D'-tecting the Beat: Refining d' in Heartbeat Detection Tasks

Matias Fraile-Vazquez[1], Alisa Zoltowski[1], Facundo Emina[3,4], Grace Zamora[2,5], Jessica Hazelton[6,7], Paula Salamone[8], Jellina Prinsen[9], William Quackenbush[2,5], Caitlin Convery[2], Martin Dottori[10], and Carissa Cascio[1,2]

[1]Life Span Institute, University of Kansas
[2]Vanderbilt University Medical Center
[3]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Buenos Aires, Argentina
[4]Leloir Institute-IIBBA/CONICET, Buenos Aires, Argentina
[5]Vanderbilt University
[6]The University of Sydney, School of Health Sciences, Sydney, NSW, Australia
[7]The University of Sydney, Brain and Mind Centre, Camperdown, NSW, Australia
[8]Center for Social and Affective Neuroscience (CSAN), Linköping University
[9]Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium
[10]Universidad de San Andres, Buenos Aires, Argentina

***Abstract*—Cardiac interoception—the ability to perceive and interpret one's own heartbeat—is increasingly recognized as fundamental to emotion regulation, cognition, and allostasis, as well as a potential transnosologic biomarker. Interoceptive accuracy—performance on objective behavioral tests of heartbeat detection—can be assessed using behavioral sensitivity indices derived from experimental tasks that leverage temporal information, such as *d'* and mean distance. Despite the richness of temporal dynamics captured during the measurement of interoception, they are often overlooked in current behavioral indices to assess performance. As such, current analytical approaches are unrefined and insufficiently tailored to capture the full complexity of interoceptive processing, impeding our theoretical and clinical understanding of interoception. For instance, while *d'* is theoretically suited for measuring sensitivity, its application is problematic in heartbeat detection tasks (HBT) as it lacks a clear signal-versus-noise distinction. To overcome this limitation, we refined *d'* for HBT by: 1) modifying how signal detection theory (SDT) outcomes—used to calculate *d'*—are extracted, and 2) implementing a pooled z-score approach to avoid dubious statistically assumptions. Thirty-three healthy participants performed a HBT, consisting of two exteroceptive blocks (tapping to a recorded heartbeat sound) followed by two interoceptive blocks (tapping to their own heartbeat), with concurrent electrocardiogram heartbeat recordings. We calculated *d'* using the original design—which uses temporal windows for SDT outcome extraction—, a window-free design, and the window-free design with pooled z-scoring, alongside mean distance—a response frequency-based interoceptive accuracy measure. Compared to the standard approach, our revised metric exhibited a more balanced influence of SDT outcomes on the pooled z-score *d'* ($p < 0.001$, $d > 1$), sensitivity differences across the three *d'* approaches ($p < 0.001$), stronger correlations between *d'* and mean distance ($p < 0.001$), and a more robust, assumption-light framework for interoceptive accuracy assessment. This refined *d'* enhances the precision of accuracy measurements in future interoception studies—enabling more sensitive detection of sensory processing differences in clinical populations—and advances the methodological rigor of cardiac interoception research. Furthermore, new metrics combining both *d'* and mean distance are now possible and should be pursued, given their shared underlying sensitivity factor.**

I. Introduction

Interoception is defined as the conscious and non-conscious processing of internal bodily states (e.g., heartbeat fluctuations, changes in breathing) [5], [9], as opposed to exteroception, which describes the perception of exogenous stimuli (e.g., vision, hearing) [8], [28]. Heartbeat detection is the most studied interoceptive phenomenon, likely due to the easy access researchers have to heartbeat signals through the electrocardiogram (ECG). Crucially, alterations in interoception have been reported in neurodevelopmental, psychiatric, and neurodegenerative populations [3], [34], [36], [43], highlighting the urgency to obtain accurate and sensitive measures of interoception. Over the years, several tasks have been developed to test cardiac interoceptive ability, including counting [13], rate adjustment [33], and tapping tasks [6], [7], [31]. To classify this ability, Garfinkel et al. [19] described the multifaceted nature of cardiac interoception, comprising three interoceptive dimensions that have since become the most common way of organizing the literature: interoceptive accuracy (*IAcc*), sensibility, and awareness. Despite the promise of these dimensions, obtaining accurate and sensitive measures of interoception remains challenging.

