Journal: Biological Psychology

Running head: Heartbeat counting and heartbeat discrimination: a meta-analysis

Word count: 6926

Submission date: 06/04/2020; **Revision date:** 13/08/2020

The relationship between heartbeat counting and heartbeat discrimination: a meta-analysis.

Lydia Hickman[1*], Aida Seyedsalehi[2], Jennifer L Cook[1], Geoffrey Bird[3,4], Jennifer Murphy[5]

[1]School of Psychology, University of Birmingham

[2]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London

[3]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London

[4]Department of Experimental Psychology, University of Oxford

[5]Department of Psychology, Royal Holloway, University of London

[*]Corresponding author: LXH856@student.bham.ac.uk

School of Psychology,

University of Birmingham,

Edgbaston, Birmingham,

United Kingdom, B15 2TT

Abstract

Interoception concerns the perception of the body's internal state. Despite the importance of this ability for health and aspects of higher-order cognition, its measurement remains problematic. Most studies of interoception employ one of two tasks: the heartbeat counting or heartbeat discrimination task. These tasks are thought to index common abilities, an assertion often used to justify the use of a single measure of cardiac interoception. However, mixed findings regarding the relationship between performance on these tasks raises the question of whether they can be used interchangeably to assess interoceptive accuracy, confidence and awareness ('metacognition'). The present study employed a meta-analytical approach to assess the association between these tasks. Pooled findings from 22 studies revealed a small relationship between accuracy scores on the measures. Additional analyses demonstrated a moderate relationship between confidence ratings but no association between measures of interoceptive awareness. These findings question the interchangeable use of the two tasks.

Key words: Heartbeat counting; heartbeat discrimination; heartbeat detection; cardiac interoception; interoceptive accuracy

Introduction

In recent years the importance of interoception, the perception of the body's internal state (Craig, 2002, 2003, 2009), for health and higher-order cognition has begun to be appreciated (Barrett & Simmons, 2015; Khalsa et al., 2018; Murphy, Brewer, Catmur, & Bird, 2017). Indeed, numerous theoretical models posit a fundamental role for interoception in various aspects of health and cognition (Barrett & Simmons, 2015; Brewer, Cook, & Bird, 2016; Murphy et al., 2017; Paulus & Stein, 2006; Quattrocki & Friston, 2014). These models are supported by a growing body of evidence demonstrating links between interoception and fundamental cognitive abilities including learning and decision making (Werner, Jung, Duschek, & Schandry, 2009), emotional processing (Füstös, Gramann, Herbert, & Pollatos, 2013; Herbert, Pollatos, Flor, Enck, & Schandry, 2010; Schandry, 1981), and social cognition (Quattrocki & Friston, 2014; Seth, 2013). Furthermore, atypical interoception has also been observed across several mental health conditions, including Autism Spectrum Disorder (ASD; Garfinkel et al., 2016b), alexithymia (Brewer et al., 2016), depression (Harshaw, 2015; Pollatos, Traut-Mattausch, & Schandry, 2009), anxiety (Domschke, Stevens, Pfleiderer, & Gerlach, 2010; Pollatos et al., 2009), and eating disorders (Herbert & Pollatos, 2014; Klabunde, Acheson, Boutelle, Matthews, & Kaye, 2013; Pollatos et al., 2008) as well as physical health conditions such as obesity (Herbert & Pollatos, 2014) and diabetes (Pauli, Hartl, Marquardt, Stalmann, & Strian, 1991). Such evidence has led to suggestions that atypical interoception may represent a common risk factor for poor mental and physical health (Barrett & Simmons, 2015; Brewer et al., 2016; Murphy et al., 2017).

Increasing recognition of the importance of interoception for our understanding of pathology and cognition has prompted much research (Khalsa & Lapidus, 2016); however, progress in the field has been hampered by difficulties with the measurement of interoception (Brener & Ring, 2016; Murphy, Brewer, Hobson, Catmur, & Bird, 2018; Zamariola,

Maurage, Luminet, & Corneille, 2018). Indeed, whilst there are many aspects of interoception that may be quantified (e.g., the perception of respiratory, gastric or urinary signals; Khalsa et al., 2018), most studies of interoception have utilised one of two measures of cardiac interoceptive accuracy, the heartbeat counting task (HCT; Schandry, 1981) or the heartbeat discrimination[1] task (HDT; Katkin, Reed, & Deroo, 1983; Whitehead, Drescher, Heiman, & Blackwell, 1977). In the HCT, participants are asked to count the number of heartbeats they can feel during a series of time intervals (typically 3-6 intervals). Their response is compared to an objective record to determine accuracy. In the HDT, participants are required to determine whether an auditory or visual signal is presented synchronously or asynchronously with their heartbeat (typically across 15 to 60 trials). For the purposes of the present study it is relevant to note that the HDT can be administered in several variant forms (Brener & Ring, 2016), including the two-alternative forced choice procedure (2AFC; e.g., Whitehead et al., 1997), 6-alternative forced choice (Brener-Kluvitse) procedure (6AFC; e.g., Brener & Kluvitse, 1988) and the method of constant stimuli (MCS; e.g., Brener, Liu, & Ring, 1993; Yates, Jones, Marie, & Hogben, 1985). Whilst all of the above HDT variants require synchronicity judgements, they differ in terms of the delays at which the signal is presented with respect to the heartbeat and the analysis method used to determine accuracy (although notably moderate correlations have been observed between these HDT variants; > r = .50; Brener et al., 1993). Importantly, despite the existence of these variants and other tasks of cardiac interoceptive accuracy (e.g., heartbeat tapping, adjustment methods and perturbation methods; Carroll & Whellock, 1980; Gannon, 1980; Khalsa, Rudrauf, Sandesara, Olshansky, & Tranel, 2009; McFarland, 1975), it is the HCT and the 2AFC HDT

[1] Sometimes referred to as the 'heartbeat detection task' (e.g., Kleckner, Wormwood, Simmons, Barrett, & Quigley, 2015).

that are used most frequently, and interchangeably, as measures of cardiac interoceptive accuracy.

It is evident from the above descriptions that the HCT and HDT likely make different demands on cognitive processes. Indeed, whilst both presumably involve the perception of cardiac signals, the HCT requires sustained attention to heartbeat sensations over time whereas the HDT requires participants to integrate the cardiac signal with an external stimulus (Garfinkel, Seth, Barrett, Suzuki, & Critchley, 2015). Given these differences, it is perhaps unsurprising that several factors are thought to influence performance on the HCT and HDT selectively; for example, good performance on the HCT can be achieved through the use of non-interoceptive strategies. Indeed, better performance on the HCT has been associated with participants' beliefs regarding their resting heart rate (Brener & Ring, 2016; Ring & Brener, 1996; Ring, Brener, Knapp, & Mailloux, 2015; Windmann, Schonecke, Fröhlig, & Maldener, 1999) and their time estimation abilities (Murphy et al., 2018), factors that are unrelated to performance on the HDT (e.g., Knoll & Hodapp, 1992; Phillips, Jones, Rieger, & Snell, 2003). These dissociations suggest that different abilities may be quantified by the HCT and HDT and question the validity of interoceptive accuracy scores obtained from the HCT (Desmedt, Luminet, & Corneille, 2018; Zamariola et al., 2018; Murphy et al., 2018; but see Ainley, Tsakiris, Pollatos, Schulz, & Herbert, 2020).

The suggestion that the tasks may index slightly different abilities is supported by the differential impact of pathology on task performance; it is not always the case that the HCT and HDT exhibit the same patterns across clinical groups. For example, Hina and Aspell (2019) reported that non-smokers performed better on the HCT compared to smokers, but this difference was not seen for the 2AFC auditory HDT. Similar dissociations have been observed with other populations such as individuals with ASD (Garfinkel et al., 2016b) and hypermobile individuals (Mallorquí-Bagué et al., 2014). Additionally, Rae, Larsson,

Garfinkel, and Critchley (2019) reported a positive association between tic severity in Tourette syndrome and interoceptive accuracy as indexed by the 2AFC auditory HDT, but no such relationship was observed when interoceptive accuracy was indexed by the HCT. Such evidence again questions whether a common ability is quantified by these tasks of cardiac interoceptive accuracy and whether they can be used interchangeably, as one would expect a similar impact of pathology on task performance if the tasks index a common ability.

Despite indirect evidence suggestive of dissociations between performance on the HCT and HDT, studies directly comparing the two tasks are inconclusive regarding the presence or absence of a relationship; for example, early reports by Knoll and Hodapp (1992) suggested a moderate correlation (r = .59) between performance on the HCT and 2AFC auditory HDT. Similarly, other studies suggest a small but significant correlation between accuracy scores on the tasks (r = .36; Hart, McGowan, Minati, & Critchley, 2013). Such evidence of a small-to-moderate correlation between these measures is often used to justify the use of a single measure of cardiac interoceptive accuracy, as performance is presumed to generalise from one task to the other (e.g., Borhani, Ladavas, Fotopoulou, & Haggard, 2017; Herbert, Blechert, Hautzinger, Matthias, & Herbert, 2013; Pollatos, Traut-Mattausch, Schroeder, & Schandry, 2007; Scarpazza, Sellitto, & di Pellegrino, 2017; Werner et al., 2009). However, there are instances where performance on the HCT and HDT has not been found to correlate; for example, Forkmann et al. (2016) found no significant association between performance on the HCT and the 2AFC auditory HDT. This lack of an association was replicated by Schulz, Lass-Hennemann, Sutterlin, Schachinger, and Vogele (2013) who tested participants on the HCT and both the auditory and visual versions of the 2AFC HDT. Whilst a significant correlation was found between performance on the two versions of the HDT (r = .63; i.e. 39.7% of variance in one task is explained by the other), no relationship was found between the HCT and either version of the HDT. Finally, a study by Ring and

Brener (2018) which tested participants on the HCT and MCS auditory HDT also observed no significant association between performance on the two measures. It is clear that these inconsistent reports from single studies must be considered together before concluding whether there is a relationship between performance on the two tasks, and in turn whether they might index a common ability. Indeed, quantifying the relationship between these two tasks is important for determining whether the HCT and HDT can be used interchangeably as measures of cardiac interoceptive accuracy, and the generalisability of studies that have employed one task.

Thus far, we have focused on interoceptive accuracy, but there are other aspects of interoceptive ability that may be quantified using the HCT and HDT. In addition to accuracy it is now common for studies to obtain confidence ratings during tasks of interoceptive accuracy in order to assess both one's interoceptive sensibility (self-reported beliefs regarding interoceptive accuracy) and to calculate interoceptive awareness (a metacognitive measure reflecting the correspondence between interoceptive accuracy and interoceptive sensibility; Garfinkel et al., 2015; Murphy, Catmur, & Bird, 2019b). For both tasks, interoceptive sensibility is calculated by averaging the confidence ratings obtained across trials. However, it is notable that there are differences in the assessment of interoceptive sensibility and awareness for the HDT and HCT; for example, 1) far fewer trials are used for the HCT (typically 3-6) compared to the HDT (typically 15-60) thus reducing the reliability of the HCT accuracy, sensibility and awareness indices, and 2) the analysis strategy for calculating interoceptive awareness differs for the HCT and HDT. Whilst HDT interoceptive awareness is usually calculated using Receiver Operating Characteristic (ROC) curves (but see Palser, Fotopoulou, Pellicano, and Kilner (2018) for an alternative method for calculating HDT interoceptive awareness), confidence-accuracy correlations are generally used to calculate interoceptive awareness for the HCT (but see Murphy et al. (2020) for an alternative

scoring method for calculating HCT interoceptive awareness). In terms of the relationship between these aspects of interoception, confidence ratings for the HCT and HDT (indexing interoceptive sensibility) are often correlated with one another, but the strength of this association has been found to vary substantially across studies, with Forkmann et al. (2016) reporting a relatively low correlation ($r = 0.348$) and Garfinkel et al. (2015) reporting a much stronger correlation ($r = 0.711$). Conversely, awareness scores obtained using these two tasks have not been found to correlate (Forkmann et al., 2016; Garfinkel et al., 2015) and show different relationships across pathologies. For example, Ewing et al. (2017) found that interoceptive awareness on the HCT was predicted by an interaction between sleep effectiveness and mixed anxiety and depressive disorder, but no such relationship was observed for HDT interoceptive awareness. With increasing interest in these aspects of interoception (Forkmann et al., 2016; Garfinkel et al., 2015), understanding the generalisability of interoceptive sensibility and awareness scores calculated using the HCT and HDT is a priority.

It is clear from the above review that questions exist as to the relationship between the HCT and HDT, which has implications for whether they can be considered to be testing the same ability (or set of abilities). Lack of clarity regarding the relationship between these measures is potentially problematic for cases where only one task is utilised as a measure of cardiac interoception, or where both tasks are employed but show differential relationships with a third variable. As such, in this study we investigate the relationship between the HCT and HDT in order to clarify the extent to which using these measures interchangeably should be a concern. Specifically, evidence from studies that utilised both the HCT and HDT was collated to determine the relationships between accuracy, confidence and awareness scores obtained using the two different tasks. This was achieved by employing a meta-analytical strategy to obtain the pooled effect sizes of the reported correlations.

Methods

Search strategy

A systematic literature search was conducted in PubMed, Web of Science, PsycINFO and Medline. All searches were restricted to the year 1976 onwards, 2 years prior to the first description of the HCT (Dale & Anderson, 1978). All searches were conducted on the 28th October 2019. The following search was employed across the 4 search engines:

("interoceptive sensitivity" OR "interoceptive accuracy" OR "heartbeat perception" OR "heartbeat interoception" OR "cardiac perception" OR "cardiac interoception" OR "cardioception" OR "cardioceptive" OR "cardiac awareness" OR (("heartbeat tracking" OR "heartbeat counting" OR "Schandry") AND ("heartbeat discrimination" OR "heartbeat detection" OR "Whitehead"))).

The search terms were designed to ensure that articles mentioning concepts relating to interoceptive accuracy, or that used both tasks, would be identified. The terms "Schandry" and "Whitehead" were included to identify studies using the authors names to refer to the HCT and HDT respectively (Schandry, 1981; Whitehead et al., 1977). This search returned 1583 results. Of these, 410 were from PubMed, 543 from Web of Science, 369 from PsycINFO and 261 from Medline. Following the removal of 922 duplicates, 661 articles remained.

Following a Reviewer's recommendation, an additional search was conducted across all 4 search engines, replacing 'heartbeat' with 'heart beat' in the original search. This search yielded 11 further results from the specified time period.

Study selection

The remaining articles were screened in two phases by two researchers. First, the titles and abstracts were assessed to identify whether the content of the article was relevant to the meta-analysis, with one researcher conducting a light screening and another researcher

conducting a more thorough screening. If a paper was deemed relevant, the full text was then examined by a researcher to identify whether the article should be included. Ambiguous cases were discussed by the two researchers. The initial title and abstract screening process removed a total of 470 articles. Removed articles were either not written in English, not peer reviewed, did not present empirical data (e.g., review articles) or were deemed not relevant (e.g., they did not focus on cardiac interoception). The full text screening stage resulted in the removal of a further 177 articles. Of these articles, 145 were removed as they only utilised one task, 27 did not employ either task, 3 did not assess cardiac interoception, 1 did not present empirical data, and 1 was determined not to be a measure of interoceptive accuracy as the participants were permitted to feel for their pulse during the tasks.

A total of 25 articles utilised both the HCT and HDT, with 18 reporting the correlation between HCT accuracy and HDT accuracy in the paper or providing open access data which enabled the calculation of this correlation. The authors of the remaining 7 articles were contacted for the correlation statistics for accuracy, confidence and awareness (where applicable), with data available for 4 of the 7 aforementioned papers. Consequently, data from 22 articles were used in the analysis assessing the relationship between accuracy scores on the HCT and HDT (hereafter 'accuracy analysis'). For the assessment of the relationship between confidence ratings on the HCT and HDT (hereafter 'confidence analysis'[2]), data from 7 of the 25 articles were used. Of the 18 studies excluded from analyses, 17 did not include confidence ratings and data were unavailable from 1 of the 4 authors contacted who did not report the correlation in the article. For the analysis of the relationship between awareness scores on the HCT and HDT (hereafter 'awareness analysis'), a total of 6 of the 25 articles were included. Of the excluded articles, 17 did not measure awareness and data were

[2] The 'confidence analysis' is named as such for clarity due to differences in the literature with regards to the naming structure of interoceptive abilities and the frequent use of the term 'sensibility' to relate to both confidence ratings and questionnaires of interoceptive sensibility which may index different abilities (Murphy et al., 2019b).

unavailable from 2 of the 3 authors contacted who did not report the correlation in the article. The process of article selection is displayed in Figure 1, following the Preferred Reporting Items for Systematic review and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009).

