The New Measures of Interoceptive Accuracy: A Systematic Review and Assessment

Olivier Desmedt[1], Olivier Luminet[2,3], Marta Walentynowicz[4], & Olivier Corneille[1]

[1] Institute of Psychology, University of Lausanne, Lausanne, Switzerland

[2] Psychological Science Research Institute, UCLouvain, Louvain-la-Neuve, Belgium

[3] Fund for Scientific Research – Belgium (FRS-FNRS)

[4] Centre for the Psychology of Learning and Experimental Psychopathology, KULeuven, Belgium

PREPRINT VERSION

Author Note

Olivier Desmedt https://orcid.org/0000-0002-9496-5465

Olivier Luminet https://orcid.org/0000-0002-1519-2178

Marta Walentynowicz https://orcid.org/0000-0002-7049-7943

Olivier Corneille https://orcid.org/0000-0003-4005-4372

We have no conflict of interest to disclose. Pierre Maurage (Research Associate) and Olivier Luminet (Research Director) are funded by the Fund for Scientific Research – Belgium (FRS-FNRS). Olivier Desmedt is funded by the Swiss National Science Foundation.

Correspondence concerning this manuscript should be addressed to Olivier Desmedt, Institute of Psychology, University of Lausanne, Géopolis, CH-1015 Lausanne, Switzerland. Email: olivier.desmedt@unil.ch.

Abstract

Conscious interoception, the perception of internal bodily states, is thought to contribute to fundamental human abilities (e.g., decision-making and emotional regulation). One of its most studied dimensions is interoceptive accuracy: the objective capacity to detect internal bodily signals. In the past few years, several labs across the world have started developing new tasks aimed at overcoming limitations inherent in traditional measures of interoceptive accuracy. In this systematic review, we identified these tasks (since 2015) for the cardiac, respiratory, and gastrointestinal domains. For each identified task, we discuss their strengths and weaknesses, and make constructive suggestions for further improvement. In the general discussion, we discuss the (potentially elusive) possibility of reaching high validity in the measurement of interoceptive accuracy. We also point out that interoceptive accuracy may not be the most critical dimension for informing current theories, and we encourage researchers to investigate other dimensions of conscious interoception.

Keywords: interoception, interoceptive accuracy, measurement, validity, reliability

The New Measures of Interoceptive Accuracy: A Systematic Review and Constructive Assessment

Interoception, the conscious and non-conscious processing of internal bodily states (Berntson & Khalsa, 2021; Chen et al., 2021), has received increasing interest in recent years. At the conscious level, several dimensions of interoception have been proposed. Nowadays, the most popular terminology comes from the three-dimension model of interoception (Garfinkel et al., 2015), which is composed of interoceptive accuracy (the objective ability to detect internal bodily signals), interoceptive sensibility (the self-reported ability to detect these signals and the tendency to focus on them) and interoceptive awareness (the correspondence between objective and self-reported abilities in detecting internal bodily signals). Interoceptive accuracy, which is measured by performance-based tasks, is the most studied dimension of interoception. It is hypothesized that deficits in interoceptive accuracy are risk factors for low emotional abilities (e.g., awareness and regulation), dysfunctional decision-making and behaviors (e.g., substance abuse and over-eating), and psychopathology (e.g., depression and anxiety).

However, measures of interoceptive accuracy are characterized by important limitations threatening robust theoretical conclusions and effective clinical interventions (Desmedt, Van Den Houte, et al., 2022). In this paper, we focus on the three most studied interoceptive domains: cardiac, respiratory, and gastrointestinal. Below, we provide a discussion of interoceptive accuracy tasks classically used in interoceptive research. In a second step, we report the results of a systematic review of new interoceptive accuracy tasks designed in these bodily domains. In doing so, we discuss the relative strengths and weaknesses of these new tasks and make constructive recommendations for further improvement.

A brief discussion of classical interoceptive accuracy tasks

Cardiac domain.

The cardiac domain has been the most frequently studied one, presumably because heartbeats are easy to measure (contrary to gastric signals, for instance) and are not under participants' direct control (contrary to breathing, for instance). In 2016, Brener and Ring reviewed the existing cardiac tasks. They proposed to classify these tasks into three categories: (1) heartbeat tracking tasks, (2) two alternative forced-choice tasks, and (3) multi-interval tasks.

In heartbeat tracking tasks, participants have to give an accurate estimate of their heart rate by reporting the number of counted heartbeats (the Heartbeat Counting Task; Dale & Anderson, 1978; Schandry, 1981), adjust the rate of an external stimulus (e.g., Carroll & Whellock, 1980; Gannon, 1980), or tap on each heartbeat (e.g., Canales-Johnson et al., 2015; Kleinman, 1970; McFarland, 1975). The main limitation of these tasks is that participants can achieve high performance by guessing their heart rate instead of relying on their felt heartbeats. This has been repeatedly demonstrated for the Heartbeat Counting Task (HCT), the most frequently used measure of interoceptive accuracy (e.g., Desmedt et al., 2018; Phillips et al., 1999; Ring et al., 2015; Ring & Brener, 1996; Windmann et al., 1999), which is also characterized by other limitations (see Corneille et al., 2020; Ferentzi et al., 2022).

In two-alternative forced-choice (2AFC) tasks, participants have to judge whether an external signal (e.g., a tone or a light) is synchronous or asynchronous with their heartbeats (Whitehead et al., 1977). Contrary to the above-mentioned tasks, participants cannot use informed guessing to perform these tasks. They can indeed not guess the synchronicity of external signals with their heartbeats without detecting their heartbeats. However, the main limitation is that only one in four participants generally has an accuracy score (d') equal to or greater than 0.75, the threshold to be classified as a heartbeat detector (e.g., Brener et al., 1993). One potential explanation is that this paradigm implies assuming the moment relative to the R-wave at which individuals should (not) feel their heartbeat, although important individual differences exist regarding the time interval between the R-waves and heartbeat sensations. Consequently, participants may have low scores because they feel their heartbeats within time windows that are mistakenly considered asynchronous to their heartbeats.

To counter this limitation, Brener and colleagues (1993) developed the third category of tasks, based on the method of constant stimuli (MCS). In these tasks, a series of tones are presented at different intervals relative to the R-wave (e.g., R + 0, R + 100, R + 200, R + 300, R + 400, or R + 500 ms). Participants have to judge whether the tones are synchronous or asynchronous with their heartbeats. Scores are indexed by the interquartile range of the distribution of preferred intervals; the more consistent participants are when selecting intervals, the higher their scores. This method is an improvement compared to previous tasks, as the researcher does not need to assume the moment (relative to the R-wave) at which the individual should feel their heartbeats. Unfortunately, this multi-interval task has several limitations, too (e.g., Knapp et al., 1997; Knapp-Kline & Kline, 2005; Schneider et al., 1998). The main one is that it remains difficult to perform, as indicated by the small proportion of individuals (+/- 30%) that can perform the task above the chance level (e.g., Schneider et al., 1998). Furthermore, this task is characterized by long testing times, which represents a burden for participants and makes it difficult to reach high sample sizes.

In addition to these tasks reported by Brener and Ring (2016), we can cite the standardized isoproterenol sensitivity test (Khalsa et al., 2009) in which participants receive isoproterenol or saline infusions in a double-blinded protocol. They are then asked to rate the intensity of perceived cardiorespiratory sensations by rotating a dial. Cardiac interoceptive accuracy corresponds to the within-subject correlation between heart rate and dial ratings. The main advantages of this task are that it allows experimental control over the cardiac signal and prevents informed guessing strategies. However, this task is invasive and requires physician supervision (Khalsa et al., 2009). Moreover, it is not clear whether the effect of bolus infusions on cardiac signal intensity is similar across participants. Also, the effect of bolus infusions is not specific to the cardiac system. Finally, while participants are asked to report the *intensity* of cardiac sensations, the objective measure is heart *rate*.

In summary, as already stated in 2005, “no one method [of cardiac interoceptive accuracy] has emerged in the literature as a clear favorite” (Knapp-Kline & Kline, 2005, p. 394).

Respiratory domain.

In the respiratory domain, resistance-detection paradigms have often been used to assess subjects' respiratory interoceptive accuracy (Zechman & Davenport, 1978). In these tasks, participants are required to breathe into a tube and do one of the following: (1) detect when resistance has been introduced during the inspiration or the expiration (Gottfried et al., 1978; Lopata et al., 1977); (2) discriminate between different degrees of resistive loads (Webster & Colrain, 2000); or (3) detect CO2 level changes (Lopata et al., 1977). These tasks, however, suffer from important limitations (Nikolova et al., 2022; Van Den Houte et al., 2021). First, resistive loads are often aversive, making it difficult to distinguish between the detection of respiratory signals and the experience of negative affect. Consequently, they are also inappropriate for clinical populations (e.g., anxiety disorder patients and those suffering from respiratory distress disorders) that would benefit from them the most. Second, resistances are often characterized by a lack of granularity; the derived thresholds are therefore unreliable and require many trials to stabilize (Harver et al., 1993). Third, these tasks require long testing times (i.e., more than 60 minutes).

