HHS Public Access

Author manuscript

Psychophysiology. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

Psychophysiology. 2015 November ; 52(11): 1432–1440. doi:10.1111/psyp.12503.

Methodological Recommendations for a Heartbeat Detection-Based Measure of Interoceptive Sensitivity

Ian R. Kleckner[1], **Jolie Baumann Wormwood**[1], **W. Kyle Simmons**[2,3], **Lisa Feldman Barrett1**[1,4,*], and **Karen S. Quigley**[1,5,*]

[1]Department of Psychology, Northeastern University, Boston, MA.

[2]Laureate Institute for Brain Research, Tulsa, OK.

[3]Faculty of Community Medicine, The University of Tulsa, Tulsa, OK.

[4]Department of Psychiatry and the Martinos Center for Biomedical Imaging, Harvard Medical School, Massachusetts General Hospital, Boston, MA.

[5]Edith Nourse Rogers Memorial VA Hospital, Bedford, MA.

Abstract

Heartbeat detection tasks are often used to measure cardiac interoceptive sensitivity: the ability to detect sensations from one's heart. However, there is little work to guide decisions on the optimum number of trials to use, which should balance reliability and power against task duration and participant burden. Here, 174 participants completed 100 trials of a widely-used heartbeat detection task where participants attempt to detect whether presented tones occurred synchronously or asynchronously with their heartbeats. First, we quantified measurement reliability of the participant's accuracy derived from differing numbers of trials of the task using a correlation metric; we found that at least 40–60 trials were required to yield sufficient reliability. Next, we quantified power by simulating how the number of trials influenced the ability to detect a correlation between cardiac interoceptive sensitivity and other variables that differ across participants, including a variable measured from our sample (body mass index) as well as simulated variables of varying effect sizes. Using these simulations, we quantified the trade-offs between sample size, effect size, and number of trials in the heartbeat detection task such that a researcher can easily determine any one of these variables at given values of the other two variables. We conclude that using fewer than forty trials is typically insufficient due to poor reliability and low power in estimating an effect size, although the optimal number of trials can differ by study.

Keywords

heartbeat detection; interoceptive sensitivity; reliability; power calculation; simulation

Correspondence should be addressed to Ian R. Kleckner, Department of Psychology, Northeastern University, 125 Nightingale Hall, 360 Huntington Ave, Boston, MA. i.kleckner@neu.edu.

*Shared senior authorship

Interoception is broadly defined as the central nervous system processing of signals from the periphery (e.g., heart rate, temperature, blood glucose) and occurs regardless of conscious awareness (Ádám, 1998; Cameron, 2002; Vaitl, 1996). The literature on interoception has grown exponentially over the past ten years in part because of renewed theoretical interest in the links between interoception and emotion (Barrett, Quigley, Bliss-Moreau, & Aronson, 2004; Damasio, 1994; Wiens, 2005), decision-making (Dunn et al., 2010; Werner, Jung, Duschek, & Schandry, 2009), and health (Cameron, 2001; Fassino, Pierò, Gramaglia, & Abbate-Daga, 2004; Herbert, Herbert, & Pollatos, 2011; Mehling et al., 2012; Pollatos et al., 2008). In addition, recent neuroanatomical and neuroimaging studies have helped reveal how the brain supports interoception, which involves brain networks that include the insula and anterior cingulate cortex (e.g., Critchley, Wiens, Rotshtein, Ohman, & Dolan, 2004; Pollatos, Schandry, Auer, & Kaufmann, 2007; for reviews see Barrett & Simmons, 2015; Craig, 2009; Critchley & Harrison, 2013; Damasio & Carvalho, 2013). These data have been useful for clarifying hypotheses about how dysfunctional interoception can cause or exacerbate anxiety and depression (Paulus & Stein, 2010) and addiction and craving (Gray & Critchley, 2007; Naqvi & Bechara, 2010).

Much empirical research on interoception has focused on individual differences in *interoceptive sensitivity*: the ability to detect signals from the body (e.g., heartbeats). Although interoceptive sensitivity can differ across cardiac, respiratory, and gastrointestinal physiological systems (for a review, see Kleckner & Quigley, 2014), heartbeat detection tasks are widely used because they are sensitive to theoretically-relevant individual differences such as anxiety (e.g., Critchley et al., 2004), they relate to features of affective responding such as intensity (e.g., Wiens, Mezzacappa, & Katkin, 2000), and they are methodologically tractable. There are two widely-used approaches to assess cardiac interoception: *the heartbeat detection task* and *the mental tracking method* (for a review, see Jones, 1994). Although the respective outcome measures for these two tasks, which we call *cardiac interoceptive sensitivity* and *heartbeat tracking accuracy*, are frequently used interchangeably in the literature, evidence suggests that they have distinct correlates and thus should be considered to be conceptually distinct (see Kleckner & Quigley, 2014 for a discussion; also see Ceunen et al., 2013; Garfinkel & Critchley, 2013).

One of the most commonly used types of *heartbeat detection task* is called the modified Whitehead task (Whitehead, Drescher, Heiman, & Blackwell, 1977). For each trial, the participant hears a series of ten tones, each of which is triggered by the R-spike in their electrocardiogram (ECG). The participant's task is to report whether the series of tones were coincident with or not coincident with his or her heartbeats. On coincident trials the tones are presented 200 msec after each R-spike and on non-coincident trials the tones are presented 500 msec after each R-spike. We use the term *cardiac interoceptive sensitivity* to refer to the outcome measure derived from this (and related) heartbeat detection tasks. To provide just three examples, individual differences in cardiac interoceptive sensitivity are positively related with greater self-reported emotional intensity during evocative films (Wiens et al., 2000), greater tendency to focus on the arousal component of daily life emotional experiences (Barrett et al., 2004), and greater accuracy in predicting whether a shock would accompany a stimulus in a conditioning task (Katkin, Wiens, & Öhman, 2001).

In a commonly-used *mental tracking method*, participants silently count their own heartbeats over multiple epochs ranging from 25–55 sec in duration, and then report the number of heartbeats counted for each epoch. The number of beats counted for each epoch is then compared to the number of heartbeats recorded via the ECG for each epoch and the score is averaged across trials (Schandry, 1981). We use the term *heartbeat tracking accuracy* to refer to the outcome measure derived from the mental tracking method. Three examples of findings with this task include that individual differences in heartbeat tracking accuracy are positively correlated with memory for emotional words (Werner, Peres, Duschek, & Schandry, 2010) and with greater body-related self-confidence (Duschek, Werner, Reyes del Paso, & Schandry, 2015). Moreover, heartbeat tracking accuracy moderates the relationship between physiological arousal while viewing evocative pictures and subjective arousal (Dunn et al., 2010). The mental tracking method can be implemented relatively easily and the task duration is short (10–15 min). However, the validity of the mental tracking method as strictly a measure of cardiac interoceptive sensitivity has come under serious criticism because performance can be strongly affected by beliefs about one's heart rate (e.g., knowledge of basal heart rate) rather than simply due to the online, beat-to-beat detection of one's heartbeats (Jones, 1994; Knoll & Hodapp, 1992; Phillips, Jones, Rieger, & Snell, 1999; Ring, Brener, Knapp, & Mailloux, 2015).

The heartbeat detection task and mental tracking method have been directly compared in several studies (e.g., Knoll & Hodapp, 1992; Phillips et al., 1999; Pennebaker & Hoover, 1984; Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015), and we highlight three key conclusions from these studies. First, although the two heartbeat tasks can identify participants with very low or very high performance equally well, they typically do not equivalently identify participants with intermediate performance (Knoll & Hodapp, 1992). Second, as noted above, the mental tracking method is influenced by expectations and beliefs (e.g., knowledge of one's typical heart rate) whereas the heartbeat detection task is not (Phillips et al. 1999). Third, these two different methods do not yield consistent results when used in the same participant, suggesting that the outcome measures of these two tasks should be considered conceptually distinct (Pennebaker & Hoover, 1984).