The most commonly used metric of cardiac interoception is *IAcc*, which reflects performance on objective behavioral tests of heartbeat detection. Several methods to assess cardiac IAcc have been previously used; for example, counting tasks report *IAcc* as the difference between the numbers of reported and recorded heartbeats [16]; rate adjustment tasks use Gaussian mixture models to classify participants as interoceptive or non-interoceptive [33]; and tapping tasks rely on both frequency

and proximity analyses to evaluate *IAcc* [1], [6], [34]. As a consequence of the various methods used to calculate *IAcc*, its validity has been criticized over the years within the interoceptive literature [16], [17], [44]. For example, Zamariola et al. (2018), previously argued that *IAcc* within counting tasks is biased by heart rate, lacks a clear response-to-heartbeats relationship, and that *IAcc* indexes under-reporting rather than actual sensitivity to heart rate changes. More recently, Desmedt (2023) reviewed *IAcc* across different tasks and pinpointed how rate adjustment task results could be influenced by multimodalities (e.g., auditory information, interference of tapping, and breathing control). These examples underscore the inherent challenges in designing robust experimental paradigms for assessing *IAcc*.

Tapping tasks, unlike counting and rate adjustment tasks, lack most of the aforementioned biases, partly due to their ability to capture the temporal correspondence between a subject's behavioral responses and the R peaks of the cardiac cycle. The R peak, the highest amplitude point of the R-wave, is a component of the QRS complex, which represents the sequential negative and positive electrical deflections caused by ventricular activity (e.g., when the heart beats). Notably, the R-peak is widely used as a baseline in event related potential interoception research for heartbeat evoked potential analysis [11], [12], [29], [34]. Unlike previously used interoception tasks, tapping tasks allow for the rapid generation of interoceptive driven behavioral responses within short periods of time. This feature has lead to the development of two key metrics such as mean distance (MD) [1] and d' from signal detection theory [34]. First introduced by de la Fuente et al. in 2019 [14], *MD* is a frequency-base metric which was designed to quantify how consistently an individual's response periods (amount of time between events) align with their heartbeats. Lower *MD* values, approaching zero (*MD = 0*), indicate a perfect period alignment between response and heartbeat frequencies, suggesting higher *IAcc*. Conversely, higher *MD* values denote greater discrepancies between these periods, reflecting poorer *IAcc*. MD, however, doesn't consider R peak related spatial dynamics of interoception, unlike d'. Building on this temporal correspondence, Salamone et al. [34] introduced an adapted version of *d'*, a sensitivity index derived from signal detection theory (SDT) [20]. In contrast to the frequency-domain metric *MD*, *d'* provides a spatiotemporal measure of *Iacc*, based on the assumption that behavioral responses occurring in close temporal proximity to the R peak should indicate successful discrimination of one's cardiac signals. In this framework, individuals whose responses consistently align in proximity to their heartbeats are considered to exhibit stronger interoceptive discrimination ability, hence stronger *Iacc*.

The development of these domain-specific metrics (*MD* for frequency and *d'* for spatiotemporal analysis) enables tapping tasks to produce two distinct, yet linked, *IAcc* scores, each capturing different facets of interoceptive performance. However, the methods used to apply the d' metric as proposed by Salamone et al. are not without limitations. Specifically, this d' implementation in tapping tasks relies on defining hit and false alarm time windows apriori based on heart rate (HR) thresholds to later derive SDT outputs. This approach introduces two significant caveats. First, the time windows between consecutive R-peaks—within which a response is classified as a hit—are determined based on heart rate (HR) thresholds. However, these windows scale disproportionately across thresholds when expressed as a percentage of the R–R interval: for HR < 69.76 bpm, hit windows span 87% of the inter-beat interval, whereas for HR > 94.25 bpm, they cover only 63.5%. Second, because these windows are fixed in size for each HR threshold, the resulting false alarm windows vary across heartbeats, failing to account for heart rate variability (HRV) and potentially introducing systematic bias in performance estimates.

