3. https://doi.org/10.4161/cib.27122

Scheepers, D., Ellemers, N., & Sintemaartensdijk, N. (2009). Suffering from the possibility of status loss: Physiological responses to social identity threat in high status groups. *European Journal of Social Psychology*, *39*, 1075–1092. https://doi.org/10.1002/ejsp.609

Snijders, T. A. B. & Bosker, R. J. (Eds.). (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Soto, J. A., Roberts, N. A., Pole, N., Levenson, R. W., Burleson, M. H., King, A. R., & Breland-Noble, A. M. (2012). Elevated baseline anxiety among African Americans in laboratory research. *Journal of Psychophysiology*, *26*, 105–115. https://doi.org/10.1027/0269-8803/a000073

Stone, J., & McWhinnie, C. (2008). Evidence that blatant versus subtle stereotype threat cues impact performance through dual processes. *Journal of Experimental Social Psychology*, *44*(2), 445–452. https://doi.org/10.1016/j.jesp.2007.02.006

Thayer, J. F., Hansen, A. L., & Johnsen, B. H. (2010). The non-invasive assessment of autonomic influences on the heart using impedance cardiography and heart rate variability. In A. Steptoe (Ed.), *Handbook of behavioral medicine* (pp. 723–740). New York, NY: Springer.

Thorson, K. R., Forbes, C. E., Magerman, A. B., & West, T. V. (2019). Under threat but engaged: Stereotype threat leads women to engage with female but not male partners in math. *Contemporary Educational Psychology*, *58*, 243–259. https://doi.org/10.1016/j.cedpsych.2019.03.012

Toosi, N. R., Babbitt, L. G., Ambady, N., & Sommers, S. R. (2012). Dyadic interracial interactions: A meta-analysis. *Psychological Bulletin*, *138*(1), 1–27. https://doi.org/10.1037/a0025767

Uno, D., Uchino, B. N., & Smith, T. W. (2002). Relationship quality moderates the effect of social support given by close friends on cardiovascular reactivity in women. *International Journal of Behavioral Medicine*, *9*(3), 243–262. https://doi.org/10.1207/S15327558IJBM0903_06

Vigil, J. M., DiDomenico, J., Strenth, C., Coulombe, P., Kruger, E., Mueller, A. A., … Adams, I. (2015). Experimenter effects on pain reporting in women vary across the menstrual cycle. *International Journal of Endocrinology*, 1–8. https://doi.org/10.1155/2015/520719

Waters, S. F., West, T. V., Karnilowicz, H., & Mendes, W. B. (2018). Affect contagion between mothers and babies: Exploring valence and touch. *Journal of Experimental Psychology: General*, *146*(7), 1043–1051. https://doi.org/10.1037/xge0000322

West, T. V., Koslov, K., Page-Gould, E., Major, B., & Mendes, W. B. (2017). Contagious anxiety: Anxious European Americans can transmit their physiological reactivity to African Americans. *Psychological Science*, *28*(12), 1796–1806. https://doi.org/10.1177/0956797617722551

Wilson, L. C. (2010). Psychophysiology: Daunting or doable? *The Observer*. Retrieved from https://www.psychologicalscience.org

Wood, W., & Eagly, A. (2010). Gender. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 629–667). Hoboken, NJ: Wiley.

Zolnierek, K. B. H., & DiMatteo, M. R. (2009). Physician communication and patient adherence to treatment: A meta-analysis. *Medical Care*, *47*(8), 826–834. https://doi.org/10.1097/MLR.0b013e31819a5acc

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Thorson KR, Mendes WB, West TV. Controlling the uncontrolled: Are there incidental experimenter effects on physiologic responding?. *Psychophysiology*. 2020;57:e13500. https://doi.org/10.1111/psyp.13500