Gastro-intestinal domain.

In the gastrointestinal domain, one of the most common methods to measure interoceptive accuracy is mechanical distension (Andresen, 2009). Here, participants are asked to detect mechanical distensions produced by inserting a soft balloon inflated to varying degrees in the colon or esophagus. Due to the invasiveness of these procedures, drink tests have been proposed (Kim et al., 2000). In these tasks, participants are asked to ingest mineral water or a nutrient liquid until satiation or fullness. Another method is the intestinal gas challenge which assesses the tolerance to an infusion of gas (intraduodenal infusion of lipids) in the small intestine (Serra et al., 1998). We can also cite methods that induce chemical (Bernstein and Hammer tests; Bernstein & Baker, 1958; Hammer et al., 2008), thermal (Villanova et al., 1997), or electrical stimulations (Coffin et al., 2004). A research group has even proposed a multimodal stimulation protocol that includes these three types of stimulation (Brock et al., 2008). The problem with these tasks is that they are either too invasive, painful, impractical, and time-consuming or lack experimental control.

Systematic review and assessment of new interoceptive accuracy tasks

As this short introduction indicates, classic measures of interoceptive accuracy suffer from important limitations. Hence, in recent years, several labs across the world have worked on the development of new measures of interoceptive accuracy. In this article, we report the results from a systematic review that identified these new measures in the cardiac, respiratory, and gastrointestinal domains. In doing so, we assess the strengths and weaknesses of these newly developed tasks and make recommendations for further improvement.

Method

This systematic review was conducted in accordance with PRISMA guidelines (Moher et al., 2009). The pre-registration of our protocol can be found at the following address: https://osf.io/hr7jm. The CADIMA platform (https://www.cadima.info/) was used to perform every step of the systematic review.

Inclusion and exclusion criteria

We included original validation studies of new interoceptive accuracy measures and subsequent studies that used them. Because we wanted to focus on the newly developed measures, we restricted the period from 2015 until 2022. We chose 2015 because this is the year when Garfinkel and colleagues (2015) proposed a widely-endorsed terminology allowing more consistency in the terms and definitions. We limited the interoceptive accuracy measures to the cardiac, respiratory, and gastrointestinal domains because these domains are the most studied ones (and fit the current authors' expertise). We included all measures presented by the authors as meant to evaluate the objective capacity to detect cardiac, respiratory, or gastrointestinal signals, whether the term "interoceptive accuracy" was used or not. We included peer-reviewed articles and pre-prints from well-agreed repositories (i.e., OSF Preprints, which aggregates over 30 preprint servers including bioRxiv and psyarxiv). We excluded reviews, book chapters, conference papers, and letters when searching for original validation studies. Finally, we only included English and French papers.

Search strategy

A systematic literature search was performed on December 2, 2021[1] by OD on PubMed, Scopus, PsychINFO, ScienceDirect, Web of Science, and OSF preprints by restricting results to English and French articles (excluding reviews, books, chapters, conference papers, and letters). The following keywords were used: interocept* AND (task*

---

[1] Except from OSF preprints and Web of Science for which the search was made on December 3 and 8, respectively.

OR test* OR assessment* OR measure*) AND (cardiac OR resp* OR gastr* OR gut* OR oesoph* OR esoph* OR intestin* OR colon* OR heart* OR breath*). These keywords were searched in titles, abstracts, and keywords.

Once the original validation studies of interoceptive accuracy measures were identified, we identified subsequent studies that administered these new tasks. This was done by searching for citing studies (peer-reviewed articles and pre-prints) of original papers in Google Scholar. We included any paper that informed us about the validity or reliability of the task (see Extraction below).

Selection process

Titles, abstracts, and full texts of studies were screened independently by OD & MW to identify studies meeting the inclusion/exclusion criteria. Any disagreement between them over the eligibility of particular studies was resolved through discussion with a third author. The inter-rater reliability was excellent (kappa = 0.81).

Extraction

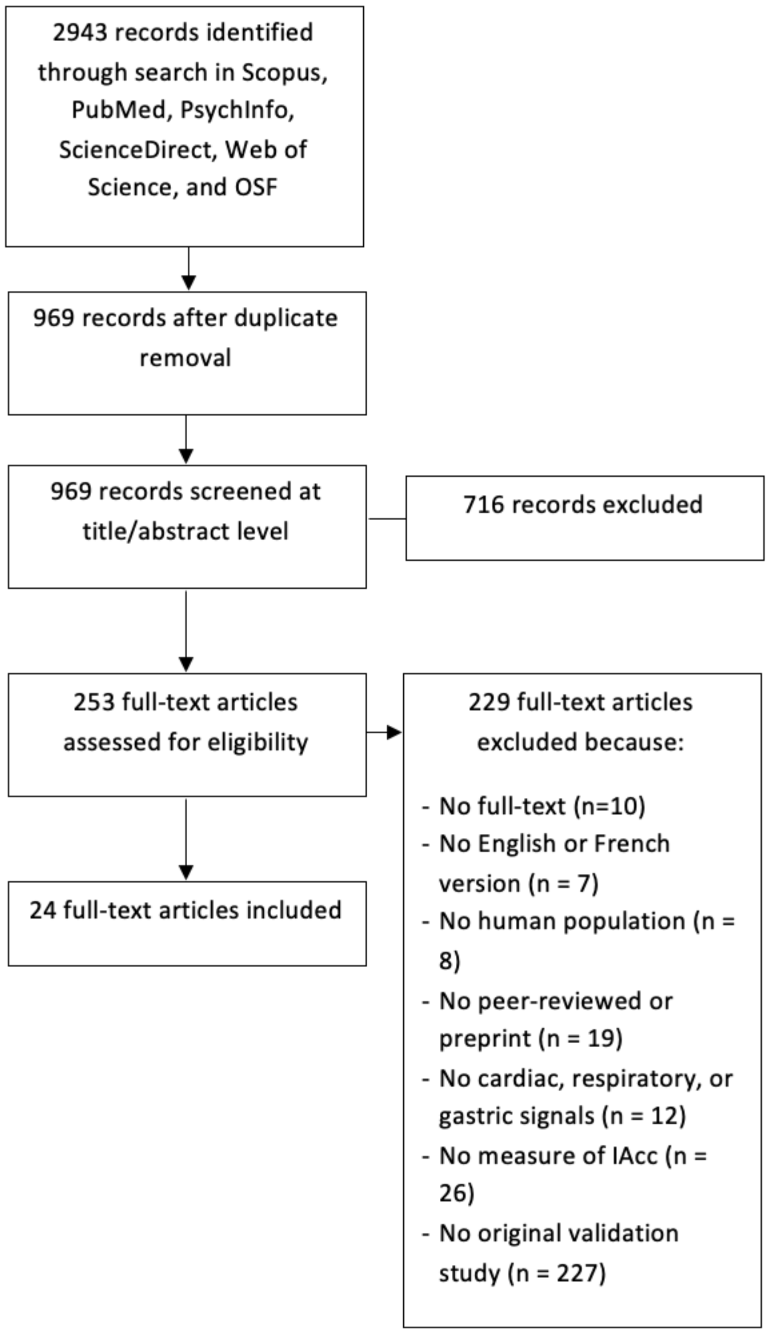
The extraction was performed by OD[2], because we were not able to test inter-rater reliability, as most questions were open. A standardized form was used to extract data from the included studies. The complete extraction of each study can be found at the following address: https://osf.io/79xcn/files/osfstorage.

---

[2] Note that we first planned that two authors extract the data (see pre-registration), but we finally decided that one author does so as no inter-rater reliability could be performed given that our extraction protocol mainly included open questions.

Results

Selection process



Note that we did not include the Heart Rate Discrimination task (HRD; Legrand et al., 2022), as it did not meet our inclusion criteria. As mentioned earlier, we only included measures that were presented by the authors as meant to measure interoceptive accuracy, defined as the objective capacity to detect internal bodily signals. Following the authors of the HRD, “in the absence of further empirical data, interoceptive thresholds and/or precisions obtained by the HRD method should not be interpreted as a straightforward measure of the objective ability to discriminate viscerosensory sensations, as a variety of different strategies utilizing, for example, semantic beliefs or tactile inputs are likely to underlie decisions on the task, in particular under resting conditions (Khalsa, Rudrauf, Feinstein, & Tranel, 2009)” (Legrand et al., 2022, p. 12). That the task is not included in our systematic review does not mean it is useless. In fact, the development of this new task was based on the observation that most studies focus on the objective capacity to detect internal signals (i.e., interoceptive accuracy) while ignoring other interoceptive-relevant processes such as interoceptive beliefs. This is consistent with some of our recommendations (see the section “Main conclusion and future perspectives”).