We focus our investigation here on the modified Whitehead heartbeat detection task instead of the mental tracking method because the heartbeat detection task is better validated for assessing individual differences in cardiac interoceptive sensitivity (for a review, see Jones, 1994). However, the modified Whitehead heartbeat detection task is relatively difficult (i.e., many participants do not perform well at the task) and can be relatively time-consuming depending on the number of trials used. Importantly, due to lack of prior research, it is not clear how many trials are needed and thus it is not clear how much the duration of the task can be reduced while maintaining sufficient reliability and power of the outcome.

More generally, selecting the number of trials to utilize in the heartbeat detection task requires making a trade-off between measurement reliability/statistical power and participant burden/time. Not surprisingly, using more trials yields a more reliable measure of cardiac interoceptive sensitivity and greater statistical power, but also increases task duration and burden. In principle, excessive participant burden can reduce the validity of the measurement if participants become disengaged and do not continue to perform well as task duration

increases. In efforts to balance these concerns, researchers utilizing this heartbeat detection task have employed anywhere from 15 trials (e.g., Garfinkel et al., 2015) to 200 trials (e.g., Whitehead et al., 1977).

To our knowledge, only one prior study has addressed the issue of the optimum number of trials to use in the heartbeat detection task. In that study, Jones and colleagues (Jones, Jones, Rouse, Scott, & Caldwell, 1987) correlated heartbeat detection accuracy, a measure of cardiac interoceptive sensitivity, calculated over 50 trials with heartbeat detection accuracy after the first 10, 20, 30, and 40 trials across 24 participants. As expected, the correlation values increased as the number of trials increased up to about 30 trials (Pearson's r values were 0.78, 0.85, 0.96, and 0.97, respectively) and thus the authors recommended using at least 30 trials. Although this analysis provides initial guidance concerning the optimum number of trials, there are additional considerations in selecting the number of trials that have not yet been assessed. First, because most researchers use heartbeat detection measures to assess relationships between cardiac interoceptive sensitivity and other variables that differ across participants, it is important to know how the number of trials influences the statistical power of these between-participant tests. Moreover, existing power analysis tools cannot fully address how many trials of the heartbeat detection task are needed because both the number of participants and the number of trials impact power. Additionally, existing power analysis tools do not account for unique characteristics of the heartbeat detection task such as the likely distribution of cardiac interoceptive sensitivity values across a typical sample (i.e., many scores reflecting low performance) and the fact that each trial yields a binary outcome (correct or incorrect). Second, because the analysis by Jones et al. was limited to 50 trials and 24 participants, we felt it was important to examine how the number of trials impacted the reliability and statistical power of cardiac interoceptive sensitivity scores in a more comprehensive sample using a larger number of trials.

The goal of the present investigation is to allow researchers to make an informed decision concerning the number of trials to use in the heartbeat detection task. To accomplish this goal, we used data from 174 participants who performed 100 trials of the heartbeat detection task and we performed three key analyses. First, we quantified how the number of trials (20–100) influenced the reliability of the cardiac interoceptive sensitivity score. Second, we quantified how the number of trials influenced the statistical power of a correlation between cardiac interoceptive sensitivity and body mass index (BMI) because cardiac interoceptive sensitivity is inversely related to percent body fat (Rouse, Jones, & Jones, 1988). Finally, we simulated cardiac interoceptive sensitivity scores for a large number of "individuals" to allow researchers to conduct a between-participants power analysis for a wider range of sample sizes, effect sizes, and numbers of trials. Importantly, our simulations uniquely address statistical power in this heartbeat detection task, which would not be pragmatically possible using empirical data alone or existing power analysis techniques.

Method

Participants

Two hundred and five participants (120 females, 85 males) were pooled across five studies in our lab that used heartbeat detection with 100 trials per participant with data acquired

between 2012 and 2014. All studies enrolled English-speaking healthy adults who reported no history of mental or cardiovascular illness. Participants were asked to abstain from ingesting caffeine and alcohol for either 12 or 24 hours prior to testing. Dataset one was collected at Northeastern University and included 113 participants (43 males). Dataset two was collected at Northeastern University and included 46 participants (24 males) who were recruited because they were regularly physically active. Dataset three was collected at the Laureate Institute for Brain Research and included 16 participants (8 males). Dataset four was collected at the Laureate Institute for Brain Research and included 16 participants (7 males). Dataset five was collected at Northeastern University and included 14 participants (3 males). Participants were removed from the analysis for any of the following reasons: (i) they did not follow task instructions (4/205; e.g., they felt their heartbeats with their fingers), (ii) they did not complete all 100 heartbeat detection trials (10/205), (iii) they exhibited an anomalous ECG that precluded performing the heartbeat detection task (4/205), or (iv) they were already tested in another study within this sample (15/205, where we used the data acquired in the first test). The final dataset consisted of 174 participants (77 males, 97 females) age 18–57 years (M±SD = 24.09±7.05 years). Participants' mean height (M±SD) was 1.71±0.09 m (range = 1.54–1.91 m), mean weight was 70.78±18.12 kg (range = 40.91–140.45 kg), and mean BMI was 24.22±5.45 kg/m^2 (range = 15.48–47.08 kg/m^2). These BMI scores are typical of young adult American samples (e.g., the median and modal BMI was 25 in a large adult sample from New York, NY in 2004; Van Wye et al., 2008).

Procedure

Participants were greeted and consented in accordance with the research site's institutional review board. Participants also completed a demographics questionnaire (dataset one only) and a brief health questionnaire about intake of caffeine, alcohol, and medications, whether they were suffering from any illnesses, and the number of hours they slept the prior night. Participants' height and weight was then measured (these were self-reported in datasets three and five). Participants were fitted with pre-gelled ConMed Cleartrace Ag/AgCl sensors (Westborough, MA) to record a modified lead II ECG. Participants also wore a respiration belt around the chest, impedance cardiographic sensors (except dataset four), and electrodermal sensors on the hand (except dataset four). Those data are not reported here. Physiological channels were sampled at 1000 Hz using BioLab v. 3.0.8–3.0.13 (Mindware Technologies LTD; Gahanna, OH). After being connected to physiological recording equipment, participants sat quietly for 2–10 minutes (in datasets three and five, participants completed a demographics questionnaire and the PANAS-X [Watson, Clark, & Tellegen, 1988] during this period). Next, participants completed a five minute baseline during which they were asked to sit still (dataset four had no baseline). They then completed 100 trials of the heartbeat detection task described below. Finally, participants completed additional tasks not related to this study. Participants were compensated $5 per half hour (datasets one, two, and five) or $10 per half hour (datasets three and four).