From a statistical perspective, the use of inverse of the standard normal cumulative distribution function (INCDF) to compute z-values for hit and false alarm rates could also be critiqued due to expected non-normal distributions. In this study, we assume that responses exhibit temporal variability, arising from both predictive coding [4], [30]—where the heartbeat is anticipated before it occurs—and motor action delays, reflecting the time between conscious heartbeat perception and motor execution. We propose that this variability introduces sufficient noise in response times to generate individually random-like hit and false alarm distributions, ultimately resulting in an aggregated uniform distribution that appears stochastic. Because we assume predictive tapping varies by subject, we challenge the assumption that both hit and false alarm rates are derived from the same underlying distribution, thus complicating the interpretation of *d'* as the difference between their INCDF-transformed values (eq. 3). As a result, the standardization process inherent to *d'* and subsequent rate subtraction may not accurately reflect interoceptive performance.

In this study, we systematically evaluated the limitations of *d'* as a measure of interoceptive accuracy. Following this evaluation, we proposed methodological refinements to improve its validity. Finally, we compared the correlations of the original and revised metrics with *MD*, assuming both should reflect the same underlying interoceptive process. By directly contrasting different versions of *d'*, we aimed to determine which provides a more accurate representation of interoceptive accuracy. This enables more precise characterization of bodily sensitivity, with the potential to inform and enhance targeted interventions—such as exercise and mindfulness-based therapies—for clinical populations exhibiting interoceptive deficits.

II. Signal Detection Theory in tapping tasks

A. Sensitivity index - d'

Signal detection theory provides a rigorous mathematical framework for measuring perceptual sensitivity, independent of response bias by isolating a sensitivity metric (*d'*) that reflects the ability to discriminate signal from noise [40].

In a typical SDT experimental paradigm, participants engage in trials containing either target signals or noise, producing one of four possible outcomes: a hit (correctly detecting a target signal), a miss (failing to detect a target), a false alarm (incorrectly identifying noise as a target), or a correct rejection (accurately identifying noise as non-target) (Figure 1a). These outcomes allow for the computation of *d'*, which measures

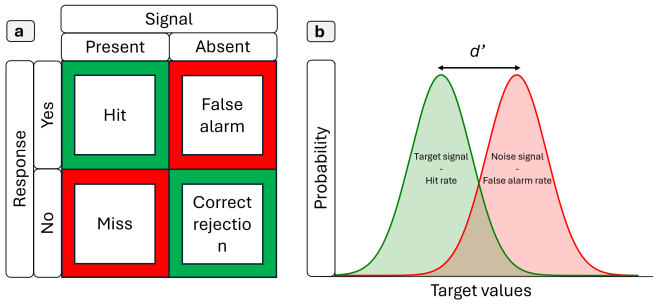

Fig. 1: **(a)** ***hit***: response yes - signal present; ***miss***: response no - signal present; ***false alarm***: response yes - signal absent; ***correct rejection***: response no - signal absent. **(b)** Visual representation of *d'* as the distance in between the hit and false alarm rates

the distance between the standardized hit rate (HitR) and false alarm rate (FAR) (Figure 1b).

Importantly, *d'* encapsulates the subject's ability to distinguish target signals from noise. Positive values (*d'* > 0) indicate successful signal detection with a hit-oriented discrimination pattern, values near zero (*d'* ≈ 0) suggest chance-level performance or inability to discriminate, and negative values (*d'* < 0) reveal a bias towards false alarms. The calculation of *d'* is derived as follows:

Let $x_1, x_2, ..., x_n$ denote n binary observations of trial outcomes, the HitR and FAR are calculated as

$$Rate = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

where *Rate* represents either the HitR or FAR, x_n = 1 corresponds to a hit/false alarm and x_n = 0 a miss/correct rejection for HitR and FAR respectively. Then, HitR represents the probability that a signal is correctly detected and FAR represents the probability that the subject incorrectly detects a signal when no signal is present.

We then defined d' as

$$d' = Z(HitR) - Z(FAR), \tag{2}$$

where $Z(p)$ represents the z-score of a probability p. Finally, in SDT, it is usually assumed that both signal and noise distributions come from normal distributions with different means. That is why the inverse of the normal cumulative distribution function (INCDF) is used, implying that

$$Z(p) = \sqrt{2}\text{erf}^{-1}(2p - 1), \tag{3}$$

where p represents the probability HitR or FAR, and $\text{erf}(p) = \frac{2}{\sqrt{\pi}} \int_0^p e^{-t^2} dt$.