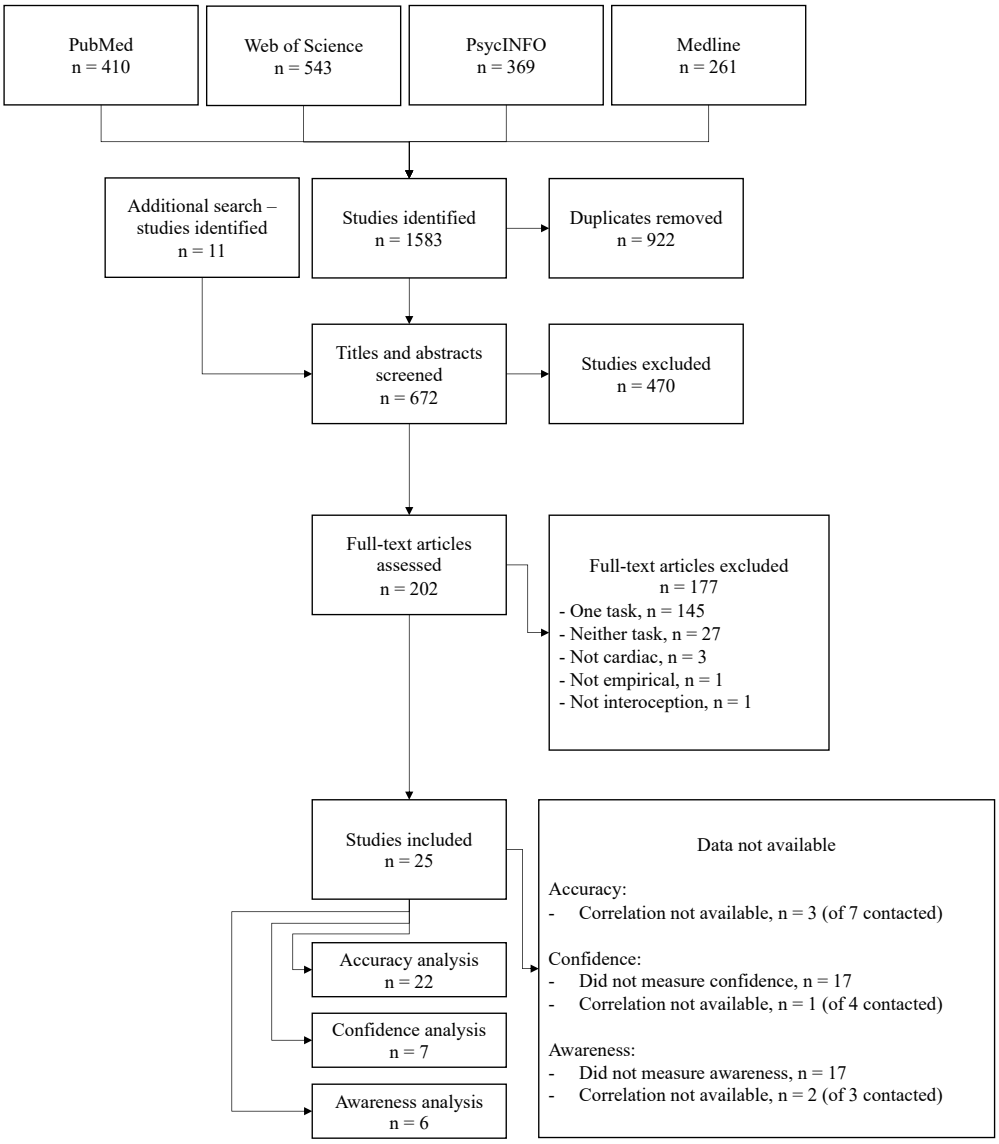

Figure 1. A PRISMA flowchart depicting the screening procedure employed for the meta-analyses. "Additional search – studies identified" refers to the newly identified studies resulting from a Reviewer's suggestion to replace 'heartbeat' with 'heart beat' in the original search. Overall, 22 studies were identified for the accuracy analysis, 7 for the confidence analysis and 6 for the awareness analysis.

Data extraction

Relevant data for the meta-analysis, including details of the experimental design, implementation of tasks and the correlation statistics between tasks, were extracted and subsequently checked by a researcher. The data extracted for each of the studies is presented in Tables 1-4.

Author	Participants*	Age	Gender	Counterbalancing	Device
Betka (2018)	Heavy alcohol users (n=32)	M=25.1	0F	Not reported	PO
Ewing (2017)	MH diagnoses (n=138) and controls (n=42)	MH: M=34.21, SD = 14.25; Controls: M=28.2, SD = 9.8	MH: 92F, 1 other, 2 undisclosed; Controls: 34 F	HCT before HDT	PO
Forkmann (2016)	Typical population (n=159)	M=23.9, SD=3.3	118F	Yes	ECG
Garfinkel (2015)	Typical population (n=80)	M=25.1, SD=4.44	30F	Not reported	PO
Garfinkel (2016)	ASD (n=20) and controls (n=20)	ASD: M=28.06, SD=8.8; Controls: M=27.81, SD=3.4	ASD: 2F; Controls: 2F	HCT before HDT	PO
Hart (2013)	BPD (n=24) and controls (n=30)	BPD: M=37, SD=11; Controls: M=31, SD=5	BPD: 21F; Controls: 24F	Yes	PO
Herman (2019)	Typical population (n=60)	M=22.33, SD=3.75	44F	HCT before HDT	PO
Hina (2019)	Smokers (n=48) and non-smokers (n=51)	M=25.67, SD=8.71	Smokers: 28F; Non-smokers: 32F	Not reported	ECG
Kandasamy (2016)	Traders (n=18)	Not reported	0F	HCT before HDT	PO
Knoll (1992)	Typical population (n=59)	M=22.3	64F	HCT before HDT	ECG
Leganes-Fonteneau (2019)	Typical population (n=50)	M=21.8, SD=3.88	30F	HCT before HDT	PO
Michal (2014)	DPD (n=24) and controls (n=24)	DPD: M=27.8, SD=7.5; Controls: M=26.4, SD=1.6	DPD: 11F; Controls: 12F	Yes	ECG
Mul (2018)	ASD (n=26) and controls (n=26)	ASD: M=25.4, SD=7.3; Controls: M=25.4, SD=7.6	ASD: 7F; Controls: 7F	Yes	ECG
Palser (2018)	ASD (n=30) and controls (n=45)	ASD: M=12.5, SD=2.88; Controls: M=11.26, SD=3.16	ASD: 5F; Controls 22F	HCT before HDT	PO
Rae (2019)	TS (n=21) and controls (n=22)	TS: M=34; Controls: M=34 (SD not reported)	TS: 9F; Controls: 10F	HCT before HDT	PO
Ring (2018)	Typical population (n=48)	M=18.69, SD=0.78	30F	Yes	ECG
Schaefer (2012)	SFD (n=23) and controls (n=27)	SFD: M=45.26, SD=13.57; Controls: M=41.74, SD=12.52	SFD: 16F; Controls: 16F	HCT before HDT	ECG
Schroeder (2015)	NCCP (n=42), CPP (n=36) and controls (n=52)	NCCP: M=51.7, SD=10.5; CCP: M=59.4, SD=9.2; Controls: M=49.1, SD=9.6	NCCP: 20F; CCP: 9F; Controls: 33F	HCT before HDT	ECG
Schulz (2013)	Cold pressor group (n=21) and controls (n=21)	Cold pressor group: M=23.2, SD=2.6; Controls: M=22.7, SD=2.5	Cold pressor group: 15F; Controls: 14F	Yes	ECG
Villani (2019)	Typical population (n=51)	M=21.1, SD=3.1	34F	HCT before HDT	ECG
Weitkunat (1996)	Panic patients (n=9) and controls (n=20)	Panic patients: M=35.8, SD=7.7; Controls: M=25.9, SD=5.3	Panic patients: 4F; Controls: 10F	HCT before HDT	ECG
Wittkamp (2018)	Typical population (n=60)	M=23.4, SD=3.5	40F	Yes	ECG

Table 1. Demographics and general task administration. MH = Mental Health, ASD = Autism Spectrum Disorder, BPD = Borderline Personality Disorder, DPD = Depersonalization Disorder, TS = Tourette's Syndrome, SFD = Somatoform Disorders, NCCP = Noncardiac Chest Pain, CCP = Cardiac Chest Pain, M = Mean, SD = Standard Deviation, F = Female, HCT = Heartbeat Counting Task, HDT = Heartbeat Discrimination Task, PO = Pulse Oximeter, ECG = Electrocardiogram. *See forest plot for post-exclusion sample sizes

Author	Cue	HDT Trials	HDT Scoring	Number of stimuli	Type	Delays	Confidence Scale	Awareness Measure
Betka (2018)	Auditory	20	Accuracy	10	2AFC	S="rising edge of finger pulse pressure wave", A=300 ms later	-	-
Ewing (2017)	Auditory	20	d prime	10	2AFC	S="rising edge of finger pulse pressure wave", A=300 ms later	VAS	ROC
Forkmann (2016)	Auditory	40	d prime	6	2AFC	S=230ms, A=530ms	Numerical scale (0-8)	ROC
Garfinkel (2015)	Auditory	15	Not provided	10	2AFC	S=250 ms, A=550 ms	VAS	ROC
Garfinkel (2016)	Auditory	15	Accuracy	10	2AFC	S=250 ms, A=550 ms	VAS	ROC
Hart (2013)	Auditory	50	Accuracy	10	2AFC	S=250 ms, A=550 ms	-	-
Herman (2019)	Auditory	20	Accuracy	10	2AFC	S="rising edge of finger pulse pressure wave", A=300 ms later	VAS	ROC
Hina (2019)	Auditory	16	Accuracy & d prime	20	2AFC	S="R-wave of QRS complex", A="80% or 120% of the speed of the two preceding R-wave"	-	-
Kandasamy (2016)	Auditory	15	Accuracy	10	2AFC	S="rising edge of finger pulse pressure wave", A=300 ms later	-	-
Knoll (1992)	Auditory	90	2 . arcsin (sqrt(P(A))	8	2AFC	S="after 1/4 of estimated duration of IBI", A="after 3/4 of estimated duration of IBI"	-	-
Leganes-Fonteneau (2019)	Auditory	20	Accuracy	10	2AFC	S = heartbeat, A = 300 ms later	VAS	ROC
Michal (2014)	Auditory	20	d prime	10	2AFC	S=230ms, A=530ms	-	-
Mul (2018)	Auditory	8	Accuracy	20	2AFC	S="R-wave of QRS complex", A="80% or 120% of the speed of the two preceding R-wave"	-	-
Palser (2018)	Auditory	10	Accuracy	10	2AFC	S=250 ms, A=550 ms	Numerical scale (1-5)	ANOVA analysis
Rae (2019)	Auditory	20	Accuracy	10	2AFC	S="rising edge of finger pulse pressure wave", A=300 ms later	VAS	ROC
Ring (2018)	Auditory	120	IQR of distribution of simultaneous judgements across 6 intervals	10	MCS	6 delays: 0, 100, 200, 300, 400, 500 ms	-	-
Schaefer (2012)	Auditory	60	d prime	10	2AFC	S=200 ms, A=500 ms	-	-
Schroeder (2015)	Auditory	60	d prime	10	2AFC	S=determined manually by participants prior to testing (or 200ms if participants unsure), A=S delay + IBI/2	-	-
Schulz (2013)	Auditory and Visual	20	d prime	6	2AFC	S=230ms, A=530ms	-	-
Villani (2019)	Auditory	50	Accuracy	10	2AFC	S=200 ms, A=500 ms	VAS	-
Weitkunat (1996)	Auditory	100	2 . arcsin (sqrt(P(A))	10	2AFC	S=130ms, A=N + 30i ms after wave peak (n in 0:200 (random), i in 1:10 random heartbeats)	-	-
Wittkamp (2018)	Visual	40	d prime	6	2AFC	S=230ms, A=530ms	-	-

Table 2. Heartbeat discrimination task administration and scoring. IQR = Inter-Quartile Range, AFC = Alternative Forced Choice, MCS = Method of Constant Stimuli, S = Synchronous, A = Asynchronous, IBI = Interbeat Interval, VAS = Visual Analogue Scale (from total guess/no heartbeat awareness to complete confidence/full perception of heartbeat), ROC = Receiver Operating Characteristic

Author	HCT trials	HCT scoring	Time Intervals	Counterbalancing	Confidence Scale	Awareness Measure
Betka (2018)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	-	-
Ewing (2017)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	Pearson's
Forkmann (2016)	3, 4, or 6	Schandry & Hart	30, 45, 60 s; 25, 35, 45, 55 s; 25, 35, 45, 55, 65, 75 s	Yes	Numerical scale (0-8)	Pearson's
Garfinkel (2015)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	Pearson's
Garfinkel (2016)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	-
Hart (2013)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	-	-
Herman (2019)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	Pearson's
Hina (2019)	4	Schandry	25, 35, 45, 55 s	Yes	-	-
Kandasamy (2016)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	-
Knoll (1992)	3	Schandry	26, 21, 36 s	Not reported	-	-
Leganes-Fonteneau (2019)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	Pearson’s
Michal (2014)	7	Schandry	20, 25, 35, 45, 55, 65, 75 s	Yes	-	-
Mul (2018)	4	Schandry	25, 35, 45, 55 s	Yes	-	-
Palser (2018)	6	Hart	25, 30, 35, 40, 45, 60 s	Yes	Numerical scale (1-5)	-
Rae (2019)	6	Hart	25, 30, 35, 40, 45, 50 s	Yes	VAS	Pearson's
Ring (2018)	3	Schandry	25, 35, 45 s	Not reported	-	-
Schaefer (2012)	3	Schandry	25, 35, 45 s	Not reported	-	-
Schroeder (2015)	3	Error Score*	25, 35, 45 s	No	-	-
Schulz (2013)	3	Schandry	30, 45, 60 s	Yes	-	-
Villani (2019)	6	Schandry	21s, 25s, 33s, 47s, 55s, 74 s	Yes	-	-
Weitkunat (1996)	3	Error Score*	35, 25, 45 s	No	-	-
Wittkamp (2018)	3	Schandry	35, 45, 55 s	Yes	-	-

Table 3. Heartbeat counting task administration and scoring: VAS = Visual Analogue Scale (from total guess/no heartbeat awareness to complete confidence/full perception of heartbeat), Schandry = 1/n[1-((|objective-subjective|)/objective)], Hart = 1/n[1-((|objective-subjective|)/((|objective+subjective|)/2))]. * r value reversed in calculation due to the use of an error score for the HCT.