Heartbeat detection task

We assessed cardiac interoceptive sensitivity for each participant using the modified Whitehead heartbeat detection task described in Barrett et al., (2004), which is a previously established modification of the task described in Whitehead et al. (1977). Participants were

seated and viewed instructions on a computer monitor. The researcher instructed the participant to focus on feeling their heart beating in their chest without using the chair, their fingers, or other objects. Participants were instructed to remain still during each trial. For each of the 100 trials, participants heard 10 tones of 50 msec duration each (200 msec duration in dataset one) which were triggered by their own ECG. For a given trial, all 10 tones were presented either 200 msec after the occurrence of each R-spike in the participant's ECG (perceived as coincident with the heartbeats) or 500 msec after each R-spike (perceived as non-coincident with the heartbeats). Equal numbers of trials were presented at each delay time. After the 10 tones, participants indicated whether the series of tones were coincident or non-coincident with their heartbeats. This task is "modified" in that the delay times of 200 ms and 500 ms were found to be superior (e.g., Wiens & Palmer, 2001) in distinguishing coincident from non-coincident timing compared to the delay times originally proposed by Whitehead et al. (128 ms and 384 ms). Participants in datasets two and four also rated their confidence in each response, and participants in datasets one, two, and four were given 30 second breaks at 25%, 50%, and 75% completion. For 25/205 participants there were technical difficulties, which resulted in a 1–3 minute break at some point during the task. The task was self-paced (approximately 20–40 min) and participants pressed a button to begin each trial. This task was implemented using BioLab v. 3.0.8–3.0.13 and an in-house MATLAB program (Mathworks, Natick, MA) that utilized PsychoPhysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Invalid trials were removed if observational data recorded by the researchers during the study indicated the participant moved excessively during a trial (e.g., yawning, stretching) or if the researcher had to talk with the participant during the trial (e.g., reminding them of instructions). Only 2% of trials were lost for these reasons from datasets 2–5. It was not possible to remove trials for behavioral reasons from dataset one because observational data were not recorded. The first three (practice) trials were not analyzed.

Analysis plan

To calculate cardiac interoceptive sensitivity, we used the fraction of trials correct, which we designate *accuracy* in the following text. We utilized accuracy as a measure of cardiac interoceptive sensitivity because of its simplicity and tractability in performing the power analysis simulations and because it is established in studies of heartbeat detection (e.g., Wiens et al., 2000; Critchley et al., 2004; for a review, see Jones, 1994).

In the results that follow, we first assessed the reliability of accuracy over differing numbers of trials and we compared our results to those of Jones et al. (1987). Second, we examined how statistical power was influenced by the number of trials by examining how the number of trials impacted the observed correlation between accuracy and BMI measured in our sample. Third, we used simulated data to quantify how an observed effect size was influenced by both the number of trials and the sample size across a wider range of effect sizes, sample sizes, and numbers of trials. Importantly, these simulations take into account the stochastic and binary nature of the outcome measure (correct or incorrect at each trial) and the measured distribution of cardiac interoceptive sensitivity scores in the current sample. Finally, we utilized these simulations to recommend the optimal number of trials to

reliably observe a between-participants correlation linking cardiac interoceptive sensitivity and a second variable of interest.

Results

The Influence of the Number of Trials on Reliability

Rationale—Measurement reliability is an important consideration when selecting the number of trials used to derive accuracy. If too few trials of the heartbeat detection task are performed, accuracy may not be a good measure of the participant's actual cardiac interoceptive sensitivity due to excessive noise in the measurement. The goal of this analysis is to determine a minimum number of trials needed to meet a reasonable criterion for reliability of accuracy.

Approach—To replicate the reliability analysis used by Jones et al. (1987), we computed the between-participants correlation of accuracy values calculated from performance on all 100 trials and accuracy values calculated from performance on a smaller number of trials (e.g., first 20 trials, first 30 trials, etc)."

Results—The across participants correlation between accuracy at 100 trials and accuracy at another trial (20 or 21 or 22, …, or 100) increased over trials (Figure 1). After 30–40 trials the reliability was well over $r = 0.7$, which is often considered sufficient; and after 60–70 trials the reliability was well over $r = 0.9$, which is extremely high. Thus, while Jones et al. recommended at least 30 trials to achieve $r > 0.9$ with results at 50 trials, our results suggest that researchers should use at least 60 trials to achieve $r > 0.9$ with 100 trials.

Limitations—The "discrepancy" between our results and the results of Jones et al. reflects a limitation of this analysis: specifically, the recommended number of trials needed to achieve $r > 0.9$ increases based on the maximum number of trials in the analysis. Both analyses suggest that $r > 0.9$ is exceeded at 3/5 the total number of trials included in the analysis. So although this approach is a good start toward addressing the number of trials to use, it is a limited measure of reliability. This emphasizes why it is important to use other criteria to determine an optimum number of trials, such as statistical power to detect effects of interest.

The Influence of the Number of Trials on a Between-Participants Correlation: An Exemplar Effect

Rationale—Researchers often use a correlation to test for theoretically relevant relationships between cardiac interoceptive sensitivity and another variable of interest, "*Y*," which can be a behavioral, peripheral physiological, or central neural measure. Despite growing interest in cardiac interoceptive sensitivity there is no systematic research on how the number of trials in a heartbeat detection task influences the ability to assess a between-participants correlation. Although it is established that using more trials increases power, existing power calculators do not capture the unique features of the heartbeat detection task that influence power, such as the typical sample distribution of cardiac interoceptive sensitivity scores and that the outcome of each trial is binary (correct or incorrect).

Approach—To determine how the number of trials in the heartbeat detection task influences the ability to assess a correlation involving cardiac interoceptive sensitivity, we assessed the correlation between accuracy and BMI. We examined how the correlation between accuracy and BMI within our sample changed when accuracy was calculated using trials 1–20, 1–21, 1–22, …, and 1–100. This analysis should demonstrate when we have enough trials for a reliable measure of the relationship between cardiac interoceptive sensitivity and BMI because we have a wide range of scores on both variables.

We also examined whether accuracy differed by gender ($t = -0.39$, $p = 0.70$ at 100 trials), and we assessed the correlations between accuracy and age ($r = -0.05$, $p = 0.49$ at 100 trials), and accuracy and resting baseline heart rate ($r = 0.06$, $p = 0.44$ at 100 trials). Because none of these effects were statistically significant, we do not address them further.

Results and recommendations—Cardiac interoceptive sensitivity was negatively correlated with BMI at 100 trials ($r = -0.20$, $p = 0.009$; Figure 2a), consistent with prior research (Rouse et al., 1988). Moreover, the correlation coefficient was slightly more negative with more trials included (Figure 2b). At 20 trials Pearson's r was about −0.10, and after 25–35 trials Pearson's r consistently exceeded the statistical threshold ($p = 0.05$). Further, the variability in r across trials became notably more stable after 40–50 trials (Figure 2b), suggesting that the observed r was more reliable once the number of trials exceeded 40–50. Importantly, the relationship between cardiac interoceptive sensitivity and BMI is not stronger with more trials, rather the estimate of accuracy improved with more trials, and the observed effect size more closely approached the true effect size. Altogether, these data suggest that the observed effect size for this relationship is $r = -0.20$ at 100 trials using 174 participants, but this relationship can be reliably observed using as few as 25–35 trials with this same number of participants.

Limitations—This analysis is limited to a single observed effect size ($r \sim 0.2$) with 174 participants and 100 trials, and thus cannot be used to extrapolate to other effect sizes, or more than 174 participants and/or more than100 trials. However, data simulations are well-suited for providing a more comprehensive assessment of different effect sizes, sample sizes, and the number of trials (for related data simulation approaches to assessing statistical power, see e.g., Zhang, 2014). In the next two sections, we address these issues of statistical power using simulations.

The Influence of the Number of Trials on the Ability to Estimate the True Effect Size of a Between-Participants Correlation

Rationale—It is useful to know how sample size and the number of trials together influence the effect size of correlations across a wide range of effect sizes both for interpreting existing results and for planning future studies.