Identification and assessment in the cardiac domain

We identified eleven new measures of interoceptive accuracy (among thirteen papers), which can be classified into two main categories (see Table 1).

The first category includes synchronicity judgment paradigms. These tasks evaluate individuals’ capacity to synchronize (or judge the synchronicity of) their heartbeats with an external stimulus. In these tasks, participants have to adjust the synchronicity of an external sound with their heartbeats (Plans et al., 2021), tap with their finger in synchrony with their heartbeats (Körmendi et al., 2022; Smith et al., 2020; Smith, Feinstein, et al., 2021), or indicate if they felt a heartbeat during one or more of three very short time intervals (Herman et al., 2021). These tasks are likely to be contaminated by multi-modality processing (participants have to judge or adjust the synchronicity of an external signal or motor response with their heartbeats); as a consequence, they may also be more difficult to perform.

The second category includes heart rate estimation paradigms. These tasks evaluate the subject’s capacity to give an accurate estimate of their heart rate (changes). In these tasks, participants have to count their heartbeats (Larsson et al., 2021; Pohl et al., 2021), adjust the rate of an external stimulus to match their heart rate (Palmer et al., 2019), identify the external stimulus (i.e., a rabbit image, blue ball, or character) that bounces at the same rate as their heart rate (Maister et al., 2017; Yang, Zhou, Wei, et al., 2021; Yang, Zhou, Zheng, et al., 2021), judge whether their heart rate is slower or higher after making jumping jacks (Schaan et al., 2019), or judge several times a day whether their heart rate is higher than usual (Ponzo et al., 2021). These tasks are more likely to be contaminated by informed guessing (participants could use their knowledge about heart rate to perform the tasks), except for two tasks (Larsson et al., 2021; Maister et al., 2017).

Larsson et al.'s (2021) task is less likely to be contaminated by guessing strategies because they evaluate participants' ability to detect heart rate changes in time (vs. heart rate), which cannot be guessed with prior knowledge. Although Maister et al.'s (2017) task would probably be contaminated by guessing strategies in older individuals, their task is meant to be administered to infants (e.g., 5 months) and, more recently, monkeys (Charbonneau et al., 2022). Infants and monkeys may not have the prior knowledge that would allow them to identify the external stimuli that are at a similar pace as their heart rate, but future studies are needed to confirm this assumption.

Based on our systematic assessment of these tasks, more than half of them (54.54%) may be contaminated by informed guessing (Palmer et al., 2019; Pohl et al., 2021; Ponzo et al., 2021; Schaan et al., 2019; Yang, Zhou, Wei, et al., 2021; Yang, Zhou, Zheng, et al., 2021). This is because these tasks request participants to detect their heart *rate*. Instructions that ask participants to only rely on their felt heartbeats (vs. using guessing strategies) to perform the task are effective in *reducing* the contribution of guessing strategies (Desmedt et al., 2020). However, it remains problematic for a performance-based task to be dependent on participants' compliance with instructions. Before we can recommend these tasks, future studies are needed to determine whether strict instructions are sufficient to make these tasks valid. We should also find other ways to control the contribution of guessing strategies.

We also cannot *yet* recommend Herman et al.'s (2021) task, although it is included in the first category, because future studies are needed to determine whether it effectively disentangles accuracy from response bias[3] and whether it is contaminated by heart rate, memory capacity, and multimodality processing (as supported by the significant association with the somatosensory versions of this task ($r(25) = .40$, $p = .043$).

The more promising cardiac tasks and their strengths.

Three new tasks are less likely to be contaminated by informed guessing strategies (Körmendi et al., 2022; Larsson et al., 2021; Plans et al., 2021; Smith, Feinstein, et al., 2021).[4] The first of these, devised by Larsson et al. (2021), requires participants to count their heartbeats, but interoceptive accuracy scores correspond to the intra-individual association between reported and actual beats per minute (bpm). Therefore, participants have to estimate their heart rate variability (as opposed to heart rate), which may go a long way to reducing the probability that they will use their knowledge about their heart rate. In a second task, the Phase Adjustment Task (PAT; Plans et al., 2021), participants are asked to adjust tones (to a similar rate as their heartbeats) until they are synchronized with their heartbeats. The third task of interest is the heartbeat tapping task (as administered by Körmendi et al., 2022; Smith et al., 2020; Smith, Feinstein, et al., 2021) where participants are asked to tap in synchrony with their heartbeats.

---

[3] One epoch duration (750ms) is equal to the cardiac cycle of someone with 80 bpm which is a common heart rate. Consequently, for participants having a heart rate of 80 bpm minimum, at least one heartbeat is present in each epoch. These participants can thus not have false alarms and correct rejections. It follows that, among those participants, the more they report having felt a heartbeat, the higher their scores at the task.

[4] For a task that can be administered in infants, see Maister et al. (2017).

In the PAT and heartbeat tapping task, interoceptive accuracy is indexed via a continuous measure of the beat-to-tap/tone consistency across trials (i.e., tap consistently at the time relative to the R-wave or adjust the tones so that their delay with heartbeats is consistently similar). These tasks assume that to achieve high consistency across trials, one cannot use informed guessing and, instead, needs to accurately detect heartbeats. Smith and colleagues (2020) provided evidence for the absence of informed guessing by showing that instruction modification does not significantly influence interoceptive precision scores. The consistency measure is also a great improvement compared to other measures and computation methods that often require estimating the time windows (relative to the R-wave) in which individuals should feel their heartbeats. Finally, these three tasks have the common advantage of being relatively quick to complete (they last under 5 minutes) compared to, for example, the MCS variant of the HDT (Brener et al., 1993).

Limitations of the more promising cardiac tasks.

Although these three tasks have strengths, it is important to note that they are still characterized by significant limitations and that a more complete psychometric assessment has yet to be performed.

The major limitation of these tasks is their difficulty. In Larsson et al.'s (2021) task, participants have to give an accurate estimate of heart rate variability, which might be very difficult given the relatively low variability of heart rate at rest. Consistently, the mean score for the task was -0.22 (SD = 0.62), which indicates that participants tended to report fewer heartbeats as their heart rate increased. In the PAT and the heartbeat tapping task, participants have to process both their heartbeats and an external stimulus (i.e., their motor response or tones), which might prevent them from detecting their heartbeats. This limitation might be even more pronounced in the heartbeat tapping task in which participants tend to report that button pressing actively interferes with the detection of heartbeats (Flynn & Clemens, 1988;

Pennebaker & Hoover, 1984). Accordingly, some researchers have observed that few participants (from 0.5 to 15.6%) demonstrate a statistically significant beat-to-tap consistency across trials (Flynn & Clemens, 1988; Körmendi et al., 2022). Conversely, Plans et al. (2021) reported that approximately 30% of participants achieve significant consistency across trials in the PAT. It is, therefore, possible that the PAT interferes less with heartbeat detection than the heartbeat tapping task. Future studies are needed to answer this question.

A solution to the high task difficulty is to administer the tasks in conditions of physiological activation to facilitate the detection of heartbeats. In particular, tasks could be administered with varying degrees of cardiac signal intensity (i.e., difficulty). This would maximize the possibility of discriminating between individuals concerning cardiac interoceptive accuracy. One possibility is to implement these tasks at different levels of physiological activation. One way of doing so is to make participants cycle on a stationary bike at different speeds and loads. The tasks could then be administered at rest and just after each small physical exertion. The obvious limitation is that the impact of physical exercise on participants' cardiac signal intensity will differ (see below for a discussion of signal intensity contamination). We, therefore, propose the implementation of a repeated assessment of cardiac signal intensity (e.g., before and after each heartbeat tapping task session). This can be indexed by systolic blood pressure or, more ideally, by "momentum" (i.e., the product between velocity and mass of the blood ejected) which requires impedance cardiography (Schandry et al., 1993; Schandry & Montoya, 1996).

It should, however, be noted that, in addition to myocardial contraction strength, other morphological (e.g., body fat) and mechanical (e.g., postures and devices used to measure heartbeats) parameters could influence signal intensity. Body fat could decrease the possibility for skin receptors to be in contact with the blood conveying cardiac signals. This assumption is supported by studies showing that individuals with less body fat perform better on heartbeat detection tasks (Jones et al., 1987; Montgomery et al., 1984; Rouse et al., 1988), but caution is needed given the limitations of these tasks. We could also expect that some postures are associated with more or less contact between skin and vasculature, but empirical evidence is needed. Furthermore, it has recently been shown that the devices used (e.g., ECG and pulse oximeter) influence participants' performance on the HCT (Murphy et al., 2019). The pulse oximeter was associated with higher performance than ECG, presumably because this former device put more pressure on the participant's skin which increased the contact between the skin receptors and blood. Finally, physical exercise can also trigger other bodily sensations (e.g., muscle pain) that could interfere with heartbeat detection. To mitigate the impact of these confounding variables, researchers could control for body fat, use heart rate devices that minimize the pressure on participants' skin (e.g., ECG), implement physical exercises that are as short as possible, and control for the intensity of other sensations caused by the exercise.