Author	Correlation Type	Accuracy correlation in total sample	Accuracy correlation in subsamples	Confidence correlation	Awareness correlation
Betka (2018)	Pearson's	$r = 0.071$, $p > .05$	-	-	-
Ewing (2017)	Not reported	$r = 0.15$, $p = .049$	Not reported	$r = 0.584$, $p = .000$	$r = 0.086$, $p = .262$
Forkmann (2016)	Not reported	$r = 0.072$, $p = .691$	-	$r = 0.348$, $p = .047$	$r = -0.160$, $p = .399$
Garfinkel (2015)	Pearson's	$r = 0.316$, $p = .004$	-	$r = 0.711$, $p = .000$	$r = 0.103$, $p = .362$
Garfinkel (2016)	Pearson's	$r = 0.36$, $p = .021$	Not reported	Not available	Not available
Hart (2013)	Not reported	$r = 0.36$, $p = .008$	Not reported	-	-
Herman (2019)	Pearson's	$r = 0.098$, $p = .462$	-	$r = 0.652$, $p < .001$	$r = 0.144$, $p = .276$
Hina (2019)	Spearman	$r = 0.278$, $p < .05$	Not reported	-	-
Kandasamy (2016)	Not reported	$r = 0.318$, $p = .198$	-	-	-
Knoll (1992)	Spearman	$r = 0.59$, $p < .001$	-	-	-
Leganes-Fonteneau (2019)	Pearson's	$r = 0.153$, $p = .288$	-	$r = 0.520$, $p < .001$	$r = 0.140$, $p = .333$
Michal (2014)	Not reported	Not reported	DPD: $r = 0.102$, $p > .05$; Controls: $r = 0.332$, $p > .05$	-	-
Mul (2018)	Spearman	$r = 0.10$, $p = .49$	Not reported	-	-
Palser (2018)	Spearman	$r = -0.114$, $p = .328$	ASD: $r = -0.172$, $p = .363$; Controls: $r = -0.043$, $p = .779$	$r = 0.473$, $p < .001$	-
Rae (2019)	Pearson's	$r = 0.327$, $p = .033$	Not reported	$r = 0.784$, $p = .000$	$r = 0.114$, $p = .468$
Ring (2018)	Pearson's	$r = -0.04$, $p = .77$	-	-	-
Schaefer (2012)	Not reported	$r = 0.43$, $p < .01$	SFD: $r = 0.37$, $p = .09$; Controls: $r = 0.50$, $p < .01$	-	-
Schroeder (2015)	Spearman	$r = -0.284$, $p = .001$	Not reported	-	-
Schulz (2013)	Pearson's	Auditory: $r = 0.22$, $p = .15$; Visual: $r = 0.08$, $p = .60$	Not reported	-	-
Villani (2019)	Not reported	Not reported	Sham: $r = 0.068$, $p = .65$	-	-
Weitkunat (1996)	Pearson's	$r = -0.061$, $p > .05$	Not reported	-	-
Wittkamp (2018)	Not reported	$r = 0.26$, $p < .05$	-	-	-

Table 4. Correlation coefficients and associated p values for accuracy, confidence and awareness. DPD = Depersonalization Disorder, ASD = Autism Spectrum Disorder, SFD = Somatoform Disorders.

Meta-analyses

The primary aim of this paper was to quantify the effect size of the correlation between accuracy scores obtained via the HCT and HDT. As such, the correlation coefficients for the relationship between HCT and HDT accuracy scores extracted from the included papers were used to obtain a pooled effect size. The data were analysed using R with the packages meta (Schwarzer, 2007) and dmetar (Harrer, Cuijpers, Furukawa & Ebert, 2019). There was a total of 23 correlation coefficients pooled in the meta-analysis due to Michal et al. (2014) reporting separate statistics for the two groups within their study. In cases where one of the two scores related to error as opposed to accuracy (e.g., Schroeder, Gerlach, Achenbach, & Martin, 2015; Weitkunat, 1996), the sign of the correlation coefficient was reversed prior to running the meta-analysis. It should also be noted that one paper (Schulz et al., 2013) investigated the relationship between HCT and HDT accuracy using both visual and auditory versions of the HDT. As these data were from the same participants and their respective correlation coefficients with HCT accuracy did not significantly differ as determined by a Fisher r-to-z transformation (p = .526), we included only the auditory version of the HDT at it is used more frequently in the literature. However, to ensure that the version selected did not alter the pattern of results obtained, the first meta-analysis was re-run to check whether replacing the auditory HDT-HCT correlation with the visual HDT-HCT correlation changed the pooled effect size. This did not alter the pattern of results observed (see "primary meta-analysis: accuracy"). Heterogeneity of the dataset was investigated using the Q statistic, which is calculated by summing the weighted squared differences between each study's observed effect size and the fixed-effect estimate, and compared to a null hypothesis of homogeneity. The I^2 statistic, considered complementary to the Q statistic (Huedo-Medina, Sánchez-Meca, Marín-Martínez & Botella, 2006), was also calculated. This statistic indexes the percentage of effect size variability not caused by

sampling error, with values of 25%, 50% and 75% indicating low, moderate and high heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003). In accordance with recommendations in the field (Field, 2001; Hunter & Schmidt, 2000), the random-effects model was followed due to the likelihood of significant heterogeneity within the results of the studies and the Sidik-Jonkman estimator was used to assess between-study heterogeneity within the model ($\tau 2$; Sidik & Jonkman, 2007). The pooled effect size was generated using inverse variance weighting, with a Fisher's z-transformation to obtain accurate weights. Finally, a publication bias analysis was conducted to assess whether null or weak results had been excluded from publication within the interoception literature. A funnel plot was produced to enable inspection of the relationship between the standard errors and effect sizes, and Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997) was used to test asymmetry of the funnel plot. For the 6 and 7 studies reporting the correlation between HCT and HDT awareness and confidence ratings, respectively, two further meta-analyses were conducted using the methods described above in order to obtain a pooled effect size for the relationship between the scores.

Analysis scripts, data, full screening details and a PRISMA checklist are available online at https://osf.io/a32n9/.

Results

Primary meta-analysis: Accuracy

Using all data obtained from the 22 selected studies (23 correlation coefficients), we employed the above analysis to uncover the pooled effect size of the relationship between accuracy as measured by the HCT and HDT. A significant Q statistic ($Q = 41.47$, $p = .007$) and an I^2 value of 47.0% supported the use of a random-effects model meta-analysis. The meta-analysis identified a pooled effect size of 0.21 ($p < .001$). Thus, with an R^2 value of 0.044, 4.4% of the variance in accuracy on one measure was explained by accuracy on the

other. The individual effect sizes from each study and the pooled effect size are displayed in Figure 2. Inspection of the funnel plot (Figure 3) failed to indicate publication bias, and this was further supported by a non-significant Egger's test (p = .676). A power analysis using GPower (Erdfelder, Faul, & Buchner, 1996) determined that future studies would require 173 participants to find this pooled effect size of 0.21 with ~80% power. As described above, we re-ran the meta-analysis substituting the Schulz et al. (2013) auditory HDT statistics for the visual HDT statistics. The meta-analysis returned a pooled effect size of 0.20 (p < .001), which did not significantly differ from the previous pooled effect size as evidenced by the overlapping confidence intervals (Auditory: 0.21, CI [0.13, 0.29]; Visual: 0.20, CI [0.12, 0.28]).

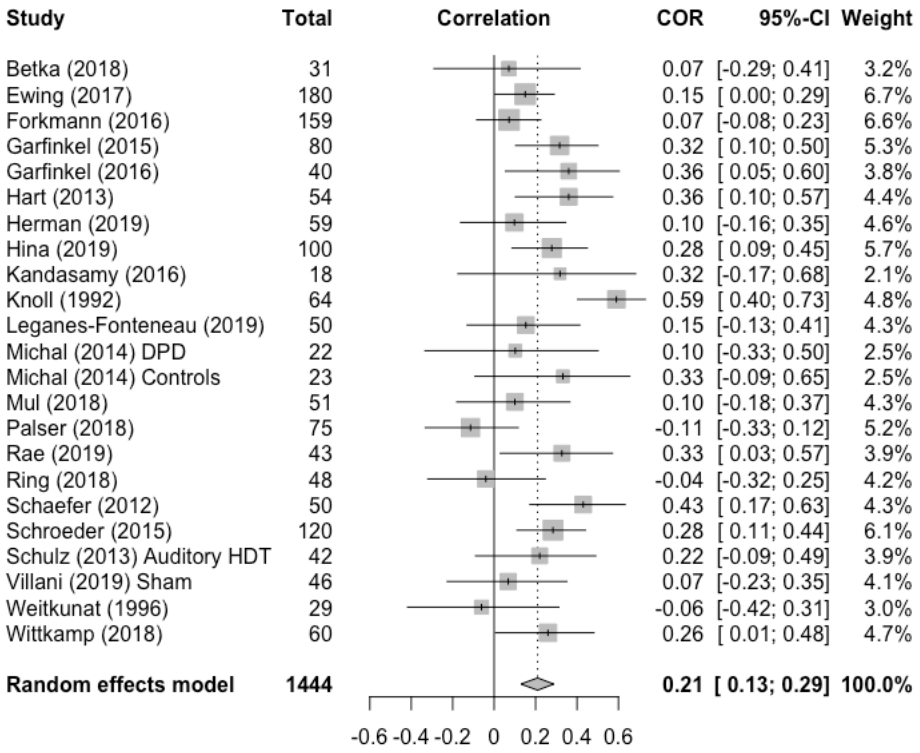

Figure 2. A forest plot displaying the individual effect sizes from each study in addition to the pooled effect size (dashed line) of the accuracy meta-analysis. As can be seen, the

random-effects model produced a pooled effect size of 0.21. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, DPD = Depersonalization Disorder, HDT = heartbeat discrimination task, Sham = data from the sham transcutaneous vagus nerve stimulation (taVNS) condition as opposed to the active condition.

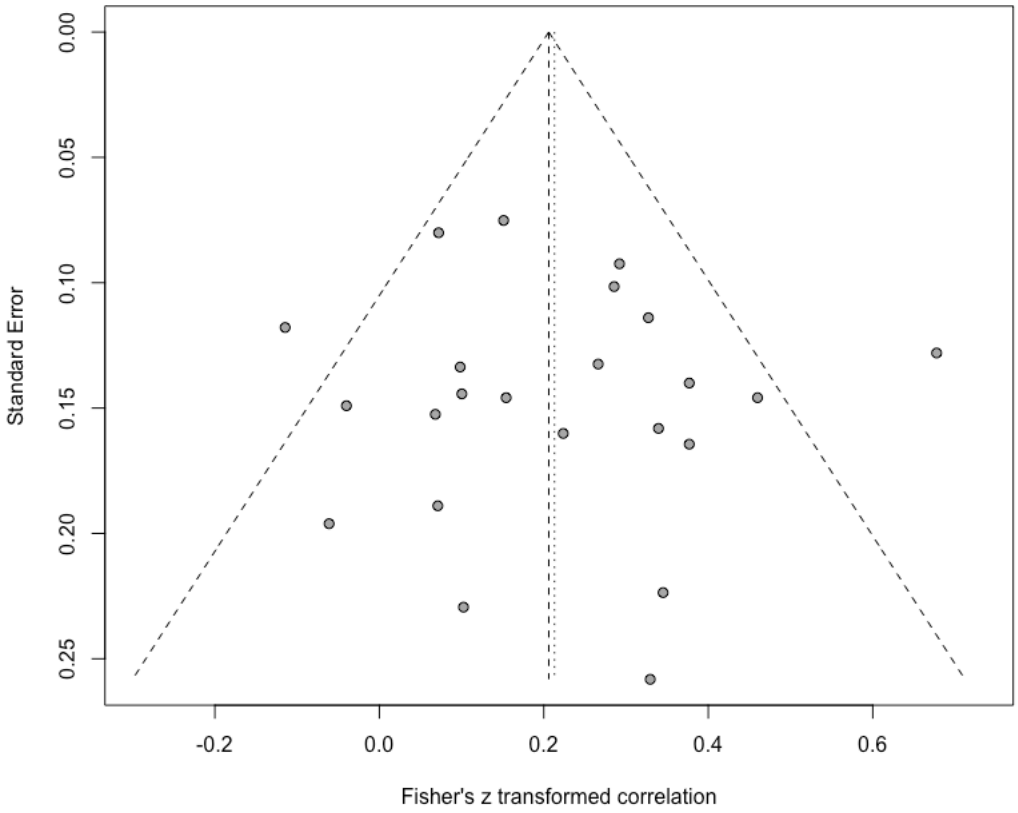

Figure 3. A funnel plot for the accuracy meta-analysis. No evidence of publication bias was observed.

As can be seen in Figure 2, the correlation coefficients reported in Knoll and Hodapp (1992) and Palser et al. (2018) were identified as outliers as they lie beyond the 95% confidence intervals of the pooled effect size (by 0.11 and 0.01 respectively). As such, the meta-analysis was run again to observe the effect of removing these correlation coefficients. This yielded a similar result of 0.20 for the pooled effect size ($p < .001$). In addition, the significance of the Q statistic was reduced and became non-significant ($Q = 20.48$, $p = .428$)

and the I^2 value was lowered to 2.3%, indicating that heterogeneity was reduced following the removal of these individual effect sizes.

Given heterogeneity in terms of the populations examined and the methods used (see Tables 1-4), exploratory analyses were conducted to assess the extent to which these differences may alter the above results (for details see Supplementary Materials [S1-S19]). In these exploratory meta-analyses, correlation coefficients were excluded based on the use of clinical and/or atypical populations (n = 11), the use of the MCS version of the HDT (n = 1; as all other included studies utilised the 2AFC HDT), the use of the visual version of the HDT (n = 1; as most other studies utilised the auditory HDT), or the use of fewer than 40 trials for the HDT (n = 14). Four further meta-analyses are also reported following 1) the separation of studies based on the device used to record heartbeats (pulse oximeter (n = 10) or electrocardiogram (ECG; n = 13)) given recent evidence that this may influence accuracy scores obtained using the HCT (Murphy et al., 2019a), and 2) the order in which the tasks were completed (HCT first (n = 13) or counterbalanced (n = 8), given that completion of the HDT prior to the HCT may provide participants with information regarding their resting heartbeat. The results of these meta-analyses ranged from a pooled effect size of 0.15 to 0.24 (all $p <= .007$; all with overlapping confidence intervals), thus this exploration of the data did not substantially change the results of the analysis.

Secondary meta-analyses

Confidence meta-analysis

Using the data obtained from the 7 studies reporting confidence ratings for both the HCT and HDT, a further meta-analysis was conducted to assess the pooled effect size of the relationship between these ratings. Heterogeneity tests revealed a significant Q statistic ($Q = 14.95$, $p = .021$) and a moderate I^2 value of 59.9%. A pooled effect size of 0.60 ($p < .001$) was identified by the meta-analysis, which is displayed in Figure 4 alongside the individual

effect sizes from each study included in the analysis. The resultant R^2 value of 0.36 indicates that 36.0% of the variance in confidence ratings on one measure was explained by confidence ratings on the other. No outliers were identified in this analysis as all of the individual effect sizes were seen to lie within the 95% confidence intervals of the pooled effect size. Figure 5 displays the funnel plot produced to assess potential publication bias. The Egger's value for this plot was non-significant ($p > .999$), which is consistent with the symmetric appearance of the funnel plot, indicating no evidence of publication bias.

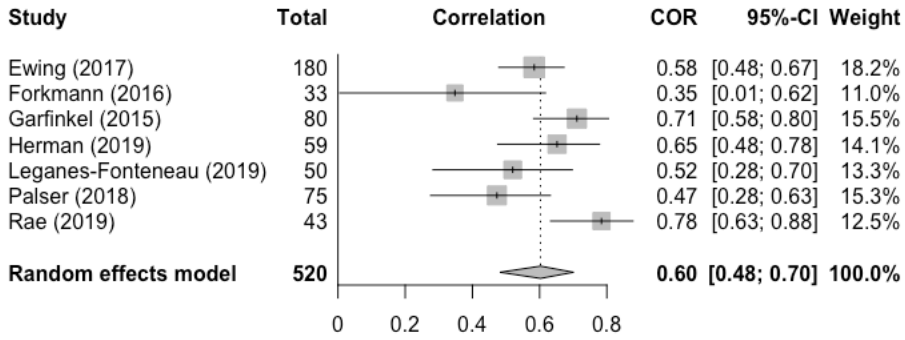

Figure 4. A forest plot displaying the pooled effect size of the confidence meta-analysis (dashed line), in addition to the individual effect sizes of each study. As can be seen, the random-effects model produced a pooled effect size of 0.60. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval.