Approach—First, we generated a population of 10,000 simulated participants where each participant was assigned a "true" cardiac interoceptive sensitivity score randomly drawn from our large ($N = 174$) empirical distribution of scores at trial 100 (Figure S1). Thus, our simulated population had the same distribution of cardiac interoceptive sensitivity values as

our empirical sample. In the second step, for each of the 10,000 simulated participants, we randomly generated a series of responses (correct or incorrect in proportion to the participant's assigned "true" accuracy) for each trial of the heartbeat detection task. For example, if the simulated participant's true accuracy was 70%, then on any given trial there was a 70% chance of being correct and a 30% chance of being incorrect (like flipping a biased coin). This step was key to capturing the stochastic nature of the task. We simulated 200 trials because it is the upper limit of published papers using this task (Jones, 1994). In a third step, we calculated how the number of trials of the heartbeat detection task influenced the observed correlation between cardiac interoceptive sensitivity and a simulated variable that differs across participants and has the same distribution as the cardiac interoceptive sensitivity values. To do this, we set up a between-participants correlation of $r_{true} = 1$ between a simulated variable (Y) and true cardiac interoceptive sensitivity, both of which differ across participants. Then, for each of the 200 simulated trials, we calculated the observed correlation ($r_{observed}$) with Y using the simulated cardiac interoceptive sensitivity values calculated for that particular trial. In a separate analysis, we found that the relationship between r_{true} and $r_{observed}$ did not depend on the value of r_{true}, and thus we just provide the ratio $r_{observed} / r_{true}$ using $r_{true} = 1$. To account for variability across different samples that is due to their finite size (25, 50, 75, 100, 150, or 200 participants), we repeated the above analysis by randomly drawing the desired sample size 1,000 times from the population of 10,000 simulated individuals. We report the 95% confidence interval of the distribution of $r_{observed} / r_{true}$ values at each trial for each sample size.

Results and recommendations—Figure 3 shows that as more trials were performed, a hyperbolic curve describes how the observed effect size ever more closely matches the true effect size. For example, if $r_{true} = 0.3$, then using 20, 50, or 100 trials would yield an observed effect size that is on average 65%, 80%, or 90% of its true value respectively, or $r_{observed} = 65\% \times r_{true} = 0.195$, $80\% \times r_{true} = 0.24$, or $90\% \times r_{true} = 0.27$, respectively (blue line in Figure 3, "infinite participants" column in Table 1). Moreover, smaller samples exhibited greater variability (larger confidence intervals) in the relationship between observed effect size and true effect size. For example, if the true effect size was $r_{true} = 0.3$, then testing 25, 50, or 200 participants using 60 trials would yield an observed effect size that is 60–93%, 69–91%, or 77–87% of its true value, respectively. Additionally, these results can help researchers infer the true effect size from an existing study using an observed effect size and sample size. For example, a study that used 20 trials with 50 participants and reported an (observed) effect size of $r = 0.25$ likely reflects a true effect size between 0.32 and 0.60 (i.e., $r_{true} = 0.25 / 65\% = 0.38$ with 95% CI = 0.32–0.60). Altogether, we recommend at least forty trials to achieve a reasonable estimate of the true effect size because the $r_{observed} / r_{true}$ curve falls steeply below forty trials (Figure 3).

Limitations—There are two primary assumptions of this simulation. First, the distribution of cardiac interoceptive sensitivity values across participants was derived from our large empirical dataset ($N = 174$; Figure S1). Therefore, this simulation is limited if these data do not generalize to other samples and populations. There are, however, several reasons to think that these data are generalizable to other samples of American young adults. In particular, our sample was relatively diverse in terms of basic demographics (i.e., height, weight, BMI,

age, gender, and race) and was collected from two different laboratories across the United States. That said, several specific sub-populations that may be of particular interest to researchers in this field were purposefully excluded from the current sample, including participants with previous or current psychiatric disorders, and those taking medications that can influence cardiac activity. If such sub-populations have dramatically different distributions of cardiac interoceptive sensitivity than the general public, then our simulation may be less generalizable for power analyses with those specific samples. The second primary assumption of these simulations is that the distribution of the second variable Y has the same distribution as the cardiac interoceptive sensitivity values, which may not be the case for certain variables (e.g., they may be normally or otherwise distributed). However, a supporting analysis showed that this assumption is reasonable. We showed that if we instead use a normally distributed Y (vs. the positively skewed distribution of accuracy values; Figure S1) there were only minimal reductions in the correlation coefficient values (0.05 units or less).

The Influence of the Number of Trials on the Ability to Reliably Observe a Between-Participants Correlation

Rationale—In this section, we address the final step to determine the minimum number of trials required to reliably observe a correlation for a given effect size and sample size. To do this, we account for how the statistical threshold for Pearson's r changes as a function of sample size.

Approach—This analysis extends the simulations from the prior section. Specifically, for each true effect size (r_{true} from zero to one in increments of 0.001) we found the smallest trial number such that $r_{observed}(trial)$ exceeded the statistical threshold at $p < 0.05$ for the given sample size using the lower bound of the 95% confidence interval for $r_{observed}(trial)$ / r_{true} from Figure 3 or Table 1. We used the lower bound of the confidence interval to be conservative in the face of variability across samples (i.e., smaller sample sizes tend to exhibit greater variability in the observed effect size than larger samples sizes).

Results and recommendations—Figure 4 and Table 2 both quantify the relationship between the number of heartbeat detection trials, sample size, and effect size for a between-participants correlation of cardiac interoceptive sensitivity and any other variable with a similar distribution as the cardiac interoceptive sensitivity values. Table 2 enables easier determination of specific values for guiding future research. There are three ways to use these results. First, for a given sample size and effect size, one can determine how many trials are needed to reliably observe a correlation of a given magnitude. For example, using 100 participants with an effect of $r_{true} = 0.3$, the study needs at least 25 trials to reliably observe the correlation. Also, because this simulation analysis builds upon the prior simulation, it accounts for how the number of trials and sample size influence the observed effect size (Figure 3). Second, for a given number of trials and effect sizes, one can determine how many participants are needed to reliably observe the correlation. For example, using 110 trials with an effect of $r_{true} = 0.2$, the study needs 100 participants to reliably observe the correlation. Finally, for a given number of trials and sample size, one

can determine the minimum effect size that can be reliably observed. For example, using 75 participants and 50 trials, one can reliably observe a correlation of $r_{true} > 0.275$.

Limitations—The limitations of the prior simulations also apply to these simulations. But, again, the assumptions made are reasonable given the large empirical sample from which these simulations were derived and the results of our supporting analyses regarding other distributions of simulated Y values.

The Influence of Time-on-Task on the Reliability of Measuring Cardiac Interoceptive Sensitivity

Rationale—One possible concern in using the heartbeat detection task is whether it is possible to have participants perform too many trials. Indeed, researchers have acknowledged that the heartbeat detection task is tedious (Jones, 1994) and suggested that participant compliance may be compromised when using many trials. If this were the case, the measurement might become invalid because the data reflects systematic reductions in observed cardiac interoceptive sensitivity over the course of the experiment.

Approach—To assess whether the participants in our dataset disengaged from the task at any point (e.g., due to fatigue, boredom or distraction), we analyzed how each participant's performance changed across the 100 trials of the heartbeat detection task. As a control analysis, we also analyzed how performance fluctuated across 100 trials simply due to chance and the binary nature of the task. See Supporting Materials for more details.

Results and recommendations—Our results indicate that anywhere between 0 and 15% of our sample exhibited systematic reductions in cardiac interoceptive sensitivity across the 100 trials of our dataset (Figure S2). To assess the effect of up to 15% of our sample exhibiting a modest performance decrement (guessing on 25% of trials), we modified the interoceptive sensitivity values in the correlation with BMI (Figure 2). Specifically, we reduced the interoceptive sensitivity scores of the best-performing 26 participants (top 15% of the sample) by an amount that was as if they were guessing (50% accuracy) on 25% of the trials. This reduced the correlation strength from $r = -0.20$ ($p = 0.009$) to $r = -0.19$ ($p = 0.013$), which is a fairly small change. Because we find limited evidence of performance decrements in our sample, it appears safe to use up to 100 trials without significant loss of participant compliance. However, for studies that need to use more than 100 trials, it may be important to consider ways to continue to enhance motivation or reduce possible task disengagement (e.g., provision of breaks between blocks of trials).