Another limitation of these tasks is the lack of (assessment of) reliability. Larsson and colleagues (2021) reported that their task had poor internal consistency with a correlation of $r = 0.22$ between odd and even trials (vs. $r = 0.99$ for the original HCT). Future studies should try to improve this task's internal consistency, which is a necessary condition for it to be useful. As for the PAT and the heartbeat tapping task, their internal consistency has not yet been evaluated. Test-retest reliability has only been assessed for the PAT for which the authors found an ICC of 0.51, which is considered poor to moderate (Portney & Watkins, 2009), fair (Fleiss, 2011), or good (Cicchetti, 1994) depending on the guidelines. Future studies should, therefore, assess the test-retest reliability of the two other tasks and try to improve it for the PAT. This improvement could be achieved by administering the PAT in a laboratory (because participants performed it remotely in the original study) as it can improve participants' compliance with the instructions[5], decrease external distractions, and increase the reliability of heart rate measurement. This potential high measurement noise can also explain the low between-subject discrimination capacity of the task. Indeed, the PAT only provides a three-category classification of individuals' cardiac interoceptive accuracy (i.e., interoceptive, non-interoceptive, and unknown).

Conclusion for tasks newly developed in the cardiac domain.

Larsson et al.'s (2021) task, the heartbeat tapping task (Körmendi et al., 2022; Smith, Feinstein, et al., 2021), and the PAT (Plans et al., 2021) are promising given their impermeability to informed guessing strategies and response bias, and their low invasiveness and low time cost. However, future studies are needed to assess or improve their reliability, decrease the difficulty of the tasks, and increase their capacity to discriminate between participants.

Table 1

Summary of the new cardiac interoceptive accuracy tasks

[5] Although participants were asked not to take their pulse, we cannot be sure that they complied with this instruction. For instance, participants may have used a smartwatch or even put more pressure on the heart rate device with their fingers so that they could feel their pulse on it.

Study	Guessing	Multi-modality processing	Internal consistency	Test-retest reliability
Maister et al. (2017)	NOT LIKELY (In infants)	YES	NA	NA
Schaan et al. (2019)	LIKELY	NO	NA	NA
Palmer et al. (2019)	LIKELY	MODERATE	NA	NA
Herman et al. (2020)	UNCERTAIN	YES	NA	NA
Larsson et al. (2021)	NOT LIKELY	NO	Split-half: r=0.22	NA
Plans et al. (2021)	NOT LIKELY	YES	NA	r = 0.51 (7 days)
Pohl et al., (2021)	LIKELY	NO	Split-half: r=0.69	r = 0.697 (NA)
Ponzo et al. (2021)	LIKELY	NO	NA	NA
Smith et al. (2021) & Körmendi et al. (2022)	NOT LIKELY	YES	NA	NA
Yang et al. (2021a)	LIKELY	YES	NA	NA
Yang et al. (2021b)	LIKELY	YES	NA	NA

Identification and assessment in the respiratory domain

We identified seven new respiratory interoceptive accuracy tasks, among eight papers (Garfinkel et al., 2016; Murphy et al., 2018; Nicholson et al., 2019; Nikolova et al., 2022; Petersen et al., 2015; Pollatos et al., 2016; Van Den Houte et al., 2021; Wang et al., 2019). Among these new tasks, however, most have not yet been validated and their scores may be contaminated by non-interoceptive processes, such as memory capacity (i.e., participants have to remember the strength of their previous exhalation), exteroceptive processing (i.e., participants could use auditory information about their breathing), breathing control (i.e., task performance could depend on participants' ability to control their breathing strength), or response bias (Murphy et al., 2018; Nicholson et al., 2019; Petersen et al., 2015; Pollatos et al., 2016; Wang et al., 2019)[6]. Two tasks are, nevertheless, more promising: the respiratory occlusion discrimination (ROD) task (Van Den Houte et al., 2021) and the respiratory resistance sensitivity task (RSST; Garfinkel et al., 2016; Harrison et al., 2021; Nikolova et al., 2022).

The more promising respiratory tasks and their strengths.

In the ROD task, participants have to breathe into a tube and discriminate between the length of two consecutive and very short inspiratory occlusions. Respiratory interoceptive accuracy in this task is defined as the discrimination capacity which is measured via an adaptive staircase paradigm to find the length difference that the participants can discriminate 70% of the time (i.e., the Just Noticeable Difference). In the RSST, participants take two breaths on each trial, where one has a resistive load applied to the breathing circuit while the other does not (pseudo-random order). They have to decide whether the first or the second breath is more difficult. Inspiratory resistive loads are delivered by a fully automated and 3D printable respiratory apparatus. A Bayesian adaptive psychophysical approach is applied to estimate interoceptive precision and priors.

The major strength of these tasks is the possibility to disentangle accuracy from response bias and the impossibility for participants to use informed guessing techniques to better perform the task. Moreover, research has shown that these tasks are rather accurate and have good internal consistency. In the ROD task, the authors found acceptable indices of split-half reliability and internal consistency. In the RSST, the researchers demonstrated that correct (vs. incorrect) trials were significantly associated with slower reaction times and higher resistive loads (i.e., easier to detect). While the test-retest reliability of the RSST still

[6] Murphy et al.'s (2018) task has been recently described as an interoceptive propensity (vs. accuracy) measure, i.e., the extent to which individuals use their internal signals to guide their behavior or higher-order cognition.

has to be tested, the ROD task is moderately stable (r = 0.53). Finally, another important strength of these tasks (vs. many previous tasks) is the low aversiveness of occlusions and resistive loads. In the ROD task, the authors argued that very short respiratory occlusions are less likely to be perceived as unpleasant compared to resistive load tasks, but they did not report an assessment of stimuli aversiveness. In the RSST, participants rated resistances as mildly aversive on average.

Limitations of the more promising respiratory tasks.

In respiratory resistance/occlusion paradigms, however, it is difficult to prevent participants from using exteroceptive information (e.g., hearing one's breathing, resistances, or occlusions) to perform the task. In both tasks, several precautions were taken to prevent the use of exteroceptive cues: (1) participants wore noise-canceling headphones (in the ROD task) or headphones playing continuous "rain" noise (in the RSST), (2) the devices were hidden from participants, and (3) participants could not feel the device vibrations. In the ROD task, no significant association was found between the ROD and auditory discrimination task using the same procedure ($r = 0.18$, $p = 0.06$), which suggests that participants did not use auditory information to perform the task. However, the correlation approached significance, which indicates that further investigation is necessary. In the RSST, no empirical test has yet been performed to demonstrate the absence of exteroceptive influences. Finally, although most researchers would agree that auditory, visual, and tactile signals should not be accessible to participants while performing the task, other respiratory-related signals are more ambiguous. Indeed, participants could potentially use a variety of signals to perform the task such as respiratory muscle effort, neck, and chest muscle contractions, and tactile sensations in the mouth (e.g., feeling the air passing through the mouth). The crucial (and open) question is whether all these signals are relevant to respiratory interoception and whether these signals represent what is generally perceived during natural respiratory sensations.

Another known issue of respiratory resistance tasks is the influence of breathing behavior (Zechman & Davenport, 1978). Participants who breathe in more deeply may detect respiratory resistances and occlusions more easily. In the ROD task, the authors have argued that, compared to resistive load tasks, their measure is less likely to be influenced by breathing behavior, as participants are asked to judge the length (vs. intensity) of the stimuli. However, a relatively small but significant association was found between ROD scores and breathing behavior ($r = -0.27$, $p = 0.008$), indicating that deeper inspiration is associated with better performance. Breathing behavior should thus be controlled for in future studies. In the RSST, the researchers designed a visual respiratory entrainment stimulus to standardize the depth of inspiration between participants. However, the authors invite caution when interpreting accuracy scores and recommend recording pressure and flow data in future studies to adjust these scores based on inspiratory pressure.

Respiratory signal intensity could also be influenced by morphological differences such as the diameter of airways (Fritz et al., 1999) and respiratory muscle strength (Campbell et al., 1980). Studies have indeed shown that patients with chronic bronchitis (i.e., increased airway resistance) need a much larger absolute amount of resistance to reach the perception threshold (Wiley & Zechman Jr, 1966). Also, individuals with weaker inspiratory muscles perceive loads as more intense (Campbell et al., 1980; Gandevia et al., 1981). Finally, high-intensity inspiratory muscle training has been shown to decrease the estimated intensity of respiratory loads via the increase of inspiratory muscle strength (Kellerman et al., 2000).

To control for these between-subject morphological and respiratory differences, two solutions have been proposed in the literature. First, instead of computing accuracy scores based on the capacity to detect absolute changes in added resistance, Zechman and Wiley (1977) recommended using the changes in added resistance relative to the subject's internal resistance (i.e., the initial load). The measurement of the subject's airway resistance can be performed by the method of forced oscillation (Fritz et al., 1999). The second solution is a new computer-driven respiratory mouthpiece that continuously measures flow and mouth pressure and maintains airflow resistance constantly by altering the tube's internal diameter (Powell & Williams, 2015). A constant resistive load can therefore be maintained between subjects.