1) Interoception and d': Similar to its application as a sensitivity index for evaluating an individual's ability to discriminate target signals from noise, *d'* can be adapted to tapping tasks to assess a subject's capacity to distinguish their heartbeat (target signal) from the broader continuum of exteroceptive and interoceptive sensory inputs (noise signals). As aforementioned, Salamone et al. (2021) developed a method to derive SDT outcomes from tapping tasks, subsequently calculating an interoception-oriented *d'* to quantify *IAcc* [34]. Applying this approach, they assessed *d'* in dementia patients and healthy controls, finding that dementia patients exhibit reduced sensitivity. To achieve this, their implementation defined the first response after an R peak as a hit if it falls within the designated hit window; any subsequent responses within the hit window and those within the false alarm window are classified as false alarms. No response is considered a miss, and no response following a hit is classified as a correct rejection. The hit window was set at 750 ms after the R peak for HR ≤ 69.76 , 600 ms for 69.76 < HR < 94.25, and 400 ms for HR ≥ 94.25 (Figure 2). From now on we will refer to Salamone's approach as the original design (OD) in contrast to our window-free design (WD) and window-free design with pooled z-score (PD) approach.

Heart rate	R-R Distance	hit window	FA Window	hit window R-R%	FA window R-R%
69.76	860 ms	750 ms	110 ms	87%	13%
82	730 ms	600 ms	130 ms	82%	18%
94.25	630 ms	400 ms	230 ms	63.5%	36.5%

TABLE I: Tapping task hit and false alarm window durations as a function of heart rate. False alarm windows defined according to the values shown. **Row 1**: Heart rates equal or lower than 69.75 bpm. **Row 2**: Heart rates between 69.76 and 94.25 bpm (mean = 82 bpm). **Row 3**: Heart rates equal or higher than 94.25 bpm.

When calculating the HitR (eq. 1), n denotes the total amount of heartbeats, and $x_1, x_2, ..., x_n$ represent n binary observations of hit and miss responses, where x_n = 1 signifies a hit and x_n = 0 a miss. In this scenario, the HitR reflects the proportion of heartbeats accurately detected (responses within a hit window) by the subject.

Regarding the FAR, $x_1, x_2, ..., , x_n$ denotes n binary observations of false alarm and correct rejection responses, where x_n = 1 represents a false alarm and x_n = 0 a correct rejection. The FAR reflects the proportion of heartbeats misidentified (hit window responses - 1 + false alarm window responses) by the subject. D' is finally calculated using the INCDF (eq. 3).

B. Time-based windows limitation

The OD presents a key drawback arising from disproportionate hit and false alarm window sizes, which result from the implemented HR thresholds. Subjects with lower HR exhibit larger R-R hit window proportions, while those with higher HR have smaller windows. To illustrate, Table 1 shows how the hit window for individuals with HR ≤ 69.76 bpm covers 87.20% of the R-R peak distance at the threshold, whereas the hit window for individuals with HR ≥ 94.25 bpm accounts for only 63%, effectively providing those with HR ≤ 69.76 bpm a 24.20% *hit* advantage.

This HR bias inherently introduces intra-group variability and can lead to spurious results when comparing groups with differing HR, especially in clinical populations such as those with depression [2], [21], dementia [26], or autism [10], [42], compared to healthy controls (HC).

A second caveat arising from the OD is that hit windows sizes are held constant across all R-R distances. This