Discussion

Our results provide guidance for choosing the number of trials needed to ensure sufficient reliability and statistical power of a measure of cardiac interoceptive sensitivity for any given effect size and sample size. Considering both our reliability analyses and power analyses, we conclude that researchers should use at least forty trials of the heartbeat detection task. Using fewer than forty trials severely compromises the ability to estimate the true effect size (Figure 3) and severely reduces reliability (Figure 1). Importantly, our simulations reflect the unique characteristics of the modified Whitehead heartbeat detection

task, such as the stochastic and binary nature of the outcome measure and the distribution of cardiac interoceptive sensitivities across individuals from our large empirical sample (Figure S1). Indeed, although existing power analysis tools can provide the minimum sample size required to observe an expected effect size (e.g., G*Power; Faul, Erdfelder, Lang, & Buchner, 2007), there are no existing power analysis tools that also consider how the number of trials for a heartbeat detection task impacts the ability to detect a given effect size for a particular sample size. Additionally, because our data came from five different samples with slight variations across protocols, our findings are likely relatively robust to methodological variations such as using mandatory breaks during the task, using different questionnaires and baseline durations before the heartbeat detection task, and providing different amounts of monetary compensation.

As with any study, there are limitations. First, our recommendations may not generalize to samples that have different distributions of interoceptive sensitivity values than our sample (e.g., due to much higher or lower percent body fat). However, because our sample is relatively large ($N = 174$) and diverse (height, weight, BMI, age, gender, and race), our data are likely representative of a community or student sample. Second, our correlation power analysis recommendations may not generalize to variables that are not normally distributed or distributed like the interoceptive sensitivity values herein (positively skewed; Figure S1). Thus, we recommend ensuring that, when a study moves to the analysis phase, variables tested for correlation with interoceptive sensitivity are linearly related to interoceptive sensitivity and/or are normally distributed (e.g., using a transformation if needed). Third, because the heartbeat detection task is so difficult (about half the sample performs at floor), the task may not be optimally sensitive for use in studying individual differences in interoceptive sensitivity (for a discussion, see Khalsa, Rudrauf, Sandesara, Olshansky, & Tranel, 2009). To address this limitation, the task can be made easier by increasing the magnitude of the heartbeat signal (e.g., using the beta adrenergic agonist isoproterenol; Khalsa et al., 2009) and/or minimizing sources of exteroceptive noise.

In future studies, it will be important to select the proper number of trials in the heartbeat detection task because heartbeat detection is being used in increasingly complex studies where more resources are at stake (e.g., neuroimaging, longitudinal, and interventional studies). Additionally, the need for specific power calculators (as we present here) will continue to be important as theoretical models involving cardiac interoceptive sensitivity become more sophisticated. For example, heartbeat tracking accuracy *moderates* the link between physiological responding and subjective experience or behavior (Dunn et al., 2010; Werner, Schweitzer, Meindl, Duschek, Kambeitz, & Schandry, 2013), suggesting the possibility that cardiac interoceptive sensitivity may also be a moderating factor. Thus, just as we incorporated the number of heartbeat detection trials into a power analysis of *correlation effects*, future work could incorporate the number of heartbeat detection trials into a power analysis of *moderation effects* (e.g., using modifications to calculations in the *pwr* package in R; Champely, Ekstrom, Dalgaard, Gill, & Rosario, 2015).

By helping researchers to select the number of trials to use in a well-established and commonly used heartbeat detection task, we hope to enhance research on cardiac interoceptive sensitivity across a wide range of topics including emotion, memory, decision-

making, health behaviors, and mental illness. Our results will prove useful to researchers studying interoception in both designing their own studies and comparing results across studies. Use of the tools we provide should result in greater power, reliability, and interpretability in the growing literature on cardiac interoceptive sensitivity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge members of the Interdisciplinary Affective Science Laboratory for discussion and critical feedback and Amber Kleckner for comments on an earlier version of this manuscript. This research was supported by the National Institute of Mental Health post-doctoral award (F32MH096533) to I.R.K., the National Institutes of Mental Health (K01MH096175-01) grant to W.K.S., the Oklahoma Tobacco Research Center grant to W.K.S., the National Institutes of Health Director's Pioneer Award (DP1OD003312) to L.F.B., the Army Research Institute for the Behavioral and Social Sciences (Contract number: W5J9CQ-12-C-0049) to L.F.B. and K.S.Q. Thanks to Joel Barcalow and Jennifer Dobson at the Laureate Institute for Brain Research for help with participant recruitment. The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

References

Ádám, G. Visceral Perception. 1st. New York: Springer-Verlag; 1998.

Barrett LF, Quigley KS, Bliss-Moreau E, Aronson KR. Interoceptive sensitivity and self-reports of emotional experience. Journal of Personality and Social Psychology. 2004; 87:684–697. [PubMed: 15535779]

Barrett LF, Simmons WK. Interoceptive predictions in the brain. Nature Reviews Neuroscience. 2015

Brainard DH. The psychophysics toolbox. Spatial Vision. 1997; 10:433–436. [PubMed: 9176952]

Cameron, O. Visceral Sensory Neuroscience: Interoception. New York, New York: Oxford University Press, Inc.; 2002.

Cameron OG. Interoception: The inside story - A model for psychosomatic processes. Psychosomatic Medicine. 2001; 63:697–710. [PubMed: 11573016]

Ceunen E, Van Diest I, Vlaeyen J. Accuracy and awareness of perception: related, yet distinct (commentary on Herbert et al., 2012). Biological Psychology. 2013; 92:423–427.

Champely S, Ekstrom C, Dalgaard P, Gill J, Rosario HD. pwr: Basic Functions for Power Analysis (R Package). 2015

Craig AD. How do you feel? Interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience. 2002; 3:655–666. [PubMed: 12154366]

Craig AD. How do you feel--now? The anterior insula and human awareness. Nature Reviews Neuroscience. 2009; 10:59–70.

Critchley HD, Harrison NA. Visceral influences on brain and behavior. Neuron. 2013; 77:624–638. [PubMed: 23439117]

Critchley HD, Wiens S, Rotshtein P, Ohman A, Dolan RJ. Neural systems supporting interoceptive awareness. Nature Neuroscience. 2004; 7:189–195. [PubMed: 14730305]

Damasio A, Carvalho GB. The nature of feelings: Evolutionary and neurobiological origins. Nature Reviews Neuroscience. 2013; 14:143–152. [PubMed: 23329161]

Damasio, AR. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Putnam; 1994.

Dunn BD, Galton HC, Morgan R, Evans D, Oliver C, Meyer M, Dalgleish T. Listening to your heart: How interoception shapes emotion experience and intuitive decision making. Psychological Science. 2010; 21:1835–1844. [PubMed: 21106893]

Duschek S, Werner NS, Reyes del Paso GA, Schandry R. The Contributions of Interoceptive Awareness to Cognitive and Affective Facets of Body Experience. Journal of Individual Differences. 2015; 36:110–118.

Fassino S, Pierò A, Gramaglia C, Abbate-Daga G. Clinical, psychopathological and personality correlates of interoceptive awareness in anorexia nervosa, bulimia nervosa and obesity. Psychopathology. 2004; 37:168–174. [PubMed: 15237246]

Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavioral Research Methods. 2007; 39:175–191.