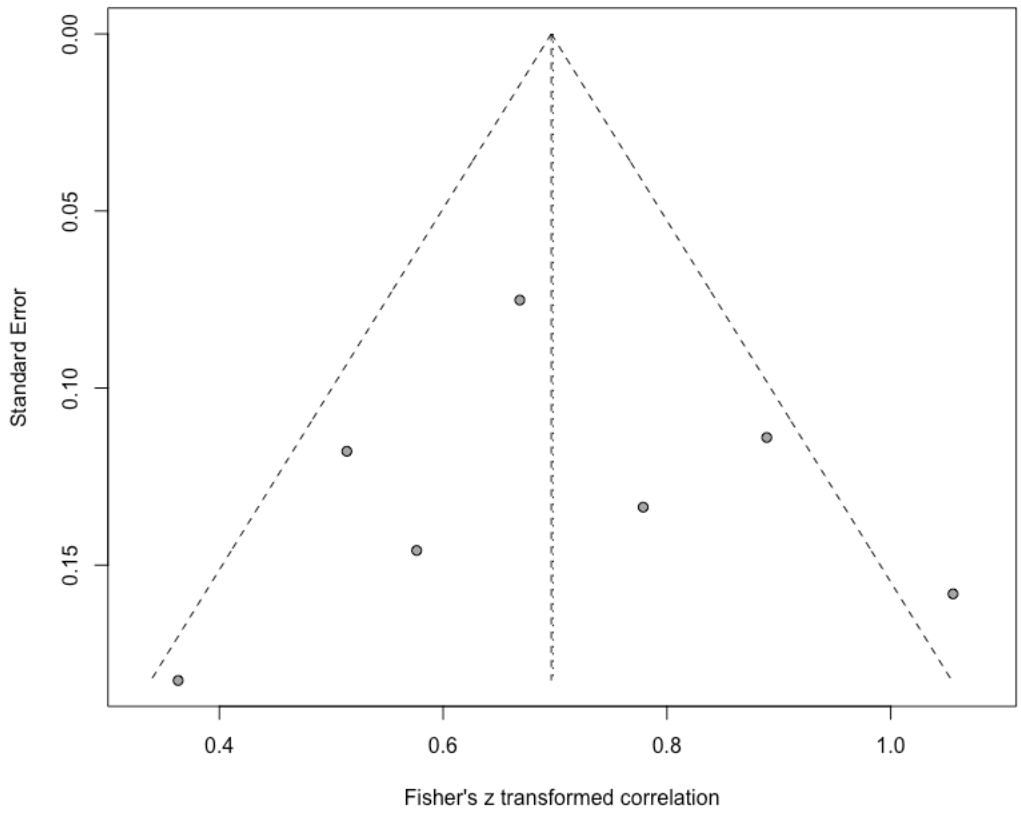

Figure 5. A funnel plot for the confidence meta-analysis. No evidence of publication bias was observed.

Awareness meta-analysis

A total of 6 studies reported the association between HCT and HDT interoceptive awareness scores. As such, a meta-analysis was performed on the data to obtain the pooled effect size of this relationship. Tests of heterogeneity revealed a non-significant Q statistic (Q = 2.04, p = .844) and an I^2 value of 0.0%. The meta-analysis identified a pooled effect size of 0.09 (p = .112), which is displayed in Figure 6. No outliers were identified as all of the individual effect sizes were seen to lie within the 95% confidence intervals of the pooled effect size. No publication bias was identified; Egger's value was non-significant (p = .563) consistent with the asymmetry of the funnel plot for this analysis (Figure 7).

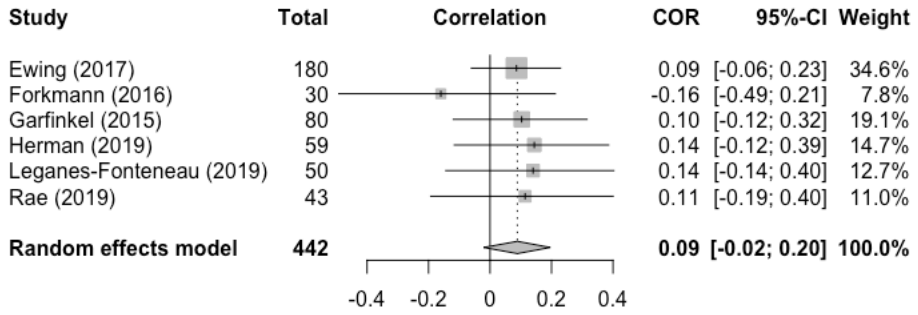

Figure 6. A forest plot displaying the pooled effect size (dashed line) identified by the awareness meta-analysis, in addition to the individual effect sizes from each study. As can be seen, the random-effects model produced a pooled effect size of 0.09. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval.

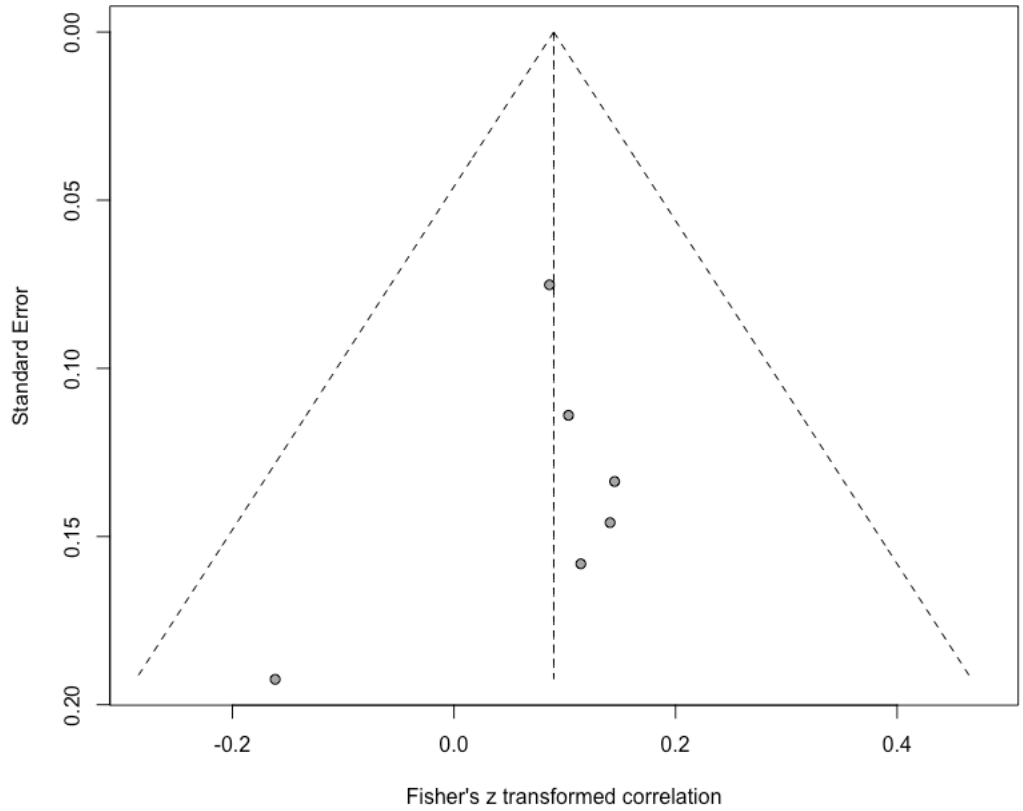

Figure 7. A funnel plot for the awareness meta-analysis. No evidence of publication bias was observed.

Discussion

This study aimed to investigate the relationship between the HCT and HDT with respect to interoceptive accuracy, confidence and awareness. Meta-analyses conducted for each of these dimensions of interoception revealed a small but significant correlation between HCT accuracy and HDT accuracy (4.4% variance shared), a moderate significant correlation between HCT confidence and HDT confidence (36.0% variance shared), and no significant correlation between HCT awareness and HDT awareness (0.8% variance shared). Comprehensive follow-up analyses indicated that this pattern of results held when accuracy analyses were restricted to correlation coefficients from typical participants, the 2AFC HDT, the auditory HDT, and the HDT administered with at least 40 trials. Consistent results for the accuracy meta-analysis were also observed following the separation of studies based on the device used to record heartbeats and the order in which the tasks were completed. Below, the results of these meta-analyses are discussed in turn.

Although the results of the accuracy meta-analysis are consistent with the proposal that at least some portion of the variance in performance on the HCT and HDT is shared (Garfinkel et al., 2015), the significant relationship observed was extremely small. Indeed, only 4.4% of the variance in accuracy on one measure was explained by accuracy on the other, offering little support for the idea that the measures are interchangeable. These data suggest that the influence of task differences on cardiac interoceptive accuracy scores is substantial, and is consistent with reported discrepancies between performance on the two tasks in previous studies, the patterning of performance on the HCT and HDT across atypical groups (e.g., Garfinkel et al., 2016b; Hina & Aspell, 2019; Mallorquí-Bagué et al., 2014), and the factors affecting performance (Phillips et al., 2003). Such discrepancies between performance on the HCT and HDT are perhaps unsurprising when considering the differing demands of these measures; in the HCT participants are required to keep track of the number

of counted heartbeats over long durations (sometimes as much as 103 seconds) which likely places demands on working memory and sustained attention, whereas in the HDT emphasis is put on the multisensory integration of exteroceptive and interoceptive stimuli. In addition, it should be noted that non-interoceptive strategies such as beliefs regarding resting heart rate (Brener & Ring, 2016; Ring & Brener, 1996; Ring, Brener, Knapp, & Mailloux, 2015; Windmann, Schonecke, Fröhlig, & Maldener, 1999) and time estimation abilities (Murphy et al., 2018) have been associated with good performance on the HCT but not the HDT (Knoll & Hodapp, 1992; Phillips, Jones, Rieger, & Snell, 2003). It is therefore understandable that performance on the two tasks diverges somewhat, not due to the underlying interoceptive ability required but rather the associated non-interoceptive task demands. Whilst it is possible that controlling for these different task demands may improve the relationship between HCT and HDT accuracy, in the absence of such control measures it appears that task differences override commonalities and that these measures cannot be treated as interchangeable tests of cardiac interoceptive accuracy. This may be particularly problematic considering that 82% of the articles that underwent full text screening for this meta-analysis were excluded because they employed only one of the two tasks. Indeed, given that our search terms were designed to identify papers that used both tasks, it is likely that a far greater percentage of studies employ only one task of cardiac interoceptive accuracy. The present results suggest that the results from studies that have used only one task may not generalise to the other.

The finding that HCT and HDT accuracy are only weakly correlated has important implications for the HCT specifically. Indeed, in recent years the HCT has been heavily criticised on the basis that non-interoceptive factors may influence performance (e.g., beliefs, time estimation; Brener & Ring, 2016; Murphy et al., 2018; Ring & Brener, 1996; Ring et al., 2015; Windmann et al., 1999), with concerns also raised regarding the psychometric properties of the task (Zamariola et al., 2018; but see Ainley et al. , 2020). Despite these

criticisms the HCT remains widely employed, with its use often justified by claiming a moderate correlation with the HDT (e.g., Borhani et al., 2017; Herbert et al., 2013; Pollatos et al., 2007; Scarpazza et al., 2017; Werner et al., 2009), a task that does not suffer from the aforementioned HCT-specific limitations (e.g., Knoll & Hodapp, 1992; Phillips et al., 2003). It is clear from the results of this meta-analysis that the relationship between HCT accuracy and HDT accuracy is far smaller than the moderate correlation reported by Knoll and Hodapp (1992), a study often used to justify the use of one task. Indeed, the correlation coefficient reported by Knoll and Hodapp (1992) was deemed to be an outlier in the accuracy meta-analysis along with one other study. This is particularly problematic for studies citing this paper to justify a relationship between HCT accuracy and HDT accuracy as it appears that the reported effect size is not representative of the overall findings in the field.

Whilst evidence of only a small association between HCT and HDT accuracy suggests that the inherent task differences strongly influence cardiac interoceptive accuracy scores, there are alternative explanations worth considering. One possibility is that the small association is driven by instability of interoceptive accuracy within individuals. Given evidence of state effects on HCT and HDT accuracy (Wittkamp, Bertsch, Vogele, & Schulz, 2018), it is possible that variations in the participants' state across the separate test phases, rather than variation that can be attributed to task effects, may account for the small association observed here. Indeed, as state variations in true interoceptive accuracy could result in a lack of consistency across testing phases (and in turn small associations across tasks), it remains a possibility that the association between the two tasks may in fact be stronger than the current meta-analysis suggests. However, as reasonable test-retest[3] reliability has been established for both tasks (e.g., HCT: $r = 0.41$ to $r = 0.60$; HDT: $r = 0.46$;

[3] Notably, few studies have examined the test-retest reliability of these tasks and, given differences in task administration, it remains a possibility that the test-retest reliability detailed here is not an accurate reflection of all variants of the HCT and HDT.

Ferentzi, Drew, Tihanyi, & Koteles, 2018; Wittkamp et al., 2018) across fairly long time periods (e.g., => 1 week), it is unlikely that the limited size of the association between accuracy scores on the two tasks is due to inconsistency in scores within the same participant across a brief testing session.

A second possibility is that differences in the administration of each task across studies contributes towards discrepancies in the relationships reported. As is evident from Tables 1-4, there are widespread differences in the exact procedures used for the HCT and HDT including the HCT time intervals used, the HDT delays used, the number of stimuli presented in HDT trials, the number of trials employed for each task, and the exact scoring methods used. The extent to which these specific administrative differences may impact between-task relationships may differ; for example, the two main HCT scoring methods (Hart et al., 2013; Schandry, 1981) are often highly correlated ($r = 0.987$, $p < .001$, Forkmann et al., 2016), meaning that it is unlikely that this difference will have influenced the observed association. Conversely, it is possible that other administrative differences (e.g., task formats or participant group) may contribute towards differences in the observed effect size of the relationship between HCT and HDT accuracy across studies. Whilst intuitive, the results of the present study are not entirely consistent with this proposal; indeed, more stringent exploratory analyses which included the removal of studies using 1) clinical and/or atypical populations, 2) less-commonly used versions of the HDT (MCS; 2AFC-visual), 3) fewer than 40 trials for the HDT, 4) different heartrate monitors (ECG vs pulse oximeter), and 5) different orders of task administration (HCT first vs counterbalanced) had little influence on the observed effect size of the relationship between HCT and HDT accuracy.

Although the results of the meta-analysis were consistent across a number of exploratory analyses, it should be acknowledged that it was not possible to account for all differences across studies. As noted, studies varied substantially in terms of the time

intervals/delays used for the HCT and HDT, the number of trials completed, as well as the instructions given to participants and scoring methods used, which are all factors that have been, or may be, associated with variability (and potentially unreliability) in scores (Brener & Ring, 2016; Desmedt et al., 2018; Kleckner et al., 2015). Unfortunately, given such heterogeneity and differences in the level of detail provided by the authors regarding their procedure, it was not feasible to run additional analyses to test all of these potential factors. As such, it is not possible to determine whether the effect size of the relationships reported here vary as a function of these factors. Nevertheless, these data serve to highlight the considerable variability in the application of both the HCT and HDT and echo recent calls to standardise the administration of these tasks within the field (Desmedt et al., 2020; Murphy et al., 2018).

As well as accuracy, this study sought to examine the relationship between the HCT and HDT on two further interoceptive dimensions: confidence and awareness. In terms of confidence ratings, a moderate correlation was observed between scores obtained by the HCT and HDT, consistent with the significant correlation often reported in the literature (Forkmann et al., 2016; Garfinkel et al., 2015). This seems intuitive as the two tasks index confidence in the same way; average confidence ratings across all trials, though far fewer trials are utilised for the HCT. Thus, despite differences in the number of ratings obtained, the demands on the participants are the same in both tasks. It is therefore fair to assume that confidence can be somewhat generalised between the HCT and HDT. What remains unclear (due to the tendency to not include control tasks) is whether confidence in the HCT and HDT is a specific interoceptive proclivity or whether it generalises to tasks in both other exteroceptive and interoceptive domains. Two recent studies contribute to this discussion; Murphy et al. (2020) observed no significant relationship between confidence scores on the HCT and a time estimation control task. This suggests that the confidence ratings are task-

specific rather than a measure of general confidence, though it should be noted that a trend emerged when participants who felt zero heartbeats were removed ($r = .36$, $p = .076$). Similarly, Garfinkel et al. (2016a) observed that confidence on the HDT was significantly correlated with confidence on a respiratory interoception task, but not with confidence on a touch acuity control task. However, respiratory interoceptive confidence was significantly correlated with touch acuity confidence. As such, whilst the results of the present study indicate some generalisability of confidence ratings from the HCT and HDT, it appears that further work is required in order to fully understand whether these confidence ratings are specific to cardiac interoception, generalisable across interoceptive domains (e.g., cardiac and respiratory), and whether confidence in performance across all domains of interoception is dissociable from general confidence in task performance.