Regarding validity indicators, the ROD task scores were not significantly related to self-report measures of interoceptive sensibility, alexithymia, dyspnea catastrophizing, and positive/negative affectivity. Either these results do not support the task validity, or they do not support the theoretical assumptions behind these tests. In the case of the RSST, no such test has been reported yet.

Finally, the ROD task also has several practical downsides: it requires advanced equipment, it is time-consuming, and it is not entirely automated. Indeed, the experimenter has to manually trigger the pair of occlusions because they should be activated at the right time during inspiration.

Conclusion for tasks newly developed in the respiratory domain.

The ROD task and RSST have several strengths; they are sensitive, reliable, and uncontaminated by response bias and informed guessing strategies. However, further empirical evidence is needed to exclude the possibility of using exteroceptive signals to perform the tasks. Moreover, better control of morphological and behavioral characteristics is desirable. Finally, the construct validity (e.g., convergent, discriminant, concurrent, and predictive validity) of these tasks should be further investigated. In the particular case of the ROD task, the most urgent question is whether the capacity to discriminate the length of inspiratory occlusions is predictive of the capacity to detect respiratory signals, such as inspiratory resistances. To test this question, it is necessary to test the convergence between both tasks.

Identification and assessment in the gastro-intestinal domain

We identified three new tasks for gastric interoceptive accuracy: The Water Load Test – II (WLT-II; Dyck et al., 2016), the glucose test (Young et al., 2021), and the gastrointestinal mechanosensory stimulation task (Smith, Mayeli, et al., 2021).

In the WLT-II, participants have to drink water until perceived satiation and then until maximum stomach fullness. Gastric interoceptive accuracy is defined as the percentage of satiation to total volume. Among the tasks included in this systematic review, the WLT-II is the most frequently used task. Its validity can, however, be legitimately questioned. As the authors themselves have noted, the WLT-II is not an accurate measure of stomach capacity, "because there is a number of confounding variables that this test cannot account for" (Dyck et al., 2016, p. 11). We cannot, therefore, consider this measure as a "standardized and non-invasive test to measure sensitivity for gastric functions, irrespective of stomach volume" (Dyck et al., 2016, p. 11). Indeed, both measures of the amount of water needed to reach satiation (first step) and fullness (second step) may be influenced by several variables, including gastric interoceptive accuracy, response bias, guessing strategies (e.g., based on the number of swallows), other beliefs or attitudes (e.g., the fear of drinking too much water), and stomach capacity.

In the glucose test (Young et al., 2021), participants have to drink glucose and report their hunger level at different times while their blood glucose is measured. Interoceptive accuracy is computed by performing the within-person correlation between blood glucose values and hunger. The limit of this test is that blood glucose is the only biomarker of satiety whereas many other neural, metabolic, and hormonal signals influence satiety. Participants could, therefore, be penalized in their measure of gastric interoceptive accuracy because they rely on other signals than blood glucose (Young et al., 2021).

The more promising gastro-intestinal task and its strengths.

In our view, the gastrointestinal mechanosensory stimulation task (Smith, Mayeli, et al., 2021) is the most promising gastro-intestinal task. In this task, participants have to ingest a vibrating capsule (controlled by the experimenter) and are asked to tap each time they feel a vibration sensation in their stomach. Computational modeling is then applied to compute participants' prior beliefs, interoceptive precision (i.e., the subject's probability to tap or not tap in reaction to the presence or absence of vibration), and learning rates.

This minimally-invasive task is promising as its precision scores are not likely to be contaminated by informed guessing, priors, or multi-modal processing. Several results have also supported its construct validity. First, significant correlations have been found between signal detection theory parameters and their Bayesian parameters. For instance, there was a positive correlation between precision and accuracy, and between priors and the number of false positives. Second, higher precision was associated with faster reaction times, which suggests that participants with higher gastric interoceptive accuracy had less difficulty detecting the vibrations and, therefore, provided faster answers. Third, higher precision was associated with lower perceived difficulty and higher confidence in task performance. Fourth, electroencephalogram (EEG) techniques demonstrated a link between stronger activations in response to capsule vibrations in several parieto-occipital regions (i.e., those generally assigned to perceptual and interoceptive processing) among participants with higher sensory precision. Fifth, more phasic skin conductance changes were observed among participants with higher interoceptive precision. This may indicate that participants who detected more vibrations had also more autonomic reactions to these vibrations.

Limitations of the more promising gastro-intestinal task.

However, this task has one main limitation: the capacity to detect vibrations may be influenced by morphological differences between participants (e.g., the stomach size), which can influence the intensity of vibrations. On a related note, the position of the capsule in the stomach could influence the ease with which participants detect its vibrations. Future studies are needed to test the possibility of controlling for these confounding variables. Another potential limitation is the practicability of the method. Only 40 participants were recruited for the validation study, which is not enough to make stable estimations of, for instance, correlation coefficients. Finally, it is an open question as to whether individual differences in the ability to detect vibratory stimulations generalize to differences in the perception of natural gastrointestinal states.

Conclusion for tasks newly developed in the gastro-intestinal domain.

The gastrointestinal mechanosensory stimulation task appears very promising as a measure of gastric interoceptive accuracy. Preliminary evidence supports its construct validity. Future studies are, nevertheless, needed to further investigate its practicability, reliability (e.g., test-retest stability), construct validity, and ecological validity.

Main conclusion and future perspectives

This systematic review indicates that many tasks have been recently developed to provide more valid and reliable measures of interoceptive accuracy. Unfortunately, our analysis suggests that many of them suffer from limitations that should be solved before we can recommend them and almost no task can yet be considered optimal for measuring interoceptive accuracy. Nevertheless, we identified a set of more promising tasks deserving closer attention and further validation efforts. In the cardiac domain, the most promising tasks are those that prevent participants from using informed guessing strategies: Larsson et al.'s (2021) task, the Phase Adjustment Task (Plans et al., 2021), and the heartbeat tapping task (as administered by Körmendi et al., 2022; Smith et al., 2020; Smith, Feinstein, et al., 2021). In the respiratory domain, we discussed the respiratory occlusion discrimination task (Van Den

Houte et al., 2021) and the respiratory resistance sensitivity task (Garfinkel et al., 2016; Harrison et al., 2021; Nikolova et al., 2022). Finally, in the gastrointestinal domain, we discussed the gastrointestinal mechanosensory stimulation task (Smith, Mayeli, et al., 2021).

Caution is, nevertheless, needed regarding our classification of the tasks as more or less promising. We endorsed a binary classification, as it allows us to select the tasks that we discussed in more depth in the present paper, but it should be clear that their psychometric and pragmatic properties differ continuously. Moreover, our assessment is subject to change as new empirical data become available.

In the remainder of this general discussion, we address two more general questions related to the (1) possibility and (2) necessity of reaching high validity in the measurement of interoceptive accuracy.

Can we design highly valid measures of interoceptive accuracy?

Although a considerable amount of work has been done by dozens of researchers for more than 40 years in the quest to develop valid and reliable measures of interoceptive accuracy, it is fair to conclude that not a single task combines high construct validity, reliability, discriminability of participants, and practicability (including acceptable aversiveness and time cost). Even the most promising tasks await further psychometric investigation and are characterized by important limitations. In particular, we have seen that all existing interoceptive accuracy measures are likely contaminated by the to-be-detected signal intensity, i.e., the strength with which the internal stimuli impact interoceptive receptors.

We consider signal intensity a confounding variable because interoceptive accuracy measures are meant to capture an individual’s ability to detect a given internal signal, which implies that the intensity of this signal should remain constant between individuals. In the exteroception literature, it has long been known that signal intensity influences the ability of individuals to detect it (Green et al., 1957). Therefore, studies in this domain usually standardize measurement based on signal intensity (i.e., participants have to detect signals of the same intensities). As an analogy, during a visual acuity test, the letters are of similar sizes regardless of the patient (Corneille et al., 2020). In contrast, current measures of interoception do not maintain signal intensity constant across individuals. This is as if the sizes of letters in the aforementioned visual acuity test changed between patients. This lack of control over signal intensity in interoception research is probably due to the difficulty of manipulating signal intensity without creating harm or discomfort. Yet, the fact that the problem is difficult to overcome does not mean it should be brushed aside.

In the cardiac domain, signal intensity can be manipulated but cannot be precisely controlled, and, therefore, statistical control (i.e., adjusting accuracy scores based on signal intensity) is necessary. The efficacy of this statistical control, however, depends on the measurement validity and reliability of cardiac signal intensity. Moreover, this statistical manipulation assumes a relationship (e.g., linear) between signal intensity and detection ability. Future studies will, therefore, need to identify the relationship type of these two variables (which has been identified as, for example, an “S” or “ogive” curve in the auditory domain; Kwan et al., 1991) and whether it differs between individuals.