Garfinkel SN, Critchley HD. Interoception, emotion and brain: new insights link internal physiology to social behaviour. Commentary on:: “Anterior insular cortex mediates bodily sensibility and social anxiety” by Terasawa et al. (2012). Social Cognitive and Affective Neuroscience. 2013; 8:231–234. [PubMed: 23482658]

Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD. Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. Biological Psychology. 2015; 104:65–74. [PubMed: 25451381]

Gray MA, Critchley HD. Interoceptive basis to craving. Neuron. 2007; 54:183–186. [PubMed: 17442239]

Herbert BM, Herbert C, Pollatos O. On the relationship between interoceptive awareness and alexithymia: Is interoceptive awareness related to emotional awareness? Journal of Personality. 2011; 79:1149–1175. [PubMed: 21241306]

Jones G. Perception of visceral sensations: A review of recent findings, methodologies, and future directions. Advances in Psychophysiology. 1994; 5:155–192.

Jones GE, Jones KR, Rouse CH, Scott DM, Caldwell JA. The effect of body position on the perception of cardiac sensations: an experiment and theoretical implications. Psychophysiology. 1987; 24:300–311. [PubMed: 3602286]

Katkin ES, Wiens S, Öhman A. Nonconscious fear conditioning, visceral perception, and the development of gut feelings. Psychological Science. 2001; 12:366–370. [PubMed: 11554668]

Khalsa SS, Rudrauf D, Sandesara C, Olshansky B, Tranel D. Bolus isoproterenol infusions provide a reliable method for assessing interoceptive awareness. International Journal of Psychophysiology. 2009; 72:34–45. [PubMed: 18854201]

Kleckner, IR.; Quigley, KS. An approach to mapping the neurophysiological state of the body to affective experience. In: Barrett, LF.; Russell, J., editors. The Psychological Construction of Emotion. New York: Guilford; 2014.

Kleiner M, Brainard D, Pelli D. What's new in Psychtoolbox-3. Perception, 36. 2007

Knoll JFF, Hodapp V. A comparison between two methods for assessing heartbeat perception. Psychophysiology. 1992; 29:218–222. [PubMed: 1635964]

Mehling WE, Price C, Daubenmier JJ, Acree M, Bartmess E, Stewart A. The Multidimensional Assessment of Interoceptive Awareness (MAIA). PLoS ONE, 7. 2012:e48230.

Naqvi NH, Bechara A. The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. Brain Structure Function. 2010; 214:435–450. [PubMed: 20512364]

Paulus MP, Stein MB. Interoception in anxiety and depression. Brain Structure Function. 2010; 214:451–463. [PubMed: 20490545]

Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision. 1997; 10:437–442. [PubMed: 9176953]

Pennebaker JW, Hoover CW. Visceral perception versus visceral detection: disentangling methods and assumptions. Biofeedback and Self-Regulation. 1984; 9:339–352. [PubMed: 6525358]

Phillips GC, Jones GE, Rieger EJ, Snell JB. Effects of the presentation of false heart-rate feedback on the performance of two common heartbeat-detection tasks. Psychophysiology. 1999; 36:504–510. [PubMed: 10432800]

Pollatos O, Herbert BM, Matthias E, Schandry R. Heart rate response after emotional picture presentation is modulated by interoceptive awareness. International Journal of Psychophysiology. 2007; 63:117–124. [PubMed: 17137662]

Pollatos O, Kurz A-L, Albrecht J, Schreder T, Kleemann AM, Schöpf V, Schandry R. Reduced perception of bodily signals in anorexia nervosa. Eating Behaviors. 2008; 9:381–388. [PubMed: 18928900]

Pollatos O, Schandry R, Auer DP, Kaufmann C. Brain structures mediating cardiovascular arousal and interoceptive awareness. Brain Research. 2007; 1141:178–187. [PubMed: 17296169]

Pollatos O, Kurz AL, Albrecht J, Schreder T, Kleemann AM, Schopf V, Schandry R. Reduced perception of bodily signals in anorexia nervosa. Eating Behaviors. 2008; 9:381–388. [PubMed: 18928900]

Ring C, Brener J, Knapp K, Mailloux J. Effects of heartbeat feedback on beliefs about heart rate and heartbeat counting: A cautionary tale about interoceptive awareness. Biological Psychology. 2015; 104:193–198. [PubMed: 25553874]

Rouse CH, Jones GE, Jones KR. The effect of body composition and gender on cardiac awareness. Psychophysiology. 1988; 25:400–407. [PubMed: 3174906]

Schandry R. Heart beat perception and emotional experience. Psychophysiology. 1981; 18:483–488. [PubMed: 7267933]

Vaitl D. Interoception. Biological Psychology. 1996; 42:1–27. [PubMed: 8770368]

Van Wye G, Kerker BD, Eisenhower D, Thorpe L, Chamany S, Matte T, Frieden TR. Obesity and diabetes in New York City. Preventing Chronic Disease. 2008:5.

Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology. 1988; 54:1063–1070. [PubMed: 3397865]

Werner NS, Jung K, Duschek S, Schandry R. Enhanced cardiac perception is associated with benefits in decision-making. Psychophysiology. 2009; 46:1123–1129. [PubMed: 19558399]

Werner NS, Peres I, Duschek S, Schandry R. Implicit memory for emotional words is modulated by cardiac perception. Biological Psychology. 2010; 85:370–376. [PubMed: 20813152]

Werner NS, Schweitzer N, Meindl T, Duschek S, Kambeitz J, Schandry R. Interoceptive awareness moderates neural activity during decision-making. Biological Psychology. 2013; 94:498–506. [PubMed: 24076035]

Whitehead WE, Drescher VM, Heiman P, Blackwell B. Relation of heart rate control to heartbeat perception. Biofeedback and Self-Regulation. 1977; 2:371–392.

Wickens, TD. Elementary Signal Detection Theory. New York, NY: Oxford University Press; 2002.

Wiens S. Interoception in emotional experience. Current Opinion in Neurology. 2005; 18:442–447. [PubMed: 16003122]

Wiens S, Mezzacappa ES, Katkin ES. Heartbeat detection and the experience of emotions. Cognition and Emotion. 2000; 14:417–427.

Wiens S, Palmer SN. Quadratic trend analysis and heartbeat detection. Biological Psychology. 2001; 58:159–175. [PubMed: 11600243]

Zhang Z. Monte Carlo based statistical power analysis for mediation models: methods and software. Behavioral Research Methods. 2014; 46:1184–1198.