In contrast to accuracy scores and confidence ratings where significant relationships were observed (with 4.4% and 36.0% shared variance respectively), the interoceptive awareness meta-analysis failed to identify a significant relationship between the HCT and HDT, consistent with previous studies (Forkmann et al., 2016; Garfinkel et al., 2015). This indicates that interoceptive awareness cannot be generalised between tasks and suggests that greater nuance is required when interpreting the results of studies assessing interoceptive awareness using one task. Such discrepancies may be driven by the different approaches for quantifying awareness across these tasks which, unlike confidence ratings, are notably different (ROC curves verses confidence-accuracy correlations). A further consideration relates to the number of trials administered. It has been suggested that at least 100 trials are required for an accurate estimate of metacognition (Fleming, 2017). As the maximum number of trials employed in the studies included in the awareness meta-analysis was 6 for the HCT and 40 for the HDT, it could be argued that both measures do not include a sufficient number of trials to precisely measure interoceptive awareness. Therefore, whilst the

present results suggest dissociation of interoceptive awareness as assessed by the HCT and HDT, it is likely that greater consideration of the measurement of interoceptive awareness is required more broadly.

Despite the relevance of these findings for our understanding of the relationship between these two commonly used tasks, it is important to acknowledge certain limitations. First, an inherent limitation to the meta-analytic approach is that inclusion is limited to published articles, a strategy that runs a risk of publication bias. Although steps were taken to mitigate possible publication bias (e.g., authors were contacted for unreported data), and no evidence of publication bias was obtained from any of the meta-analyses, there are limitations in inferring publication bias from meta-analyses conducted with few studies; first, whilst authors were contacted for data we did not contact researchers for unpublished data. As such, the meta-analysis was limited to data from published studies employing both tasks. Second, for meta-analyses with fewer than 10 studies it is difficult to assess publication bias as there is no agreed upon method (Dalton, Bolen & Mascha, 2016; Higgins et al., 2019). As such, the findings of the analyses with fewer than 10 studies should be treated with some caution as it is possible that publication bias could be present. For the confidence and awareness meta-analyses specifically, results are limited by the fact that few studies measured these aspects of interoception across both tasks. However, as the total pooled sample was relatively large at 520 and 442 participants respectively, it is likely that these meta-analyses provide an acceptable estimation of the effect size of these relationships.

A further limitation relates to the inferences that can be made with respect to the cause of the observed associations. For example, it may be that the perception of cardiac signals underlies the small correlation between accuracy scores on the HCT and HDT, with variability driven by differing task demands. However, it is also possible that non-interoceptive factors (such as motivation, attention, or IQ), which determine variance on both

tasks, drive the observed correlation. Further work which employs matched control tasks will be useful for elucidating the factors underlying these relationships and for determining whether both tasks assess cardiac interoceptive ability.

The exploratory meta-analyses attempted to reduce the impact of the different implementation strategies between studies such as device used and type of HDT employed. However, given that few studies employed alternative strategies (e.g., the visual version of the 2AFC HDT or MCS version of the HDT), little can be inferred about the specific impact of these alternative methods. One key limitation is the lack of studies employing other versions of the HDT, for example the 6AFC HDT or MCS. As only one study in the meta-analysis used an alternative approach (utilising the MCS HDT), an exploratory analysis could only be conducted using the 2AFC data. Methodological differences between the 2AFC, 6AFC and MCS versions of the HDT have been reported to affect performance; the 6AFC task or MCS thought to be preferable as these variants account for individual differences in the delay at which individuals perceive the external stimulus to be synchronous with their heartbeat (Brener & Ring, 2016). Indeed, the 2AFC HDT has been criticised on the basis that it assumes that all individuals experience heartbeat sensations as synchronous and asynchronous at the same temporal locations relative to the R-wave (Brener & Ring, 2016). Given this limitation of the 2AFC HDT, it is notable that the only study employing the MCS task observed no association with the HCT (Ring & Brener, 2018). As such, it is possible that the small correlation between HCT and HDT accuracy reported here may not generalise to all forms of the HDT.

In summary, this paper assimilated findings from 22 studies to reveal a small but significant relationship between accuracy scores on the HCT and HDT. The relationship observed was substantially smaller than studies often cited in the literature, thus highlighting that a degree of caution should be taken when generalising the results of studies that have

used only one task. Whilst it is unclear what underlies this small association, it is possible that the differing task demands of the measures contribute towards within-subject variability in cardiac interoceptive accuracy, with discrepancies in the effect sizes reported across studies potentially due to differences in the experimental protocols followed. Further research is required to assess these possibilities. For confidence ratings, a moderate relationship was observed across tasks, though further work is needed to determine whether this reflects an interoception-specific or domain-general disposition. In contrast, no evidence of an association between HCT and HDT interoceptive awareness was observed. Overall, these data suggest that whilst confidence ratings are moderately related across tasks, the HDT and HCT are not comparable when indexing interoceptive awareness, and there is little evidence for task equivalence in the measurement of interoceptive accuracy.

Acknowledgements

LH was supported by a BBSRC PhD studentship provided by the BBSRC Midlands Integrative Biosciences Training Partnership [grant reference: BB/M01116X/1]. JC was supported by the European Union's Horizon 2020 Research and Innovation Programme under ERC-2017-STG Grant Agreement No 757583. GB was supported by the Baily Thomas Trust.

References

Ainley, V., Tsakiris, M., Pollatos, O., Schulz, A, & Herbert, B. M. (2020). Comment on "Zamariola et al. (2018), interoceptive accuracy scores are problematic: evidence from simple bivariate correlations"- the empirical data base, the conceptual reasoning and the analysis behind this statement are misconceived and do not support the authors' conclusions. *Biological Psychology*, *152*, 107870. doi: 10.1016/j.biopsycho.2020.107870

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*(7), 419-429. doi:10.1038/nrn3950

Betka, S., Gould Van Praag, C., Paloyelis, Y., Bond, R., Pfeifer, G., Sequeira, H., . . . Critchley, H. (2018). Impact of intranasal oxytocin on interoceptive accuracy in alcohol users: an attentional mechanism? *Social Cognitive and Affective Neuroscience, 13*(4), 440-448. doi:10.1093/scan/nsy027

Borhani, K., Ladavas, E., Fotopoulou, A., & Haggard, P. (2017). "Lacking warmth": Alexithymia trait is related to warm-specific thermal somatosensory processing. *Biological Psychology, 128*, 132-140. doi:10.1016/j.biopsycho.2017.07.012

Brener, J., Liu, X., & Ring, C. (1993). A method of constant stimuli for examining heartbeat detection: Comparison with the Brener-Kluvitse and Whitehead methods. *Psychophysiology, 30*(6), 657-665. doi:10.1111/j.1469-8986.1993.tb02091.x

Brener, J., & Ring, C. (2016). Towards a psychophysics of interoceptive processes: the measurement of heartbeat detection. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708). doi:10.1098/rstb.2016.0015

Brewer, R., Cook, R., & Bird, G. (2016). Alexithymia: a general deficit of interoception. *Royal Society Open Science, 3*(10), 150664. doi:10.1098/rsos.150664

Carroll, D., & Whellock, J. (1980). Heart rate perception and the voluntary control of heart rate. *Biological Psychology, 11*, 169-180. doi:10.1016/0301-0511(80)90053-8

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience, 3*(8), 655-666. doi:10.1038/nrn894

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology, 13*(4), 500-505. doi:10.1016/s0959-4388(03)00090-4

Craig, A. D. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*(1), 59-70. doi:10.1038/nrn2555

Dale, A., & Anderson, D. (1978). Information variables in voluntary control and classical conditioning of heart rate: Field dependence and heart rate perception. *Perceptual and Motor Skills, 47*, 79-85. doi:10.2466/pms.1978.47.1.79

Dalton, J. E., Bolen, S. D., & Mascha, E. J. (2016). Publication bias: the elephant in the review. *Anesthesia and Analgesia*, *123*(4), 812. doi: 10.1213/ane.0000000000001596

Desmedt, O., Corneille, O., Luminet, O., Murphy, J., Bird, G., & Maurage, P. (2020). Contribution of time estimation and knowledge to heartbeat counting task performance under original and adapted instructions. Biological Psychology, 154, 107904. doi: 10.1016/j.biopsycho.2020.107904

Desmedt, O., Luminet, O., & Corneille, O. (2018). The heartbeat counting task largely involves non-interoceptive processes: Evidence from both the original and an adapted counting task. *Biological Psychology, 138*, 185-188. doi:10.1016/j.biopsycho.2018.09.004

Domschke, K., Stevens, S., Pfleiderer, B., & Gerlach, A. L. (2010). Interoceptive sensitivity in anxiety and anxiety disorders: an overview and integration of neurobiological findings. *Clinical Psychology Review, 30*(1), 1-11. doi:10.1016/j.cpr.2009.08.008

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629-634. doi:10.1136/bmj.315.7109.629

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers 28*, 1-11.

Ewing, D. L., Manassei, M., Gould van Praag, C., Philippides, A. O., Critchley, H. D., & Garfinkel, S. N. (2017). Sleep and the heart: Interoceptive differences linked to poor experiential sleep quality in anxiety and depression. *Biological Psychology, 127*, 163-172. doi:10.1016/j.biopsycho.2017.05.011

Ferentzi, E., Drew, R., Tihanyi, B. T., & Koteles, F. (2018). Interoceptive accuracy and body awareness - Temporal and longitudinal associations in a non-clinical sample. *Physiology & Behavior, 184*, 100-107. doi:10.1016/j.physbeh.2017.11.015

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*(2), 161-180. doi:10.1037/1082-989x.6.2.161

Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness, 2017*(1), nix007. doi:10.1093/nc/nix007

Forkmann, T., Scherer, A., Meessen, J., Michal, M., Schachinger, H., Vogele, C., & Schulz, A. (2016). Making sense of what you sense: Disentangling interoceptive awareness, sensibility and accuracy. *International Journal of Psychophysiology, 109*, 71-80. doi:10.1016/j.ijpsycho.2016.09.019

Füstös, J., Gramann, K., Herbert, B. M., & Pollatos, O. (2013). On the embodiment of emotion regulation: interoceptive awareness facilitates reappraisal. *Social Cognitive and Affective Neuroscience, 8*(8), 911-917. doi:10.1093/scan/nss089

Gannon, L. R. (1980). Cardiac perception and the voluntary control of heart rate. *Physiological Psychology, 8*(4), 509-514. doi:10.3758/BF03326485

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., In den Bosch, Y., Critchley, H. D., & Engels, M. (2016a). Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society B, 371*(1708). doi:10.1098/rstb.2016.0014

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology, 104*, 65-74. doi:10.1016/j.biopsycho.2014.11.004

Garfinkel, S. N., Tiley, C., O'Keeffe, S., Harrison, N. A., Seth, A. K., & Critchley, H. D. (2016b). Discrepancies between dimensions of interoception in autism: Implications for emotion and anxiety. *Biological Psychology, 114*, 117-126. doi:10.1016/j.biopsycho.2015.12.003

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D.D. (2019). dmetar: Companion R Package For The Guide "Doing Meta-Analysis in R". R package version 0.0.9000. Available from http://dmetar.protectlab.org.

Harshaw, C. (2015). Interoceptive dysfunction: toward an integrated framework for understanding somatic and affective disturbance in depression. *Psychological Bulletin, 141*(2), 311-363. doi:10.1037/a0038101

Hart, N., McGowan, J., Minati, L., & Critchley, H. D. (2013). Emotional regulation and bodily sensation: interoceptive awareness is intact in borderline personality disorder. *Journal of Personality Disorders,, 27*(4), 506-518. doi:10.1521/pedi_2012_26_049

Herbert, B. M., Blechert, J., Hautzinger, M., Matthias, E., & Herbert, C. (2013). Intuitive eating is associated with interoceptive sensitivity. Effects on body mass index. *Appetite, 70*, 22-30. doi:10.1016/j.appet.2013.06.082

Herbert, B. M., & Pollatos, O. (2014). Attenuated interoceptive sensitivity in overweight and obese individuals. *Eating Behaviors, 15*(3), 445-448. doi:10.1016/j.eatbeh.2014.06.002

Herbert, B. M., Pollatos, O., Flor, H., Enck, P., & Schandry, R. (2010). Cardiac awareness and autonomic cardiac reactivity during emotional picture viewing and mental stress. *Psychophysiology, 47*(2), 342-354. doi:10.1111/j.1469-8986.2009.00931.x

Herman, A. M., Rae, C. L., Critchley, H. D., & Duka, T. (2019). Interoceptive accuracy predicts nonplanning trait impulsivity. *Psychophysiology, 56*(6), e13339. doi:10.1111/psyp.13339

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327*(7414), 557-561. doi:10.1136/bmj.327.7414.557

Hina, F., & Aspell, J. E. (2019). Altered interoceptive processing in smokers: Evidence from the heartbeat tracking task. *International Journal of Psychophysiology, 142*, 10-16. doi:10.1016/j.ijpsycho.2019.05.012

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological methods*, *11*(2), 193. doi: 10.1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*(4), 275-292. doi:10.1111/1468-2389.00156

Kandasamy, N., Garfinkel, S. N., Page, L., Hardy, B., Critchley, H. D., Gurnell, M., & Coates, J. M. (2016). Interoceptive Ability Predicts Survival on a London Trading Floor. *Scientific Reports, 6*, 32986. doi:10.1038/srep32986

Katkin, S. D., Reed, C., & Deroo, A. (1983). A methodological analysis of 3 techniques for the assessment of individual-differences in heartbeat detection. *Psychophysiology, 20*(4), 452.