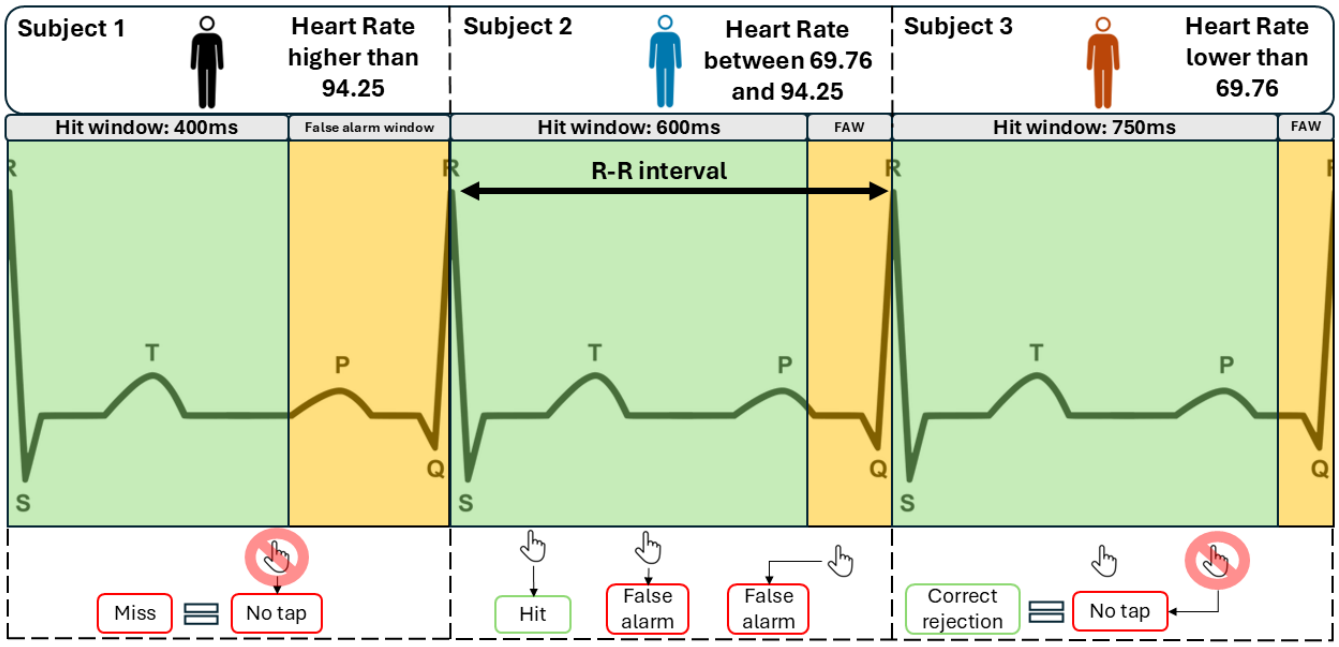

Fig. 2: **Tapping task *d'***: **Subject 1**: The hit window for subjects with a heart rate higher than 94.25 bpm is 400 ms; **Subject 2**: The hit window for subjects with a heart rate between 69.76 and < 94.25 bpm is 600 ms; **Subject 3**: The hit window for subjects with a heart rate lower than 69.76 bpm is 750 ms. Each subject is instructed to press a key whenever they perceive their heartbeat. These responses are classified as follows: ***hit*** (a response within the hit window), ***miss*** (no tap within the R-R peak distance), ***false alarm*** (any response after a hit), and ***correct rejection*** (no response after a hit).

consistency in window size, despite variability in heartbeat frequency, causes the false alarm windows to change in size from beat to beat. Intrinsically linked to the aforementioned bias, once again, such variability becomes particularly problematic when comparing HC to clinical groups known to exhibit atypical HRV values. These differences in HRV may lead to discrepancies in the accuracy of false alarm detection, potentially introducing further bias in the interpretation of interoceptive accuracy across groups.

Finally, and most relevant, the OD assumes that response proximity to the R peak signifies accurate heartbeat detection, as inferred from the definition of the *hit* time windows. However, we argue that this assumption is flawed. Anticipatory responses—arising from predictive coding-where the efference copy accompanying the motor signal predicts the sensory consequences of an action (i.e., perceiving one's heartbeat) [30], [35], [38]—along with the temporal delay required for motor execution, introduce sufficient variability for responses to appear randomly distributed between R peaks. Therefore, we argue that both the use of fixed temporal windows and the assumption of response proximity to the R-peak should be reconsidered.

1) Window-free design: We hypothesize that the OD introduces unwarranted design-driven noise into the estimation of *d'*. To address this limitation, in the current study we proposed an alternative window-free design. In this design, we defined any first response between two R peaks as a hit; any subsequent response within the same R-R distance as a false alarm; no response between R-R distances as a miss; and only one response within an R-R distance as a correct rejection (Figure 3). The formulas for calculating hit and false alarm rates remain the same as in the OD, ensuring consistency in the mathematical framework while addressing the identified biases.