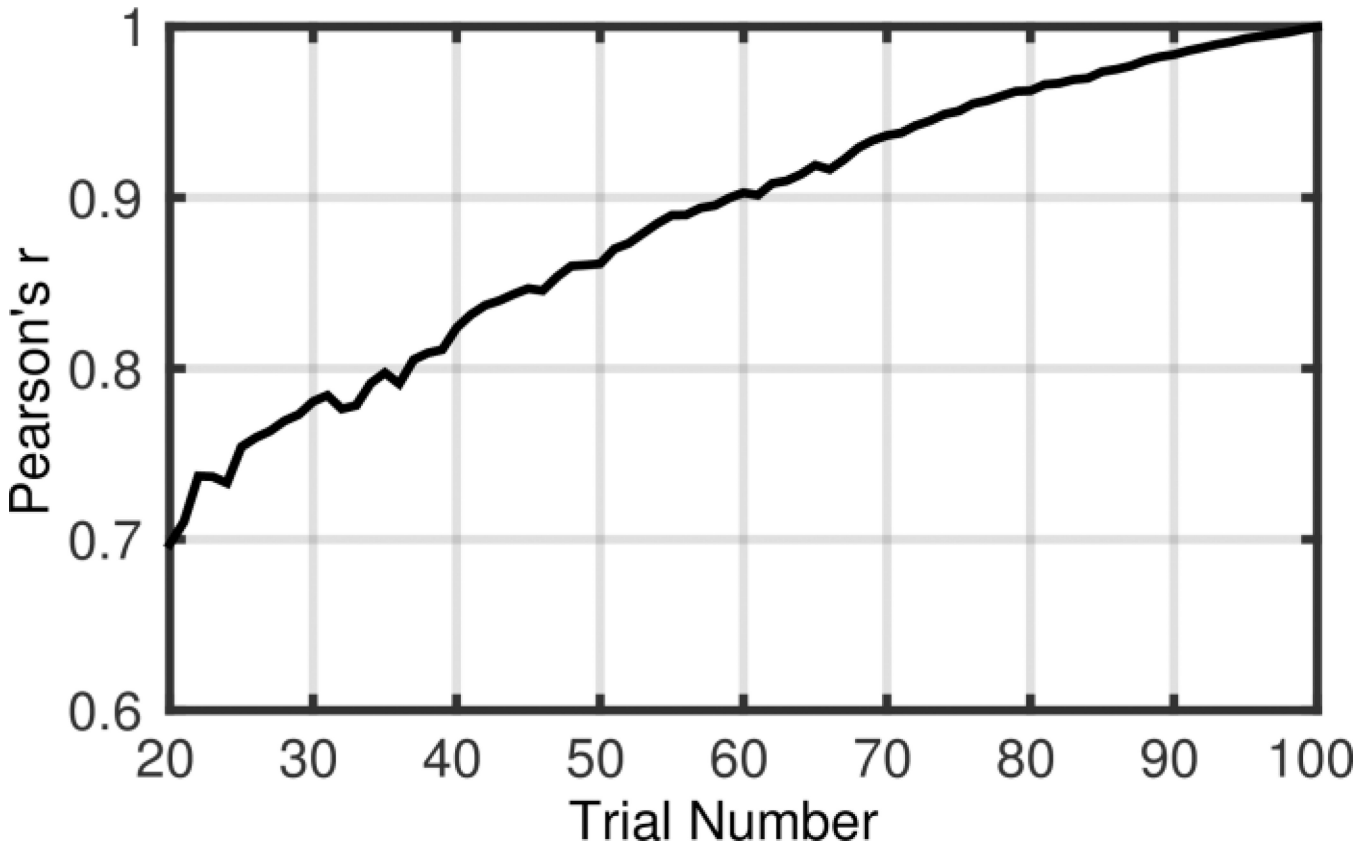

Figure 1.

This figure depicts a measure of reliability in interoceptive sensitivity, specifically, the between-participants correlations between accuracy at 100 trials and accuracy at 20–100 trials.

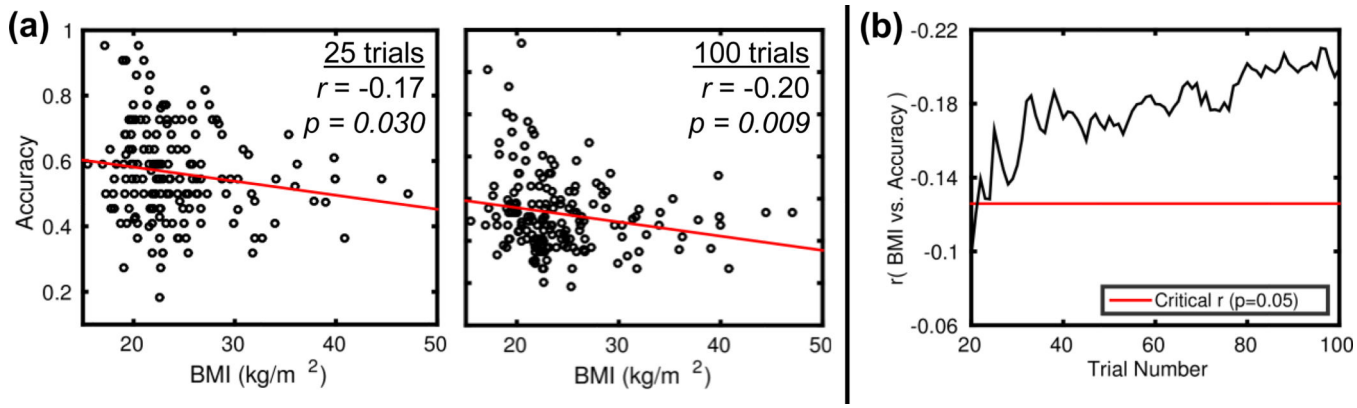

Figure 2.

(a) The correlation between cardiac interoceptive sensitivity and BMI at 25 trials and at 100 trials. (b) The observed correlation (i.e., effect size) between cardiac interoceptive sensitivity and BMI across trials. The correlation crossed the statistical threshold ($p < 0.05$) and remained near or above that threshold after 25–35 trials.

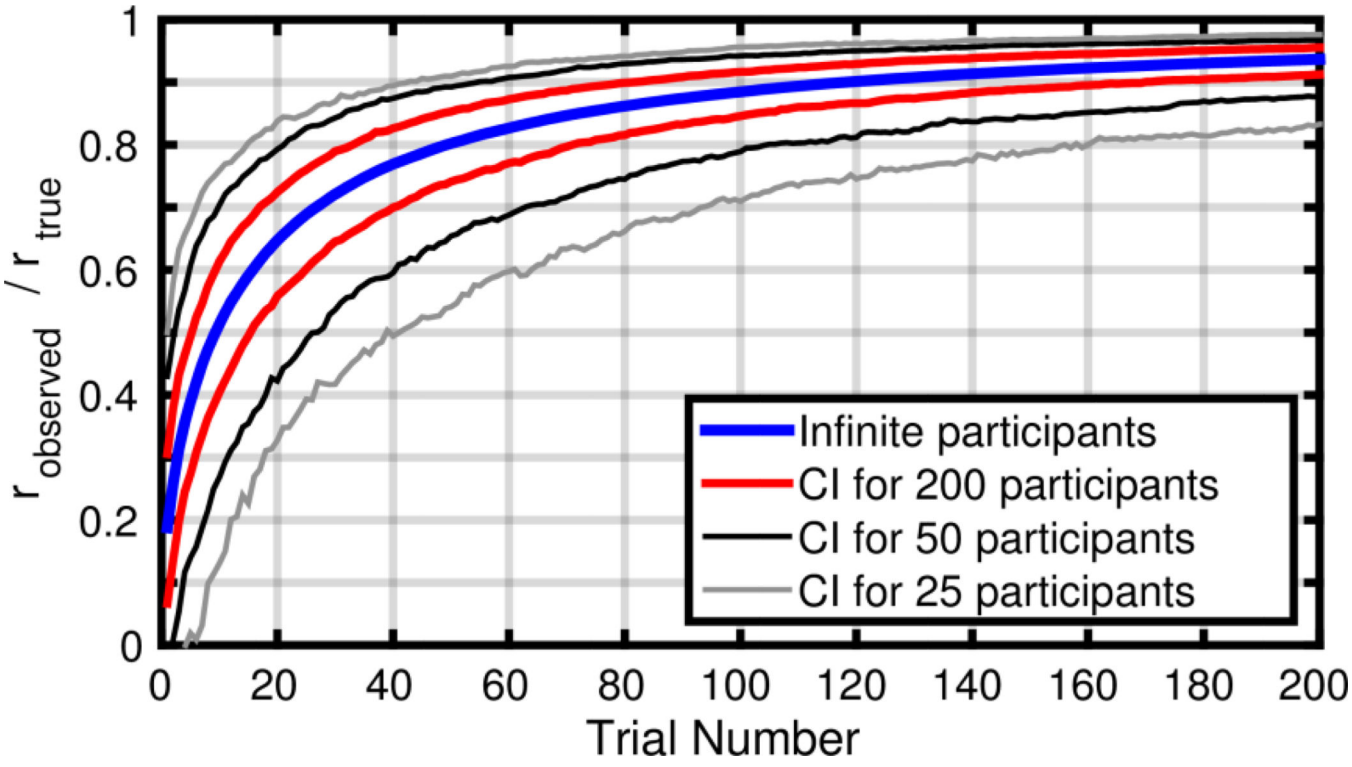

Figure 3.