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., . . . Interoception Summit, p. (2018). Interoception and Mental Health: A Roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(6), 501-513. doi:10.1016/j.bpsc.2017.12.004

Khalsa, S. S., & Lapidus, R. C. (2016). Can interoception improve the pragmatic search for biomarkers in psychiatry? *Frontiers Psychiatry, 7*, 121. doi:10.3389/fpsyt.2016.00121

Khalsa, S. S., Rudrauf, D., Sandesara, C., Olshansky, B., & Tranel, D. (2009). Bolus isoproterenol infusions provide a reliable method for assessing interoceptive awareness. *International Journal of Psychophysiology, 72*(1), 34-45. doi:10.1016/j.ijpsycho.2008.08.010

Klabunde, M., Acheson, D. T., Boutelle, K. N., Matthews, S. C., & Kaye, W. H. (2013). Interoceptive sensitivity deficits in women recovered from bulimia nervosa. *Eating Behaviors, 14*(4), 488-492. doi:10.1016/j.eatbeh.2013.08.002

Kleckner, I. R., Wormwood, J. B., Simmons, W. K., Barrett, L. F., & Quigley, K. S. (2015). Methodological recommendations for a heartbeat detection-based measure of interoceptive sensitivity. *Psychophysiology, 52*(11), 1432-1440. doi:10.1111/psyp.12503

Knoll, J. F., & Hodapp, V. (1992). A comparison between two methods for assessing heartbeat perception. . *Psychophysiology, 29*(2), 218-222. doi:10.1111/j.1469-8986.1992.tb01689.x

Leganes-Fonteneau, M., Cheang, Y., Lam, Y., Garfinkel, S., & Duka, T. (2019). Interoceptive awareness is associated with acute alcohol-induced changes in subjective effects. *Pharmacology Biochemistry and Behavior*, *181*, 69-76. doi: 10.1016/j.pbb.2019.03.007

Mallorquí-Bagué, N., Garfinkel, S. N., Engels, M., Eccles, J. A., Pailhez, G., Bulbena, A., & Critchley, H. D. (2014). Neuroimaging and psychophysiological investigation of the link between anxiety, enhanced affective reactivity and interoception in people with joint hypermobility. *Frontiers in Psychology, 5*, 1162. doi:10.3389/fpsyg.2014.01162

McFarland, R. A. (1975). Heart rate perception and heart rate control. *Psychophysiology, 12*(4), 402-405. doi:10.1111/j.1469-8986.1975.tb00011.x

Michal, M., Reuchlein, B., Adler, J., Reiner, I., Beutel, M. E., Vogele, C., . . . Schulz, A. (2014). Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS One, 9*(2), e89823. doi:10.1371/journal.pone.0089823

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine*, *6*(7), e1000097. doi: 10.1371/journal.pmed.1000097

Mul, C. L., Stagg, S. D., Herbelin, B., & Aspell, J. E. (2018). The feeling of me feeling for you: interoception, alexithymia and empathy in autism. *Journal of Autism and Developmental Disorders, 48*(9), 2953-2967. doi:10.1007/s10803-018-3564-3

Murphy, J., Brewer, R., Catmur, C., & Bird, G. (2017). Interoception and psychopathology: A developmental neuroscience perspective. *Developmental Cognitive Neuroscience, 23*, 45-56. doi:10.1016/j.dcn.2016.12.006

Murphy, J., Brewer, R., Coll, M. P., Plans, D., Hall, M., Shiu, S. S., . . . Bird, G. (2019a). I feel it in my finger: Measurement device affects cardiac interoceptive accuracy. *Biological Psychology, 148*, 107765. doi:10.1016/j.biopsycho.2019.107765

Murphy, J., Brewer, R., Hobson, H., Catmur, C., & Bird, G. (2018). Is alexithymia characterised by impaired interoception? Further evidence, the importance of control variables, and the problems with the Heartbeat Counting Task. *Biological Psychology, 136*, 189-197. doi:10.1016/j.biopsycho.2018.05.010

Murphy, J., Brewer, R., Plans, D., Khalsa, S. S., Catmur, C., & Bird, G. (2020). Testing the independence of self-reported interoceptive accuracy and attention. . *Quarterly Journal of Experimental Psychology, 73*(1), 115-133. doi:10.1177/1747021819879826

Murphy, J., Catmur, C., & Bird, G. (2019b). Classifying individual differences in interoception: Implications for the measurement of interoceptive awareness. *Psychonomic Bulletin & Review, 26*(5), 1467-1471. doi:10.3758/s13423-019-01632-7

Palser, E. R., Fotopoulou, A., Pellicano, E., & Kilner, J. M. (2018). The link between interoceptive processing and anxiety in children diagnosed with autism spectrum disorder: Extending adult findings into a developmental sample. *Biological Psychology, 136*, 13-21. doi:10.1016/j.biopsycho.2018.05.003

Pauli, P., Hartl, L., Marquardt, C., Stalmann, H., & Strian, F. (1991). Heartbeat and arrhythmia perception in diabetic autonomic neuropathy. *Psychological medicine, 21*(02), 413-421. doi:10.1017/s0033291700020523

Paulus, M. P., & Stein, M. B. (2006). An insular view of anxiety. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 60*(4), 383-387. doi:10.1016/j.biopsych.2006.03.042

Phillips, G. C., Jones, G. E., Rieger, E. J., & Snell, J. B. (2003). Effects of the presentation of false heart-rate feedback on the performance of two common heartbeat-detection tasks. *Psychophysiology, 36*(4), 504-510. doi:10.1111/psyp.12503

Pollatos, O., Kurz, A. L., Albrecht, J., Schreder, T., Kleemann, A. M., Schopf, V., . . . Schandry, R. (2008). Reduced perception of bodily signals in anorexia nervosa. *Eating Behaviors, 9*(4), 381-388. doi:10.1016/j.eatbeh.2008.02.001

Pollatos, O., Traut-Mattausch, E., & Schandry, R. (2009). Differential effects of anxiety and depression on interoceptive accuracy. *Depression and Anxiety, 26*(2), 167-173. doi:10.1002/da.20504

Pollatos, O., Traut-Mattausch, E., Schroeder, H., & Schandry, R. (2007). Interoceptive awareness mediates the relationship between anxiety and the intensity of unpleasant feelings. *Journal of Anxiety Disorders, 21*(7), 931-943. doi:10.1016/j.janxdis.2006.12.004

Quattrocki, E., & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews, 47*, 410-430. doi:10.1016/j.neubiorev.2014.09.012

Rae, C. L., Larsson, D. E. O., Garfinkel, S. N., & Critchley, H. D. (2019). Dimensions of interoception predict premonitory urges and tic severity in Tourette syndrome. *Psychiatry Research, 271*, 469-475. doi:10.1016/j.psychres.2018.12.036

Ring, C., & Brener, J. (1996). Influence of beliefs about heart rate and actual heart rate on heartbeat counting. *Psychophysiology, 33*(5), 541–546. doi:10.1111/j.1469-8986.1996.tb02430.x

Ring, C., & Brener, J. (2018). Heartbeat counting is unrelated to heartbeat detection: A comparison of methods to quantify interoception. *Psychophysiology, 55*(9), e13084. doi:10.1111/psyp.13084

Ring, C., Brener, J., Knapp, K., & Mailloux, J. (2015). Effects of heartbeat feedback on beliefs about heart rate and heartbeat counting: a cautionary tale about interoceptive awareness. *Biological Psychology, 104*, 193-198. doi:10.1016/j.biopsycho.2014.12.010

Scarpazza, C., Sellitto, M., & di Pellegrino, G. (2017). Now or not-now? The influence of alexithymia on intertemporal decision-making. *Brain and Cognition, 114*, 20-28. doi:10.1016/j.bandc.2017.03.001

Schaefer, M., Egloff, B., & Witthoft, M. (2012). Is interoceptive awareness really altered in somatoform disorders? Testing competing theories with two paradigms of heartbeat perception. *Journal of Abnormal Psychology, 121*(3), 719-724. doi:10.1037/a0028509

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology, 18*(4), 483-488. doi:10.1111/j.1469-8986.1981.tb02486.x

Schroeder, S., Gerlach, A. L., Achenbach, S., & Martin, A. (2015). The relevance of accuracy of heartbeat perception in noncardiac and cardiac chest pain. *International Journal of Behavioural Medicine, 22*(2), 258-267. doi:10.1007/s12529-014-9433-3

Schulz, A., Lass-Hennemann, J., Sutterlin, S., Schachinger, H., & Vogele, C. (2013). Cold pressor stress induces opposite effects on cardioceptive accuracy dependent on assessment paradigm. *Biological Psychology, 93*(1), 167-174. doi:10.1016/j.biopsycho.2013.01.007

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R news*, *7*(3), 40-45.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences,, 17*(11), 565-573. doi:10.1016/j.tics.2013.09.007

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine, 26*(9), 1964-1981. doi:10.1002/sim.2688

Villani, V., Tsakiris, M., & Azevedo, R. T. (2019). Transcutaneous vagus nerve stimulation improves interoceptive accuracy. *Neuropsychologia, 134*, 107201. doi:10.1016/j.neuropsychologia.2019.107201

Weitkunat, R. (1996). Cardioception and psychophysiological variables in panic patients and healthy controls. . *Cognitive and Behavioral Neurology, 9*(1), 8-15.

Werner, N. S., Jung, K., Duschek, S., & Schandry, R. (2009). Enhanced cardiac perception is associated with benefits in decision-making. *Psychophysiology, 46*(6), 1123-1129. doi:10.1111/j.1469-8986.2009.00855.x

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-regulation, 2*(4), 371–392. doi:10.1007/BF00998623

Windmann, S., Schonecke, O. W., Fröhlig, G., & Maldener, G. (1999). Dissociating beliefs about heart rates and actual heart rates in patients with cardiac pacemakers. *Psychophysiology, 36*(3), 339-342. doi:10.1017/s0048577299980381

Wittkamp, M. F., Bertsch, K., Vogele, C., & Schulz, A. (2018). A latent state-trait analysis of interoceptive accuracy. *Psychophysiology, 55*(6), e13055. doi:10.1111/psyp.13055

Yates, A. J., Jones, K. E., Marie, G. V., & Hogben, J. H. (1985). Detection of the heartbeat and events in the cardiac cycle. *Psychophysiology*, *22*(5), 561-567. doi: 10.1111/j.1469-8986.1985.tb01651.x

Zamariola, G., Maurage, P., Luminet, O., & Corneille, O. (2018). Interoceptive accuracy scores from the heartbeat counting task are problematic: Evidence from simple

bivariate correlations. *Biological Psychology, 137*, 12-17.

doi:10.1016/j.biopsycho.2018.06.006

Supplementary Materials

Meta-analyses

[S1] Exploratory Meta-Analyses

In order to assess the reliability of the effect sizes reported, additional exploratory meta-analyses were conducted using the accuracy data. For the first exploratory meta-analysis, data from clinical and/or atypical populations were removed from the meta-analysis. Where statistics were reported separately for both clinical and control groups, this was achieved by removing the statistics relating to the clinical group. Where correlation coefficients were reported for the whole sample only, these studies were removed. This resulted in the removal of 11 correlation coefficients. For the second and third meta-analyses, data were separated on the basis of the device used to record heartbeats; 13 studies used an ECG and 10 studies used a pulse oximeter. In the fourth and fifth meta-analyses, studies using alternative versions of the HDT were removed; in the fourth, one study using the MCS version of the HDT was removed, and in the fifth one study that employed the visual version of the HDT was removed. The sixth meta-analysis analysed the 9 studies using at least 40 trials for the HDT (as recommended by Kleckner et al., 2015). For the seventh and eighth meta-analyses, data were separated based on the order in which the tasks were completed; 13 studies administered the HCT first and 8 studies administered the tasks in a counterbalanced order.

Results

[S2] Excluding clinical and/or atypical populations

The accuracy analyses were re-run following the exclusion of data from clinical and/or atypical populations. Tests of heterogeneity revealed a significant Q statistic (Q = 26.94, p = .004) and an I^2 value of 59.2%. A pooled effect size of 0.21 (p = .002) was obtained from the meta-analysis. This value and the individual effect sizes of each study are

displayed in Supplementary Figure 1 [S3]. To infer the presence of publication bias in this sample, a funnel plot was produced (Supplementary Figure 2 [S4]). The Egger's test was non-significant (p = .638), indicating no evidence of publication bias.

[S3] Supplementary Figure 1. Excluding clinical and/or atypical populations forest plot

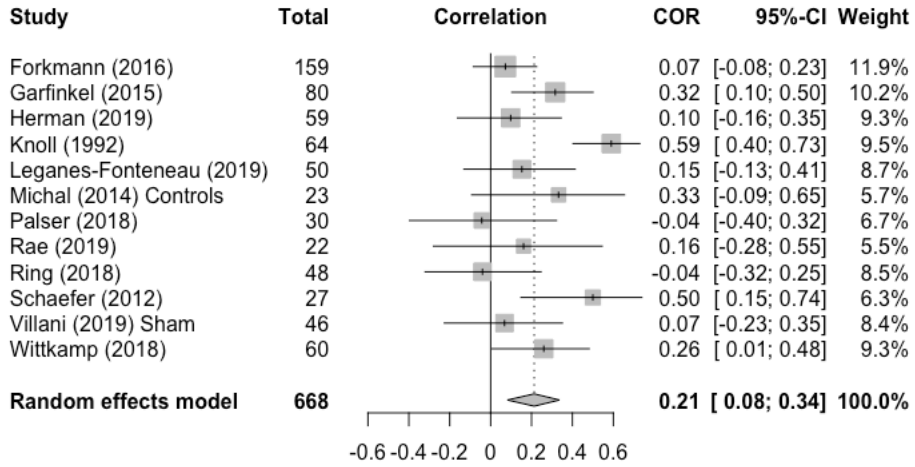

Figure 1. A forest plot displaying the pooled effect size (dashed line) and individual study effect sizes of the accuracy meta-analysis excluding clinical and/or atypical populations such as clinical populations. As can be seen, the random-effects model produced a pooled effect size of 0.21. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S4] Supplementary Figure 2. Excluding clinical and/or atypical populations funnel plot

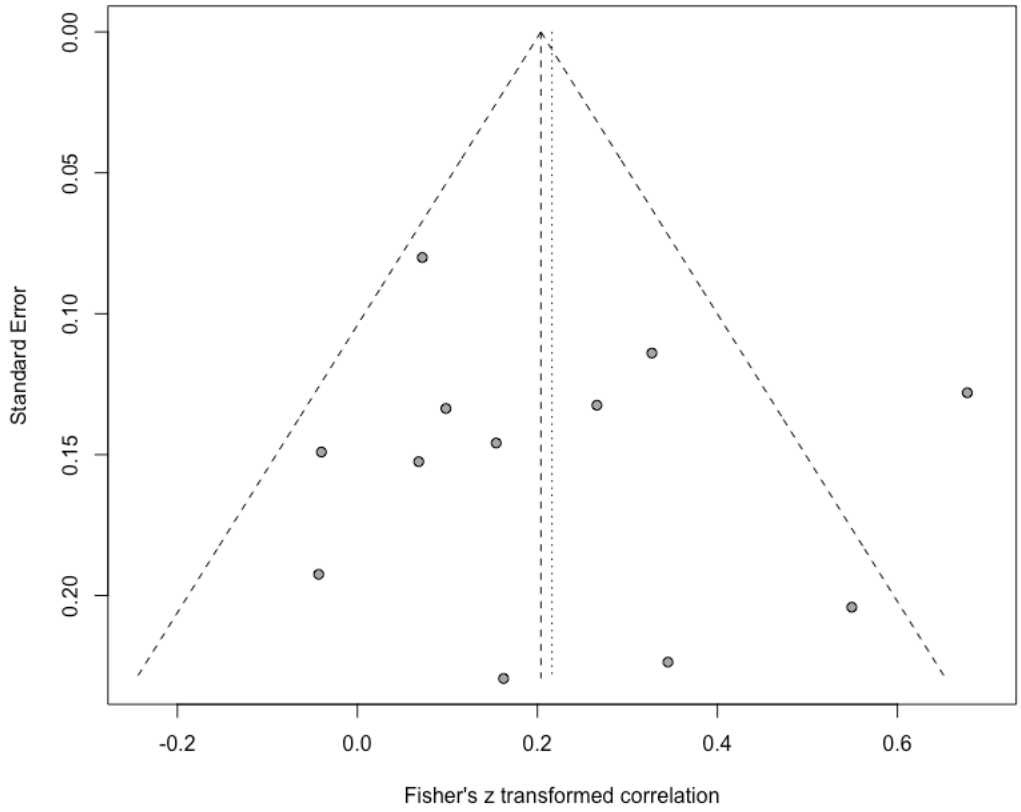

Figure 2. A funnel plot for the accuracy meta-analysis excluding clinical populations. No evidence of publication bias was observed.

[S5] Separation based on device used

Two separate meta-analyses were conducted for studies using a pulse oximeter or an ECG to record heartbeats. Only the ECG analysis produced a significant Q statistic (ECG: $Q = 27.38$, $p = .007$; pulse oximeter: $Q = 13.55$, $p = .139$), but moderate heterogeneity was indicated through I^2 values of 56.2% and 33.6% respectively. The ECG meta-analysis produced a pooled effect size of 0.22 ($p < .001$) and the pulse oximeter meta-analysis produced a pooled effect size of 0.19 ($p < .001$). Due to overlapping confidence intervals (ECG- [0.10, 0.33], PO- [0.08, 0.30]), it appears that the device used has no significant effect on the results.

Supplementary Figure 3 [S6] displays the forest plots for the two meta-analyses and the funnel plots can be seen in Supplementary Figure 4 [S7]. Both meta-analyses produced a non-significant Egger's test (ECG: $p = .947$; pulse oximeter: $p = .453$) indicating no evidence of publication bias.