In the respiratory and gastrointestinal domains, the to-be-detected signal can sometimes be precisely controlled by the experimenter. However, these experimental manipulations (e.g., resistances or occlusions applied to a tube in which participants breathe or a balloon or capsule inserted in the gastrointestinal tract of the participants) might not represent the sensations generally felt in everyday life. Perhaps even more importantly, the impact of these experimental stimulations on internal states may be moderated by morphological (e.g., the size of the stomach) and behavioral (e.g., the participant's breathing pattern) characteristics. Future studies are needed to determine whether these moderation effects can be avoided (e.g., with new technologies) or if they should be statistically controlled for. In the latter case, as explained in the previous paragraph for the cardiac domain, statistical assumptions will have to be checked.

In conclusion, it is still unclear whether we can measure interoceptive accuracy with high construct validity: even what we see as the most promising measures are not yet entirely validated and all existing measures are at least potentially contaminated by signal intensity for which we do not yet have full experimental or statistical control. Outcomes of interoceptive accuracy measures should therefore be interpreted with caution.

Is interoceptive accuracy a needed construct? Suggested ways to move forward (or backward…)

Alternatively, we can also question whether a too strong a focus on interoceptive accuracy is consistent with psychological theoretical models that posit a role of interoception. For instance, in embodied theories of emotions, the perception of bodily changes plays an important role, but these theories are often agnostic about the specific interoceptive features at play (e.g., Barrett & Lindquist, 2008). Likewise, emotional awareness has been thought to be related to the awareness and interpretation of interoceptive signals, but interoceptive accuracy does not have a central place in theoretical models (e.g., Craig, 2004; Lane et al., 2015; Smith et al., 2018). Similarly, interoception is seen as contributing to addictive behaviors, as suggested by lesion and neuroimaging studies (e.g., Naqvi et al., 2007; Noël et al., 2013), but these theoretical assumptions are often agnostic about the specific interoceptive features involved (but see Paulus et al., 2009). On a side note, these examples also illustrate the need for more formal theoretical models that include specific predictions (and the ways to empirically prove or refute them) and potential mechanisms by which interoception contributes to other phenomena (e.g., Borsboom et al., 2021).

The agnosticism of theories about the specific interoception dimension involved in other phenomena is sometimes well justified. It is indeed relevant to hypothesize that the more (frequently) individuals feel internal bodily signals, the more (frequently) they will experience, for instance, emotions, drug effects, and food effects. The reason why individuals feel internal sensations is, however, another question for which there are sometimes no specific and exclusive expectations. It can, for instance, be due to signal intensity, the lower-order processing of the nervous system, attentional processes, or interoceptive accuracy. It is, therefore, not theoretically justified to focus solely on interoceptive accuracy when evaluating the role of interoception in other phenomena. Yet, it is one of the most frequently measured interoception dimensions.

We believe that interoception research should quit over-focusing on interoceptive accuracy. Indeed, it has recently been pointed out that other important interoception dimensions should be investigated (Desmedt, Luminet, et al., 2022; Legrand et al., 2022; Murphy, 2022; Suksasilp & Garfinkel, 2022). Suksasilp and Garkfinkel (2022) have advocated for a multi-dimensional assessment of interoception encompassing lower (e.g., the preconscious impact of internal signals) and higher (e.g., interpretation of internal sensations) levels of processing (see Figure 3). In addition to these dimensions, Legrand and colleagues (2022) have recommended investigating interoceptive beliefs and developed a new paradigm meant to measure the accuracy and precision of cardiac beliefs as well as the associated meta-cognitive insight. Murphy (2022) has proposed the construct of interoceptive propensity, which she defines as the degree to which individuals utilize internal signals for gauging internal states (e.g., hunger) and for higher-order processes (e.g., emotions). Finally, we have also recently proposed a hierarchical framework of interoception that includes even more interoceptive processes at different levels of specificity (Desmedt, Luminet, et al., 2022). All of these proposed dimensions of interoception (Desmedt, Luminet, et al., 2022; Legrand et al., 2022; Murphy, 2022; Suksasilp & Garfinkel, 2022) are potential features that modulate the phenomenological end-product of interoception (i.e., interoceptive sensations) and explain its contribution to other phenomena.

References

Andresen, V. (2009). Visceral sensitivity testing. *Best Practice & Research Clinical Gastroenterology*, *23*(3), 313–324.

Barrett, L. F., & Lindquist, K. A. (2008). The embodiment of emotion. *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, 237–262. https://doi.org/10.1017/CBO9780511805837.011

Bernstein, L. M., & Baker, L. A. (1958). A Clinical Test For Esophagitis. *Gastroenterology*, *34*(5), 760–781. https://doi.org/10.1016/S0016-5085(58)80023-2

Berntson, G. G., & Khalsa, S. S. (2021). Neural Circuits of Interoception. *Trends in Neurosciences*, *44*(1), 17–28. https://doi.org/10.1016/j.tins.2020.09.011

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

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. https://doi.org/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). https://doi.org/10.1098/rstb.2016.0015

Brock, C., Nissen, T. D., Gravesen, F. H., Frøkjær, J. B., Omar, H., Gale, J., Gregersen, H., Svendsen, O., & Drewes, A. M. (2008). Multimodal sensory testing of the rectum and rectosigmoid: Development and reproducibility of a new method. *Neurogastroenterology & Motility*, *20*(8), 908–918. https://doi.org/10.1111/j.1365-2982.2008.01126.x

Campbell, E. J., Gandevia, S. C., Killian, K. J., Mahutte, C. K., & Rigg, J. R. (1980). Changes in the perception of inspiratory resistive loads during partial curarization. *The Journal of Physiology*, *309*(1), 93–100.

Canales-Johnson, A., Silva, C., Huepe, D., Rivera-Rei, Á., Noreika, V., Garcia, M. del C., Silva, W., Ciraolo, C., Vaucheret, E., & Sedeño, L. (2015). Auditory feedback differentially modulates behavioral and neural markers of objective and subjective performance when tapping to your heartbeat. *Cerebral Cortex*, *25*(11), 4490–4503.

Carroll, D., & Whellock, J. (1980). Heart rate perception and the voluntary control of heart rate. *Biological Psychology*, *11*(3–4), 169–180.

Charbonneau, J. A., Maister, L., Tsakiris, M., & Bliss-Moreau, E. (2022). Rhesus monkeys have an interoceptive sense of their beating hearts. *Proceedings of the National Academy of Sciences*, *119*(16), e2119868119. https://doi.org/10.1073/pnas.2119868119

Chen, W. G., Schloesser, D., Arensdorf, A. M., Simmons, J. M., Cui, C., Valentino, R., Gnadt, J. W., Nielsen, L., Hillaire-Clarke, C. St., Spruance, V., Horowitz, T. S., Vallejo, Y. F., & Langevin, H. M. (2021). The Emerging Science of Interoception: Sensing, Integrating, Interpreting, and Regulating Signals within the Self. *Trends in Neurosciences*, *44*(1), 3–16. https://doi.org/10.1016/j.tins.2020.10.007

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Coffin, B., Bouhassira, D., Sabaté, J.-M., Barbe, L., & Jian, R. (2004). Alteration of the spinal modulation of nociceptive processing in patients with irritable bowel syndrome. *Gut*, *53*(10), 1465–1470. https://doi.org/10.1136/gut.2003.031310

Corneille, O., Desmedt, O., Zamariola, G., Luminet, O., & Maurage, P. (2020). A heartfelt response to Zimprich et al. (2020), and Ainley et al. (2020)’s commentaries:

Acknowledging issues with the HCT would benefit interoception research. *Biological Psychology*, *152*, 107869. https://doi.org/10.1016/j.biopsycho.2020.107869

Craig, A. D. (2004). Human feelings: Why are some more aware than others? *Trends in Cognitive Sciences*, *8*(6), 239–241. https://doi.org/10.1016/j.tics.2004.04.004

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*(1), 79–85. https://doi.org/10.2466/pms.1978.47.1.79

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. https://doi.org/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. https://doi.org/10.1016/j.biopsycho.2018.09.004

Desmedt, O., Luminet, O., Maurage, P., & Corneille, O. (2022). *Discrepancies in the Definition and Measurement of Interoception: A Comprehensive Discussion and Suggested Ways Forward*. PsyArXiv. https://doi.org/10.31234/osf.io/xd3nj

Desmedt, O., Van Den Houte, M., Walentynowicz, M., Dekeyser, S., Luminet, O., & Corneille, O. (2022). How Does Heartbeat Counting Task Performance Relate to Theoretically-Relevant Mental Health Outcomes? A Meta-Analysis. *Collabra: Psychology*, *8*(1), 33271. https://doi.org/10.1525/collabra.33271

Dyck, Z. van, Vögele, C., Blechert, J., Lutz, A. P. C., Schulz, A., & Herbert, B. M. (2016). The Water Load Test As a Measure of Gastric Interoception: Development of a Two-Stage Protocol and Application to a Healthy Female Population. *PLOS ONE*, *11*(9), e0163574. https://doi.org/10.1371/journal.pone.0163574

Ferentzi, E., Wilhelm, O., & Köteles, F. (2022). What counts when heartbeats are counted. *Trends in Cognitive Sciences*, *26*(10), 832–835. https://doi.org/10.1016/j.tics.2022.07.009

Fleiss, J. L. (2011). *Design and Analysis of Clinical Experiments*. John Wiley & Sons.