The y-axis shows the ratio of the observed effect size ($r_{observed}$) to the true effect size (r_{true}). Each pair of lines with the same color shows the upper and lower 95% confidence intervals (CIs) encompassing variability in observed effect size due to finite sample size. In contrast, the "infinite participants" line does not contain any variability in $r_{observed}$ because of its infinite sample size. For a look-up table of values shown in the plot, see Table 1.

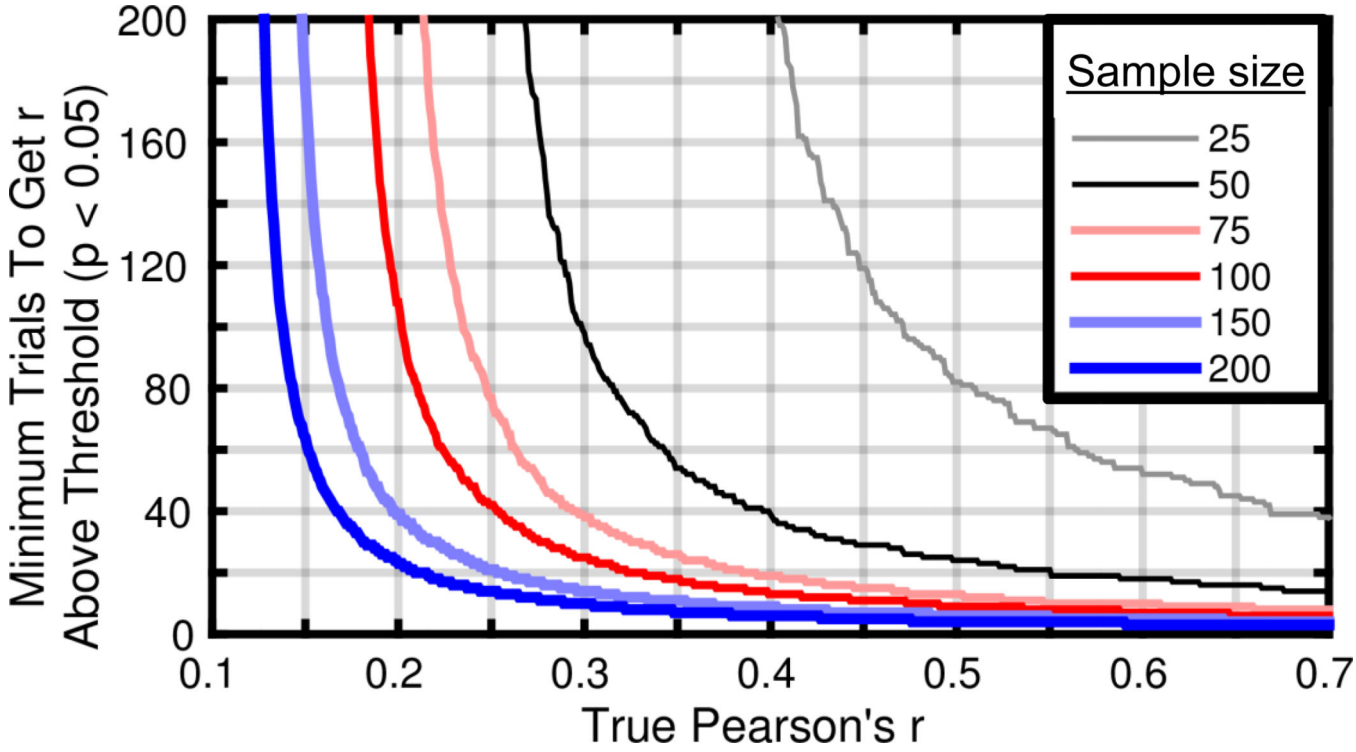

Figure 4.

The x-axis shows the true effect size (r_{true}). The y-axis shows the minimum number of trials such that r(*cardiac interoceptive sensitivity vs. Y*) exceeds the statistical threshold at $p < 0.05$ in 95% of the cases that the simulated correlation was performed under these conditions. The lines show the minimum trials as a function of r_{true} across different sample sizes (from 25 to 200 participants). For a look-up table of values shown in the plot, see Table 2.

Table 1

The proportion of true effect size determined using the number of heartbeat detection trials and the sample size N (values from Figure 3).

Number of Trials	Proportion of True Effect Size ($r_{observed}$ / r_{true})						
	N = 25	N = 50	N = 75	N = 100	N = 150	N = 200	N = Inf
20	33–84%	42–79%	50–78%	51–75%	53–74%	56–72%	65%
40	49–89%	60–87%	65–86%	66–85%	68–83%	70–83%	77%
60	60–93%	69–91%	73–90%	74–89%	75–88%	77–87%	83%
80	66–94%	74–93%	77–92%	79–91%	80–90%	82–90%	86%
100	71–96%	79–94%	81–93%	82–93%	83–92%	85–92%	88%
120	75–96%	81–95%	84–94%	85–94%	86–93%	87–93%	90%
140	77–97%	84–96%	86–95%	86–95%	88–94%	88–94%	91%
160	80–97%	85–96%	87–96%	88–95%	89–95%	56–72%	92%
180	81–97%	87–96%	89–96%	89–96%	90–95%	70–83%	93%
200	83–98%	88–97%	90–96%	90–96%	91–96%	77–87%	94%

Note. The range in each cell shows 95% confidence interval. "*N = Inf*" ignores the variability in $r_{observed}$ due to finite sample size.

Table 2

Minimum number of trials to reliably observe a between-participant correlation between cardiac interoceptive sensitivity and a second variable Y determined using effect size r_{true} and sample size N (values from Figure 4).

True Effect Size (r_{true})	Minimum Trials to Get Observed Effect Size ($r_{observed}$) Above Threshold					
	N = 25	N = 50	N = 75	N = 100	N = 150	N = 200
0.100	NA	NA	NA	NA	NA	NA
0.125	NA	NA	NA	NA	NA	268
0.150	NA	NA	NA	NA	179	64
0.175	NA	NA	NA	NA	67	35
0.200	NA	NA	NA	108	39	23
0.225	NA	NA	132	59	28	17
0.250	NA	NA	77	42	21	14
0.275	NA	167	51	31	17	12
0.300	NA	98	38	25	14	10
0.325	NA	72	30	20	12	8
0.350	NA	54	26	18	11	7
0.375	NA	45	22	15	9	7
0.400	216	39	19	13	9	6
0.425	155	32	17	12	8	6
0.450	119	29	15	11	7	5
0.475	96	26	13	10	7	5
0.500	82	24	13	9	6	4
0.525	76	22	11	9	6	4
0.550	67	21	10	8	5	4
0.575	57	19	10	8	5	4
0.600	52	18	10	7	5	3
0.625	50	17	9	7	5	3
0.650	45	16	9	6	4	3
0.675	39	15	8	6	4	3

True Effect Size (r_{true})	Minimum Trials to Get Observed Effect Size ($r_{observed}$) Above Threshold					
	N = 25	N = 50	N = 75	N = 100	N = 150	N = 200
0.700	38	14	8	6	4	3

Note. The true effect size refers to the best estimate of the effect size (using "infinite participants"; Table 1, Figure 3).

NA indicates that the sample size is too small to detect an effect of this size.