[S6] Supplementary Figure 3. ECG and pulse oximeter forest plots

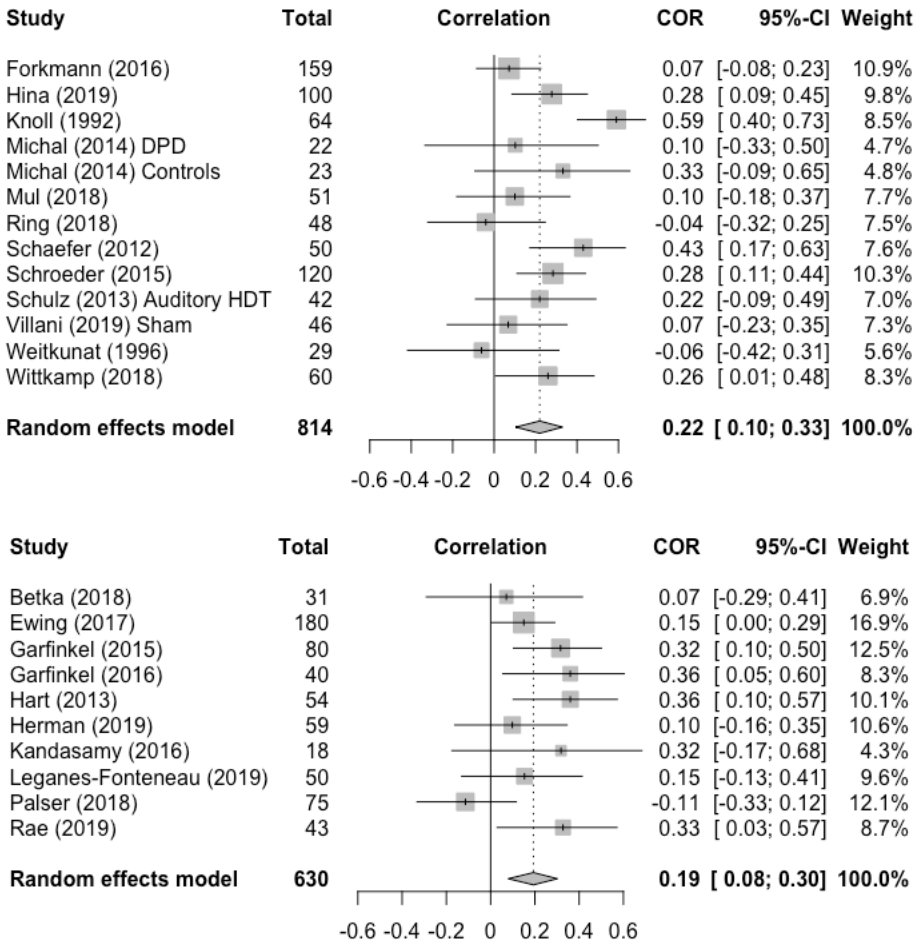

Figure 3. Forest plots for the ECG accuracy meta-analysis (top) and the pulse oximeter accuracy meta-analysis (bottom), identifying a pooled effect size of 0.22 and 0.19 respectively (dashed lines) when referring to the random-effects models. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, DPD = Depersonalization Disorder, HDT = heartbeat discrimination task, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S7] Supplementary Figure 4. ECG and pulse oximeter funnel plots

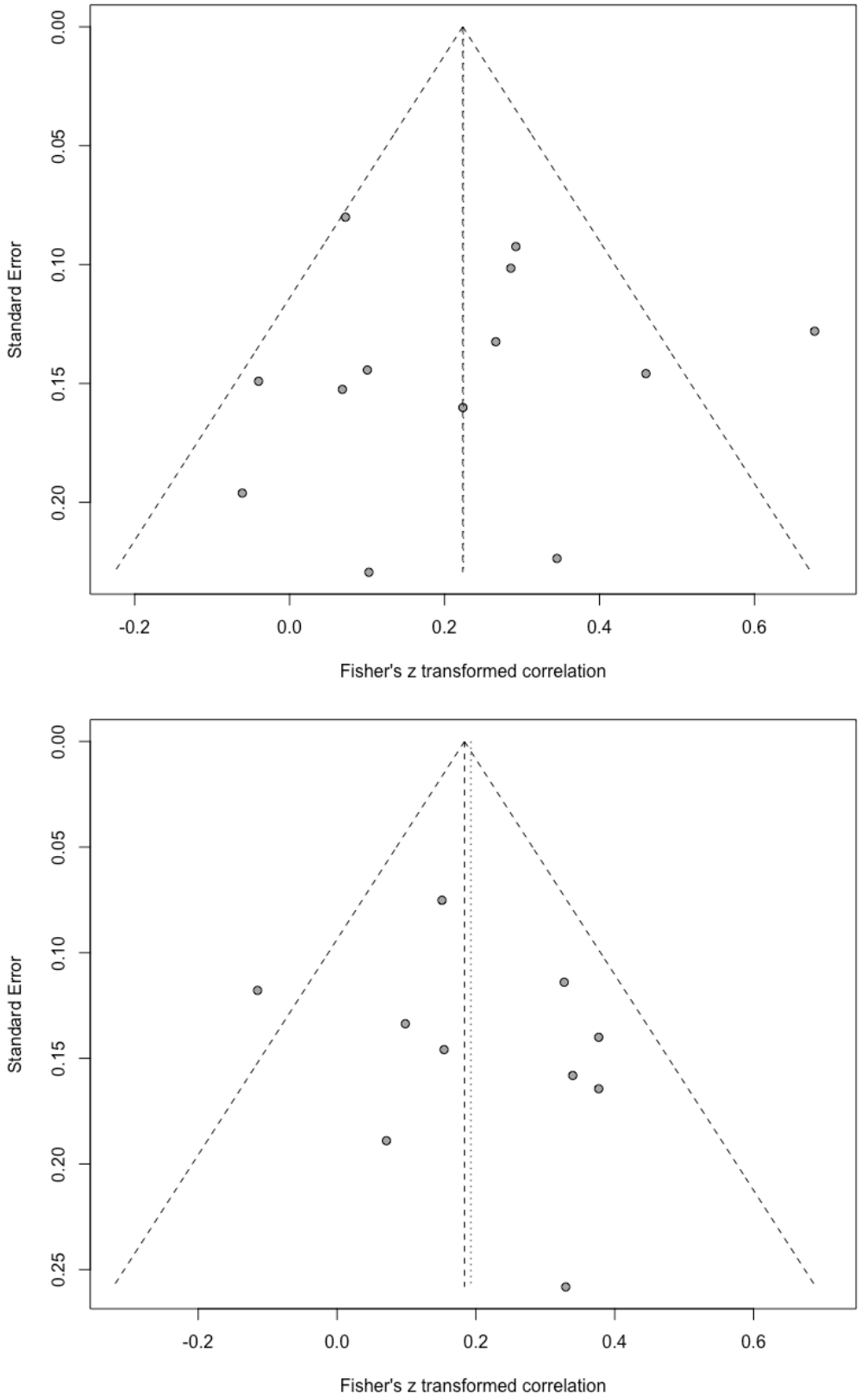

Figure 4. Funnel plots for the ECG accuracy meta-analysis (top) and the pulse oximeter accuracy meta-analysis (bottom). No evidence of publication bias was observed.

[S8] Excluding MCS HDT studies

One study reporting the MCS HDT was removed from analyses and heterogeneity was reassessed. This produced a significant Q statistic ($Q = 38.65$, $p = .011$) and an I^2 value of 45.7%, suggesting moderate heterogeneity. A pooled effect size of 0.22 ($p < .001$) was recorded, which is displayed in Supplementary Figure 5 [S9] alongside the individual effect sizes. Supplementary Figure 6 [S10] displays the funnel plot for the data which suggests no evidence of publication bias; a non-significant Egger's test was produced by the data ($p = .567$).

[S9] Supplementary Figure 5. Excluding MCS HDT studies forest plot

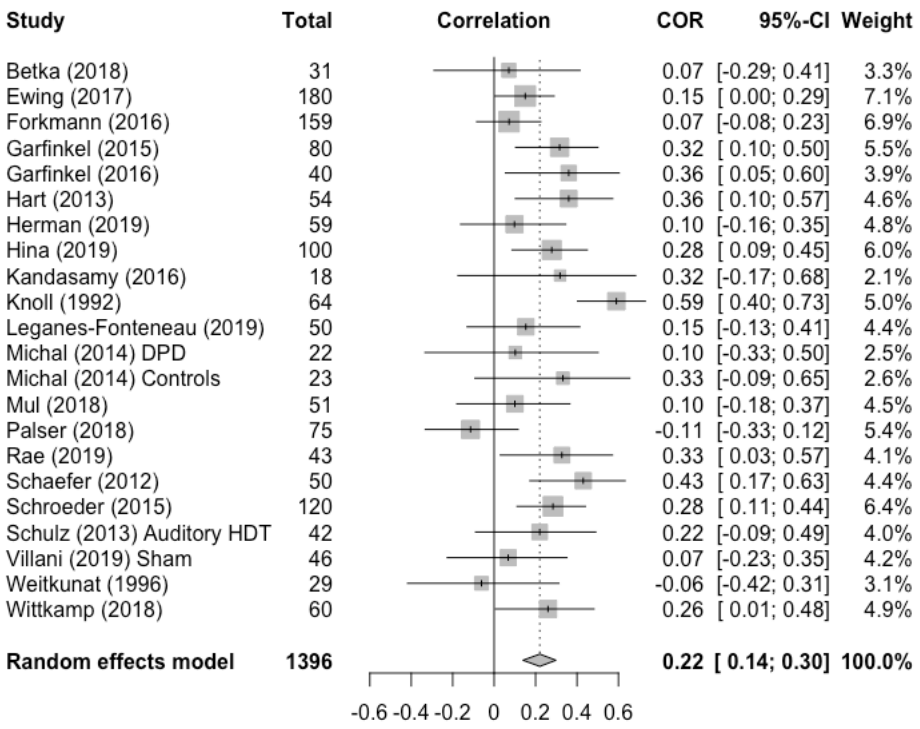

Figure 5. A forest plot for the accuracy meta-analysis excluding a study reporting the MCS HDT. As can be seen, the random-effects model identified a pooled effect size of 0.22 (dashed line). Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, DPD = Depersonalization Disorder, HDT = heartbeat discrimination

task, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S10] Supplementary Figure 6. Excluding MCS HDT studies funnel plot

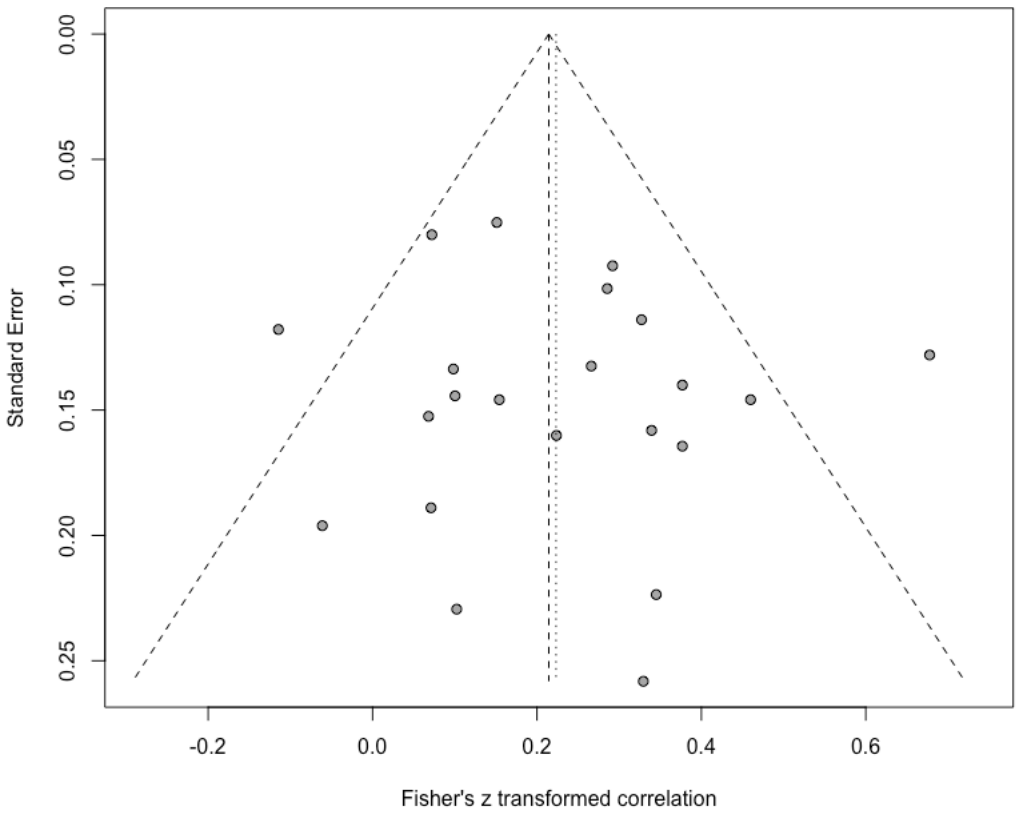

Figure 6. A funnel plot for the accuracy meta-analysis excluding a study reporting the MCS HDT. No evidence of publication bias was observed.

[S11] Excluding visual HDT studies

Following the exclusion of one correlation coefficient relating to the relationship between performance on the HCT and the visual HDT, the heterogeneity analyses returned a significant Q statistic ($Q = 41.26$, $p = .005$) and an I^2 value of 49.1%. For this sample, the meta-analysis identified a pooled effect size of 0.21 ($p < .001$). Supplementary Figure 7 [S12] displays the individual effect sizes and pooled effect size for this sample. The funnel plot for

this sample is displayed in Supplementary Figure 8 [S13]. Again, the Egger's test was non-significant (p = .694).

[S12] Supplementary Figure 7. Excluding visual HDT studies forest plot

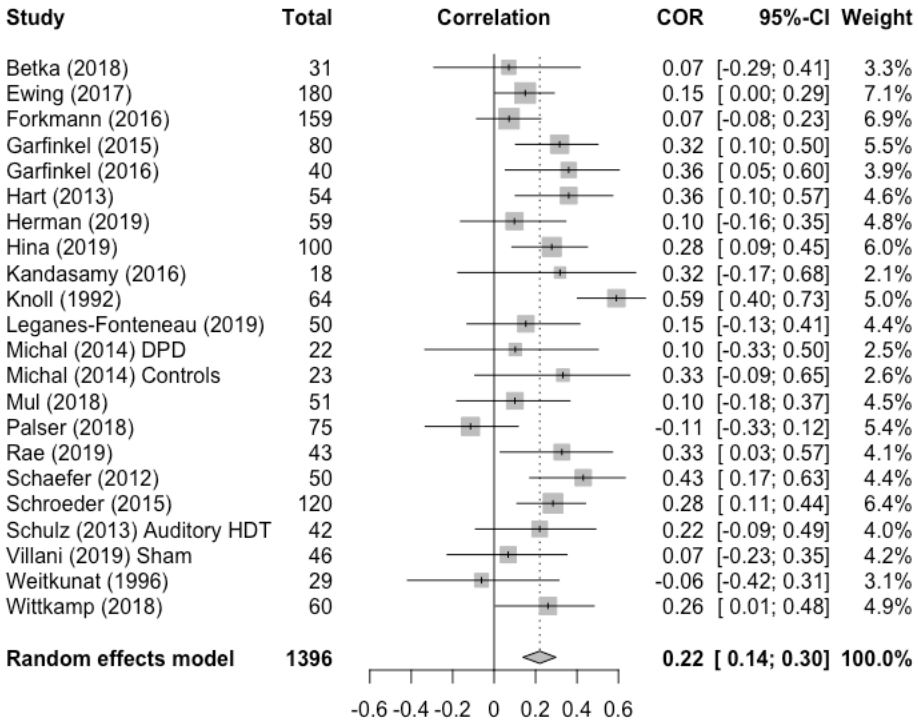

Figure 7. A forest plot for the accuracy meta-analysis following the exclusion of one correlation coefficient relating to a visual version of the HDT. As can be seen, the random-effects model identified a pooled effect size of 0.21 (dashed line). Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, DPD = Depersonalization Disorder, HDT = heartbeat discrimination task, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S13] Supplementary Figure 8. Excluding visual HDT studies funnel plot