This design eliminates the reliance on fixed windows, mitigating the biases inherent in window size discrepancies. It also provides a more direct assessment of interoceptive accuracy without the confounding influence of HRV or predefined window thresholds. Thus, by not implementing time windows to define SDT outputs, we avoided the potential biases introduced by HR and HRV in the OD. Additionally, we assumed that every first response accurately reflects the ability to sense one's own heartbeat, rather than assuming early responses represent adequate identification of a heartbeat and late responses indicative of false alarms. As a result, tapping task outcomes should be more closely aligned with those from counting tasks, yielding more hits and fewer false alarms. Moreover, by leveraging the temporal domain, d' penalizes *IAcc* when multiple responses occur within a single cardiac cycle, thereby enhancing tapping tasks reliability in estimating sensitivity indices.

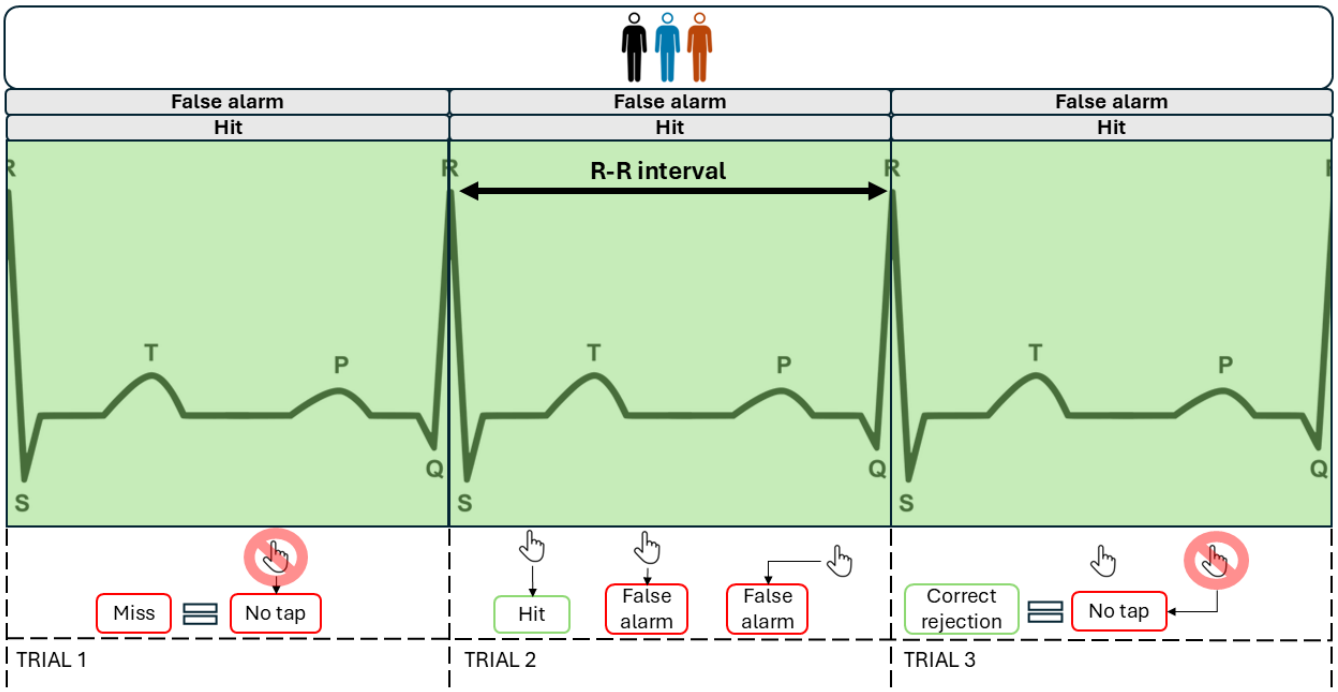

Fig. 3: **Proposed *d'*** window-free design - ***miss*** (Trial 1): no tap inside an R-R peak distance; ***hit*** (Trial 2): a response inside a R-R peak distance; ***false alarm*** (Trial 2): amount of extra responses inside R-R distance; ***correct rejection*** (Trial 3): no response after a hit within an R-R distance