Flynn, D. M., & Clemens, W. J. (1988). On the Validity of Heartbeat Tracking Tasks. *Psychophysiology*, *25*(1), 92–96. https://doi.org/10.1111/j.1469-8986.1988.tb00965.x

Fritz, G. K., McQuaid, E. L., Nassau, J. H., Klein, R. B., & Mansell, A. (1999). Thresholds of resistive load detection in children with asthma. *Pediatric Pulmonology*, *28*(4), 271–276. https://doi.org/10.1002/(SICI)1099-0496(199910)28:4<271::AID-PPUL6>3.0.CO;2-C

Gandevia, S. C., Killian, K. J., & Campbell, E. J. M. (1981). The effect of respiratory muscle fatigue on respiratory sensations. *Clinical Science*, *60*(4), 463–466.

Gannon, L. R. (1980). Cardiac perception and the voluntary control of heart rate. *Physiological Psychology*, *8*(4), 509–514.

Garfinkel, S. N., Manassei, M. F., Hamilton-Fletcher, G., den Bosch, Y. I., Critchley, H. D., & Engels, M. (2016). Interoceptive dimensions across cardiac and respiratory axes. *Phil. Trans. R. Soc. B*, *371*(1708), 20160014. https://doi.org/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. Scopus. https://doi.org/10.1016/j.biopsycho.2014.11.004

Gottfried, S. B., Altose, M. D., Kelsen, S. G., Fogarty, C. M., & Cherniack, N. S. (1978). The perception of changes in airflow resistance in normal subjects and patients with chronic airways obstruction. *Chest*, *73*(2), 286–288.

Green, D. M., Birdsall, T. G., & Tanner Jr, W. P. (1957). Signal detection as a function of signal intensity and duration. *The Journal of the Acoustical Society of America*, *29*(4), 523–531.

Hammer, J., Führer, M., Pipal, L., & Matiasek, J. (2008). Hypersensitivity for capsaicin in patients with functional dyspepsia. *Neurogastroenterology & Motility*, *20*(2), 125–133. https://doi.org/10.1111/j.1365-2982.2007.00997.x

Harrison, O. K., Garfinkel, S. N., Marlow, L., Finnegan, S. L., Marino, S., Köchli, L., Allen, M., Finnemann, J., Keur-Huizinga, L., Harrison, S. J., Stephan, K. E., Pattinson, K. T. S., & Fleming, S. M. (2021). The Filter Detection Task for measurement of breathing-related interoception and metacognition. *Biological Psychology*, *165*, 108185. https://doi.org/10.1016/j.biopsycho.2021.108185

Harver, A., Katkin, E. S., & Bloch, E. (1993). Signal-detection outcomes on heartbeat and respiratory resistance detection tasks in male and female subjects. *Psychophysiology*, *30*(3), 223–230. https://doi.org/10.1111/j.1469-8986.1993.tb03347.x

Herman, A. M., Esposito, G., & Tsakiris, M. (2021). Body in the face of uncertainty: The role of autonomic arousal and interoception in decision-making under risk and ambiguity. *Psychophysiology*, *58*(8), e13840. https://doi.org/10.1111/psyp.13840

Jones, G. E., Jones, K. R., Rouse, C. H., Scott, D. M., & Caldwell, J. A. (1987). The Effect of Body Position on the Perception of Cardiac Sensations: An Experiment and Theoretical Implications. *Psychophysiology*, *24*(3), 300–311. https://doi.org/10.1111/j.1469-8986.1987.tb00300.x

Kellerman, B. A., Martin, A. D., & Davenport, P. W. (2000). Inspiratory strengthening effect on resistive load detection and magnitude estimation. *Medicine & Science in Sports & Exercise*, *32*(11), 1859–1867.

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. https://doi.org/10.1016/j.ijpsycho.2008.08.010

Kim, D.-Y., Myung, S.-J., & Camilleri, M. (2000). Novel testing of human gastric motor and sensory functions: Rationale, methods, and potential applications in clinical practice. *The American Journal of Gastroenterology*, *95*(12), 3365.

Kleinman, R. (1970). The development of voluntary cardiovascular control. *Doctoral Dissertations*. https://trace.tennessee.edu/utk_graddiss/6127

Knapp, K., Ring, C., & Brener, J. (1997). Sensitivity to mechanical stimuli and the role of general sensory and perceptual processes in heartbeat detection. *Psychophysiology*, *34*(4), 467–473. https://doi.org/10.1111/j.1469-8986.1997.tb02391.x

Knapp-Kline, K., & Kline, J. P. (2005). Heart rate, heart rate variability, and heartbeat detection with the method of constant stimuli: Slow and steady wins the race. *Biological Psychology*, *69*(3), 387–396. https://doi.org/10.1016/j.biopsycho.2004.09.002

Körmendi, J., Ferentzi, E., & Köteles, F. (2022). A heartbeat away from a valid tracking task. An empirical comparison of the mental and the motor tracking task. *Biological Psychology*, *171*, 108328. https://doi.org/10.1016/j.biopsycho.2022.108328

Kwan, G., Terrio, L., & Oshrin, S. (1991). Fundamentals of Signal Detection. *NSSLHA Journal*, *19*, 11–21.

Lane, R. D., Weihs, K. L., Herring, A., Hishaw, A., & Smith, R. (2015). Affective agnosia: Expansion of the alexithymia construct and a new opportunity to integrate and extend Freud’s legacy. *Neuroscience & Biobehavioral Reviews*, *55*, 594–611.

Larsson, D. E. O., Esposito, G., Critchley, H. D., Dienes, Z., & Garfinkel, S. N. (2021). Sensitivity to changes in rate of heartbeats as a measure of interoceptive ability. *Journal of Neurophysiology*, *126*(5), 1799–1813. https://doi.org/10.1152/jn.00059.2021

Legrand, N., Nikolova, N., Correa, C., Brændholt, M., Stuckert, A., Kildahl, N., Vejlø, M., Fardo, F., & Allen, M. (2022). The heart rate discrimination task: A psychophysical method to estimate the accuracy and precision of interoceptive beliefs. *Biological Psychology*, *168*, 108239. https://doi.org/10.1016/j.biopsycho.2021.108239

Lopata, M., La Fata, J., Evanich, M. J., & Lourenço, R. V. (1977). Effects of flow-resistive loading on mouth occlusion pressure during CO2 rebreathing. *American Review of Respiratory Disease*, *115*(1), 73–81.

Maister, L., Tang, T., & Tsakiris, M. (2017). Neurobehavioral evidence of interoceptive sensitivity in early infancy. *ELife*, *6*, e25318. https://doi.org/10.7554/eLife.25318

McFarland, R. A. (1975). Heart rate perception and heart rate control. *Psychophysiology*, *12*(4), 402–405.

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. https://doi.org/10.1371/journal.pmed.1000097

Montgomery, W. A., Jones, G. E., & Hollandsworth, J. G. (1984). The effects of physical fitness and exercise on cardiac awareness. *Biological Psychology*, *18*(1), 11–22. https://doi.org/10.1016/0301-0511(84)90022-X

Murphy, J. (2022). Propensity to use interoceptive signals: An important individual difference. *Biological Psychology*, *171*, 108326. https://doi.org/10.1016/j.biopsycho.2022.108326

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

Murphy, J., Catmur, C., & Bird, G. (2018). Alexithymia is associated with a multidomain, multidimensional failure of interoception: Evidence from novel tests. *Journal of Experimental Psychology: General*, *147*(3), 398–408. Scopus. https://doi.org/10.1037/xge0000366

Naqvi, N. H., Rudrauf, D., Damasio, H., & Bechara, A. (2007). Damage to the Insula Disrupts Addiction to Cigarette Smoking. *Science*, *315*(5811), 531–534. https://doi.org/10.1126/science.1135926

Nicholson, T., Williams, D., Carpenter, K., & Kallitsounaki, A. (2019). Interoception is Impaired in Children, But Not Adults, with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *49*(9), 3625–3637. https://doi.org/10.1007/s10803-019-04079-w

Nikolova, N., Harrison, O., Toohey, S., Brændholt, M., Legrand, N., Correa, C., Vejlø, M., Jensen, M. S., Fardo, F., & Allen, M. (2022). The respiratory resistance sensitivity task: An automated method for quantifying respiratory interoception and metacognition. *Biological Psychology*, *170*, 108325. https://doi.org/10.1016/j.biopsycho.2022.108325

Noël, X., Brevers, D., & Bechara, A. (2013). A triadic neurocognitive approach to addiction for clinical interventions. *Frontiers in Psychiatry*, *4*, 179.