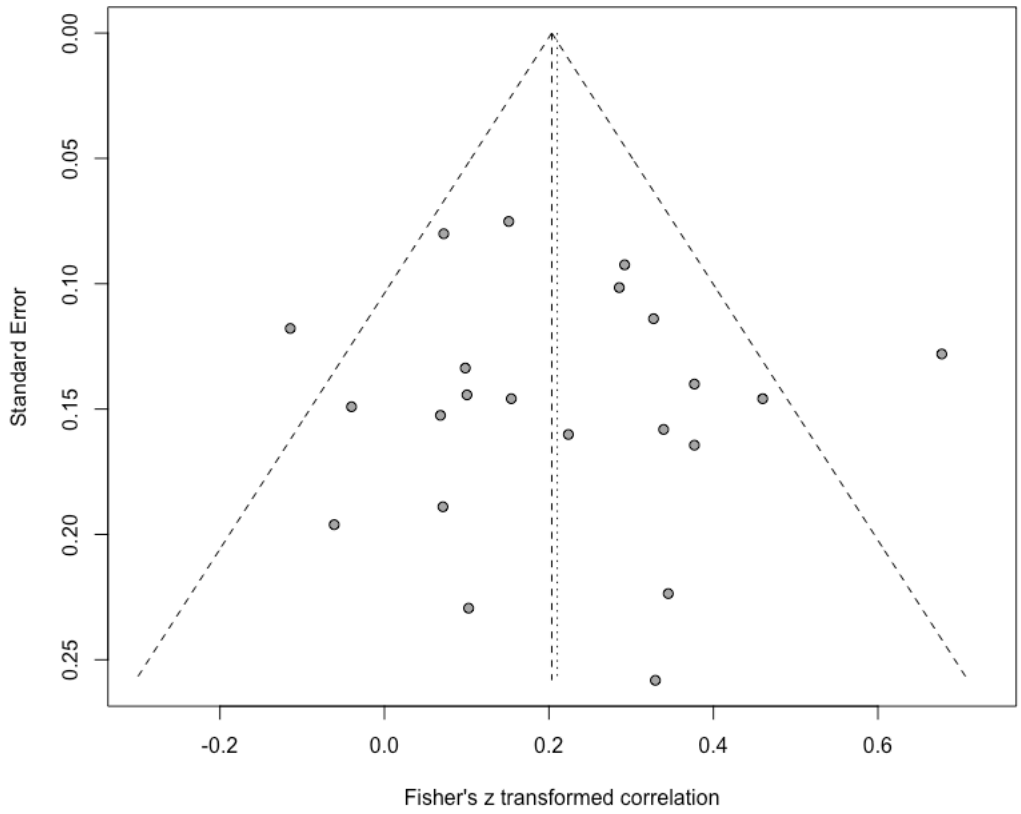

Figure 8. A funnel plot for the accuracy meta-analysis following the exclusion of one correlation coefficient relating to a visual version of the HDT. No evidence of publication bias was observed.

[S14] Excluding studies with fewer than 40 HDT trials

Studies with fewer than 40 HDT trials were removed from analyses and heterogeneity was reassessed. This produced a significant Q statistic ($Q = 26.81$, $p < .001$) and an I^2 value of 70.2%, suggesting moderate heterogeneity. A pooled effect size of 0.24 ($p = .004$) was recorded, which is displayed in Supplementary Figure 9 [S15] alongside the individual effect sizes. Supplementary Figure 10 [S16] displays the funnel plot for the data which suggests no

evidence of publication bias; a non-significant Egger's test was produced by the data (p = .847).

[S15] Supplementary Figure 5. Excluding studies with fewer than 40 HDT trials forest plot

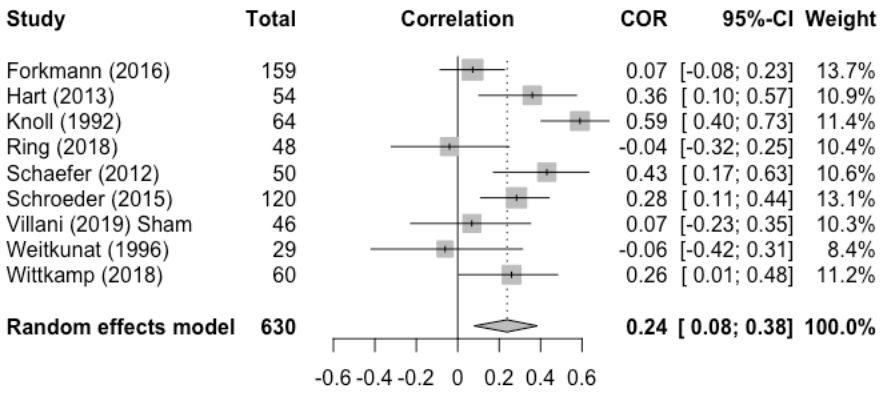

Figure 9. A forest plot for the accuracy meta-analysis excluding studies with fewer than 40 HDT trials. As can be seen, the random-effects model identified a pooled effect size of 0.24 (dashed line). Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S16] Supplementary Figure 6. Excluding studies with fewer than 40 HDT trials funnel plot

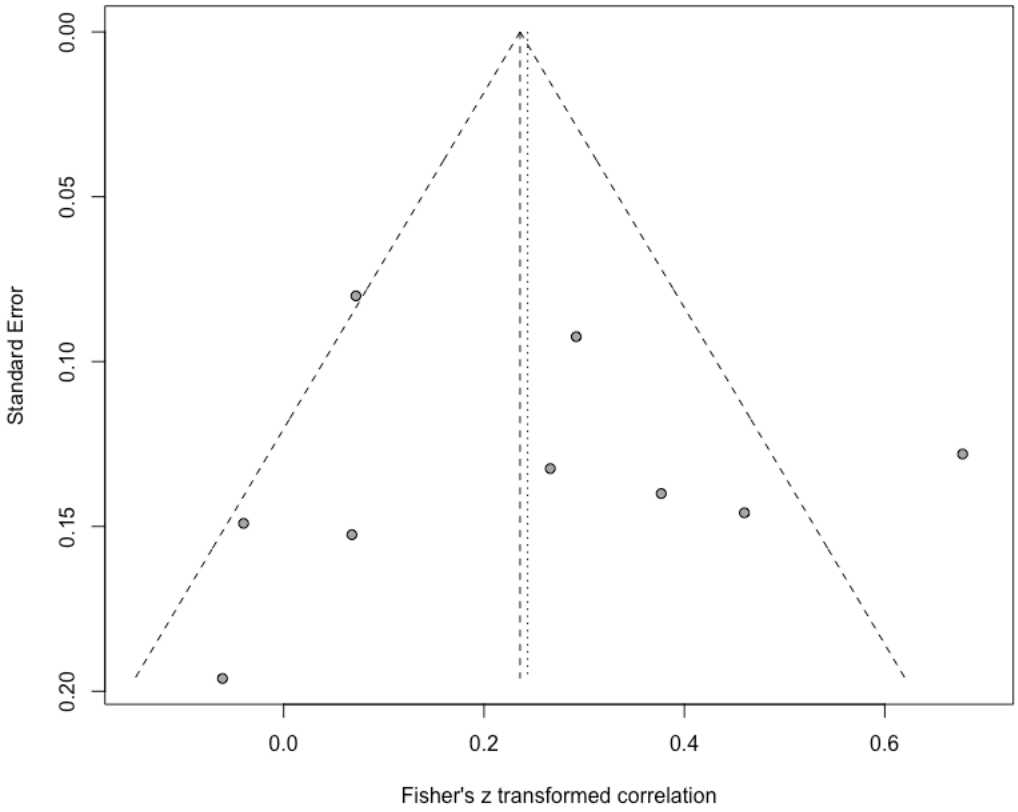

Figure 10. A funnel plot for the accuracy meta-analysis excluding studies with fewer than 40 HDT trials. No evidence of publication bias was observed.

[S17] Separation based on the order of tasks

Two separate meta-analyses were conducted for studies administering the HCT first or counterbalancing the order of the tasks. Only the HCT first analysis produced a significant Q statistic and indicated moderate heterogeneity through its I^2 value (HCT first: $Q = 31.09$, $p = .002$, $I^2 = 61.4\%$; counterbalanced: $Q = 7.10$, $p = .419$, $I^2 = 1.4\%$). The HCT first meta-analysis produced a pooled effect size of 0.23 ($p < .001$) and the counterbalanced meta-analysis produced a pooled effect size of 0.15 ($p = .007$). Due to overlapping confidence intervals (HCT first- [0.11, 0.34], counterbalanced- [0.06, 0.24]), it appears that task order

has no significant effect on the results. Supplementary Figure 11 [S18] displays the forest plots for the two meta-analyses and the funnel plots can be seen in Supplementary Figure 12 [S19]. Both meta-analyses produced a non-significant Egger's test (HCT first: $p = .789$; counterbalanced: $p = .359$) indicating no evidence of publication bias.

[S18] Supplementary Figure 9. HCT first and counterbalanced forest plots

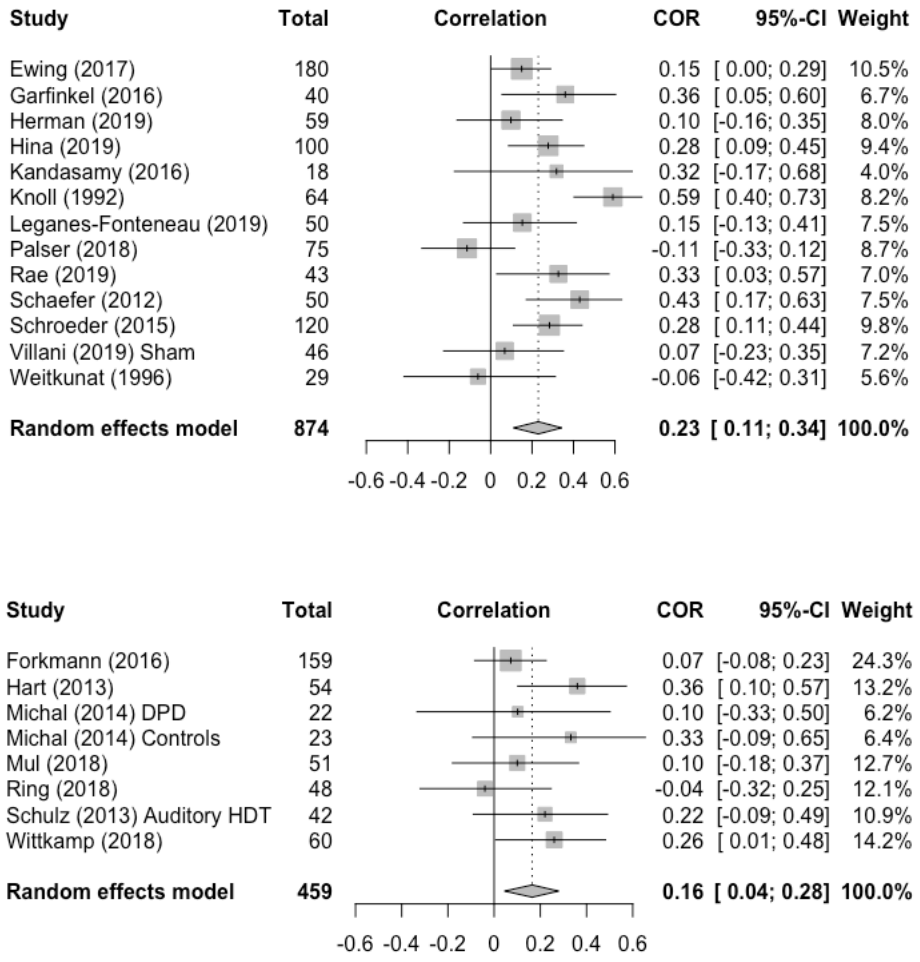

Figure 11. Forest plots for the accuracy meta-analysis separated by task order; HCT first (top) and counterbalanced (bottom) revealed pooled effect sizes of 0.23 and 0.15 respectively (dashed lines) when referring to the random-effects models. Total = the sample size for each study, COR = correlation coefficient, CI = confidence interval, DPD = Depersonalization Disorder, HDT = heartbeat discrimination task, Sham = data from the sham taVNS stimulation condition as opposed to the active condition.

[S19] Supplementary Figure 10. HCT first and counterbalanced funnel plots

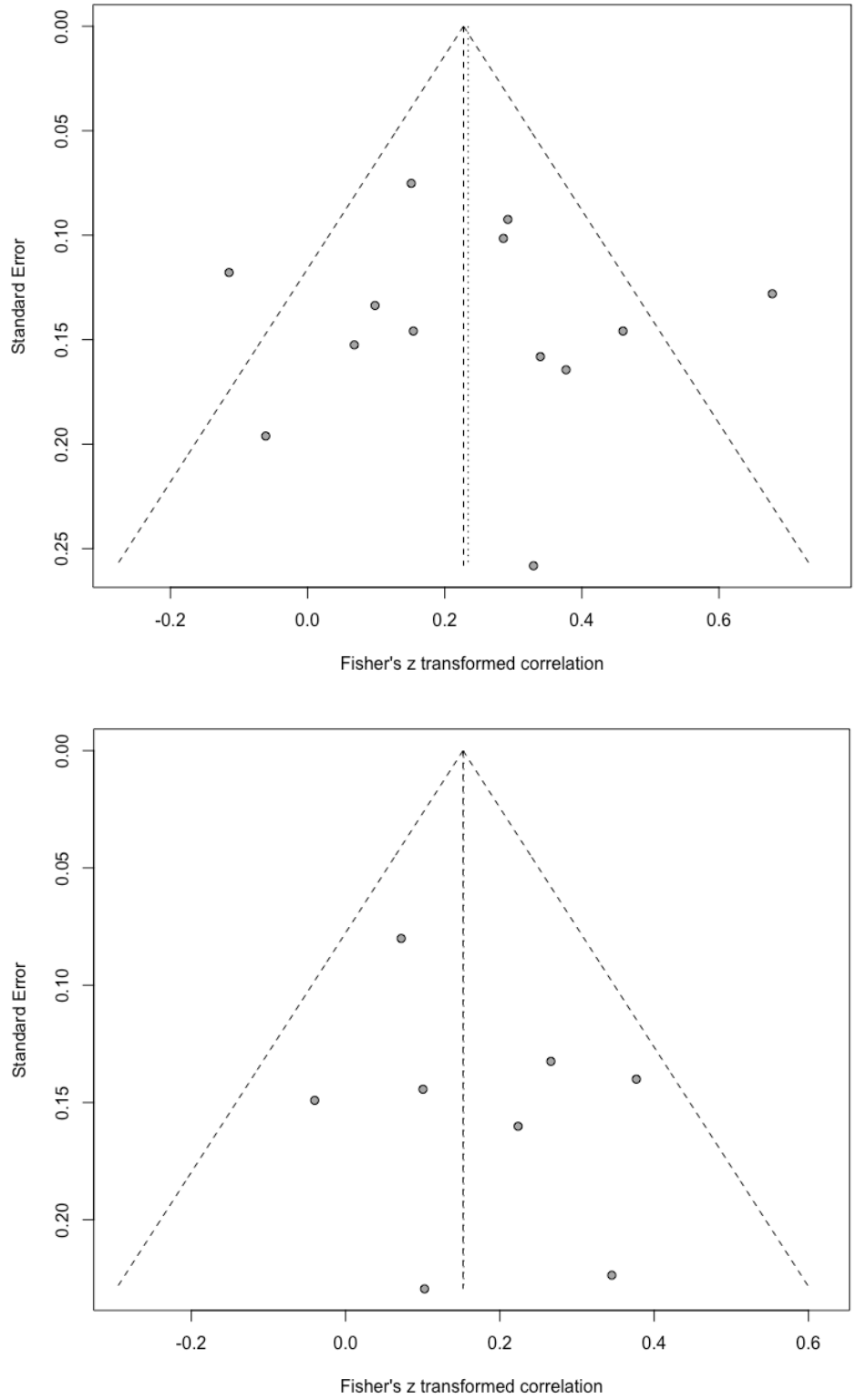

Figure 12. Funnel plots for the HCT first accuracy meta-analysis (top) and the counterbalanced accuracy meta-analysis (bottom). No evidence of publication bias was observed.