Palmer, C., Ainley, V., & Tsakiris, M. (2019). *Fine Tuning Your Heart: A novel method for measuring interoceptive accuracy*. PsyArXiv. https://doi.org/10.31234/osf.io/qz7r9

Paulus, M. P., Tapert, S. F., & Schulteis, G. (2009). The role of interoception and alliesthesia in addiction. *Pharmacology Biochemistry and Behavior*, *94*(1), 1–7. https://doi.org/10.1016/j.pbb.2009.08.005

Pennebaker, J. W., & Hoover, C. W. (1984). Visceral perception versus visceral detection: Disentangling methods and assumptions. *Applied Psychophysiology and Biofeedback*, *9*(3), 339–352.

Petersen, S., Van Staeyen, K., Vögele, C., von Leupoldt, A., & Van den Bergh, O. (2015). Interoception and symptom reporting: Disentangling accuracy and bias. *Frontiers in Psychology*, *6*, 732. https://doi.org/10.3389/fpsyg.2015.00732

Phillips, G. C., Jones, G. E., Rieger, E. J., & Snell, J. B. (1999). Effects of the presentation of false heart-rate feedback on the performance of two common heartbeat-detection tasks. *Psychophysiology*, *36*(4), 504–510. https://doi.org/10.1017/S0048577299980071

Plans, D., Ponzo, S., Morelli, D., Cairo, M., Ring, C., Keating, C. T., Cunningham, A. C., Catmur, C., Murphy, J., & Bird, G. (2021). Measuring interoception: The phase adjustment task. *Biological Psychology*, *165*, 108171. https://doi.org/10.1016/j.biopsycho.2021.108171

Pohl, A., Hums, A.-C., Kraft, G., Köteles, F., Gerlach, A. L., & Witthöft, M. (2021). Cardiac interoception: A novel signal detection approach and relations to somatic symptom distress. *Psychological Assessment*, *33*(8), 705–715. https://doi.org/10.1037/pas0001012

Pollatos, O., Herbert, B. M., Mai, S., & Kammer, T. (2016). Changes in interoceptive processes following brain stimulation. *Phil. Trans. R. Soc. B*, *371*(1708), 20160016. https://doi.org/10.1098/rstb.2016.0016

Ponzo, S., Morelli, D., Suksasilp, C., Cairo, M., & Plans, D. (2021). *Measuring interoception: The CARdiac Elevation Detection Task*. PsyArXiv. https://doi.org/10.31234/osf.io/aw8yh

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice* (Vol. 892). Pearson/Prentice Hall Upper Saddle River, NJ.

Powell, T., & Williams, E. M. (2015). Assessing Perceptual Sensitivity of Respiratory Load Using Constant Airway Resistance. *J Pulm Respir Med*, *5*(236), 2.

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

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. https://doi.org/10.1016/j.biopsycho.2014.12.010

Rouse, C. H., Jones, G. E., & Jones, K. R. (1988). The Effect of Body Composition and Gender on Cardiac Awareness. *Psychophysiology*, *25*(4), 400–407. https://doi.org/10.1111/j.1469-8986.1988.tb01876.x

Schaan, L., Schulz, A., Nuraydin, S., Bergert, C., Hilger, A., Rach, H., & Hechler, T. (2019). Interoceptive accuracy, emotion recognition, and emotion regulation in preschool children. *International Journal of Psychophysiology*, *138*, 47–56. Scopus. https://doi.org/10.1016/j.ijpsycho.2019.02.001

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

Schandry, R., Bestler, M., & Montoya, P. (1993). On the relation between cardiodynamics and heartbeat perception. *Psychophysiology*, *30*(5), 467–474. https://doi.org/10.1111/j.1469-8986.1993.tb02070.x

Schandry, R., & Montoya, P. (1996). Event-related brain potentials and the processing of cardiac activity. *Biological Psychology*, *42*(1), 75–85. https://doi.org/10.1016/0301-0511(95)05147-3

Schneider, T. R., Ring, C., & Katkin, E. S. (1998). A test of the validity of the method of constant stimuli as an index of heartbeat detection. *Psychophysiology*, *35*(1), 86–89.

Serra, J., Azpiroz, F., & Malagelada, J. (1998). Intestinal gas dynamics and tolerance in humans. *Gastroenterology*, *115*(3), 542–550. https://doi.org/10.1016/S0016-

5085(98)70133-7

Smith, R., Feinstein, J. S., Kuplicki, R., Forthman, K. L., Stewart, J. L., Paulus, M. P., & Khalsa, S. S. (2021). Perceptual insensitivity to the modulation of interoceptive signals in depression, anxiety, and substance use disorders. *Scientific Reports*, *11*(1), 2108. https://doi.org/10.1038/s41598-021-81307-3

Smith, R., Killgore, W. D. S., & Lane, R. D. (2018). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*, *18*(5), 670–692. https://doi.org/10.1037/emo0000376

Smith, R., Kuplicki, R., Feinstein, J., Forthman, K. L., Stewart, J. L., Paulus, M. P., & Khalsa, S. S. (2020). A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Computational Biology*, *16*(12), e1008484. https://doi.org/10.1371/journal.pcbi.1008484

Smith, R., Mayeli, A., Taylor, S., Al Zoubi, O., Naegele, J., & Khalsa, S. S. (2021). Gut inference: A computational modelling approach. *Biological Psychology*, *164*, 108152. https://doi.org/10.1016/j.biopsycho.2021.108152

Suksasilp, C., & Garfinkel, S. N. (2022). Towards a comprehensive assessment of interoception in a multi-dimensional framework. *Biological Psychology*, *168*, 108262. https://doi.org/10.1016/j.biopsycho.2022.108262

Van Den Houte, M., Vlemincx, E., Franssen, M., Diest, I. V., Oudenhove, L. V., & Luminet, O. (2021). The respiratory occlusion discrimination task: A new paradigm to measure respiratory interoceptive accuracy. *Psychophysiology*, *58*(4), e13760. https://doi.org/10.1111/psyp.13760

Villanova, N., Azpiroz, F., & Malagelada, J.-R. (1997). Perception and gut reflexes induced by stimulation of gastrointestinal thermoreceptors in humans. *The Journal of Physiology*, *502*(1), 215–222. https://doi.org/10.1111/j.1469-7793.1997.215bl.x

Wang, X., Wu, Q., Egan, L., Gu, X., Liu, P., Gu, H., Yang, Y., Luo, J., Wu, Y., Gao, Z., & Fan, J. (2019). Anterior insular cortex plays a critical role in interoceptive attention. *ELife*, *8*, e42265. https://doi.org/10.7554/eLife.42265

Webster, K. E., & Colrain, I. M. (2000). The relationship between respiratory-related evoked potentials and the perception of inspiratory resistive loads. *Psychophysiology*, *37*(6), 831–841.

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. https://doi.org/10.1007/BF00998623

Wiley, R. L., & Zechman Jr, F. W. (1966). Perception of added airflow resistance in humans. *Respiration Physiology*, *2*(1), 73–87.

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. https://doi.org/10.1017/S0048577299980381

Yang, H.-X., Zhou, H.-Y., Wei, Z., Wan, G.-B., Wang, Y., Wang, Y.-Y., Yang, T.-X., Lui, S. S. Y., & Chan, R. C. K. (2021). Multidimensional Interoception and Autistic Traits Across life Stages: Evidence From a Novel Eye-tracking Task. *Journal of Autism and Developmental Disorders*. Scopus. https://doi.org/10.1007/s10803-021-05155-w

Yang, H.-X., Zhou, H.-Y., Zheng, H., Wang, Y., Wang, Y.-Y., Lui, S. S. Y., & Chan, R. C. K. (2021). Individuals with Autistic Traits Exhibit Heightened Alexithymia But Intact Interoceptive-Exteroceptive Sensory Integration. *Journal of Autism and Developmental Disorders*. Scopus. https://doi.org/10.1007/s10803-021-05199-y

Young, H. A., Gaylor, C. M., de-Kerckhove, D., & Benton, D. (2021). Individual differences in sensory and expectation driven interoceptive processes: A novel paradigm with implications for alexithymia, disordered eating and obesity. *Scientific Reports*, *11*(1), 10065.

https://doi.org/10.1038/s41598-021-89417-8

Zechman, F. W., & Davenport, P. W. (1978). Temporal differences in the detection of resistive and elastic loads to breathing. *Respiration Physiology*, *34*(2), 267–277.

Zechman, F. W., & Wiley, R. L. (1977). Effect of chest cage restriction on perception of added airflow resistance. *Respiration Physiology*, *31*(1), 71–79. https://doi.org/10.1016/0034-5687(77)90066-4