C. Standardizing rates limitation

Standardization is the process of applying a linear transformation to a random variable so that its mean is zero and its standard deviation is one. To achieve this, the z-score formula (z_i) rescales the set of n measurements by subtracting the mean (μ) and dividing by the standard deviation (σ),

$$z_i = \frac{x_i - \mu}{\sigma}, \quad \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}, \quad \mu = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{4}$$

For individual probability values assumed to come from a normal distribution, the inverse of the normal cumulative distribution function (INCDF) can be utilized. The normal cumulative distribution function (NCDF) returns the probability p of X being less than or equal to a value x expressed as

$$P(X \leq x) = F(x). \tag{5}$$

Its inverse form, the INCDF, determines the value x that corresponds to a specific probability p under the assumption that x follows a normal distribution, expressed as

$$F^{-1}(p) = x. \tag{6}$$

The NCDF can be understood as answering the question: *For a given value x, what proportion (p) of the data falls below it in a normal distribution?* In contrast, the INCDF reverses this perspective, asking: *Given a proportion p, what value x corresponds to that percentage of the data being below it?*.

In the OD, the INCDF is employed to compute the z-values of HitR and FAR. The z-score function incorporates the distribution's parameters (μ, σ), (eq. 4), whereas the INCDF does not (eq. 3). As a result, non-Gaussian distributions lead to discrepancies between the outputs of these two functions. In fact, according to Stanislaw and Todorov (1999),

> "SDT states that d' is unaffected by response bias (i.e., is a pure measure of sensitivity) if two assumptions are met regarding the decision variable: (1) The signal and noise distributions are both normal, and (2) the signal and noise distributions have the same standard deviation" [39].

Under the assumption that responses are tied to R-peak latency (yielding a skewed signal distribution favoring HitR), and given our assumption of stochastic-like responses (with uniform signal and noise distributions), a conflict arises: both signal and noise distributions are not expected to be normal. This non-normality should be reflected in the distributions of HitR and FAR, which are expected to follow the underlying signal (heartbeat) and noise (interoceptive–exteroceptive continuum) distributions. Consequently, the assumption of normality inherent in the INCDF is violated, which complicates the interpretation of the resulting z-values.

We argue that hit and false alarm rates do not necessarily arise from the same distribution. In the OD, the HitR is biased toward higher values due to responses closeness to the R peak, conversely constraining the FAR to remain low. In contrast, in

the WD, within-subject variability in predicted responses and motor lag leads to continuously shifting distribution shapes leading to an aggregated uniform distribution of hits. Also, due to an expected reduction of false alarms due to the no window-design, we expected the FAR to showcase low values. These distributional differences would complicate the interpretation of *d'* sensitivity when derived using the INCDF, as the normality assumption underlying the INCDF would not hold, rendering the sensitivity estimates harder to interpret.

In conclusion, we argue that calculating *d'* for tapping task data using the INCDF sacrifices empirical information by assuming the HitR and FAR stem from the same normal distribution, thereby gaining interpretational power through reliable z-value subtraction. In contrast, the z-score function sacrifices interpretational power by reducing the reliability of z-value subtraction, but preserves empirical information by deriving z-values from the actual distributions of the populations.

1) Solving standardization: To avoid reliance on the INCDF, we propose an intermediate approach between z-values and the INCDF, leveraging the combined mean and standard deviation of HitR and FAR to compute z-values for each rate. This method treats both distributions as a unified entity, integrating their descriptive statistics into the z-value formula (eq. 4).

Let x be a vector containing n probability values of both HitR and FAR distributions per subject, we calculate the mean (μ), and standard deviation (σ) to then calculate the z-values for each rate following eq. 4.

By using this approach, we aim to balance interpretational power and empirical information usage while mitigating the limitations associated with the INCDF and standard z-score functions.

III. The Frequency Domain of *IAcc* in Tapping Tasks

A. Mean Distance (MD)

MD quantifies the oscillatory coupling between a subject's response frequencies and their cardiac frequency during motor-tracking heartbeat detection tasks [1], [14], [22], [23]. A *MD* value of zero (*MD* = 0) implies perfect alignment between the subject's response periods and their heart rate. As *MD* increases, it indicates a decrease in the interoceptive guidance of the response periods.

To compute *MD*, Abrevaya et al. (2020) [1] implemented ten-second time windows starting from each R-peak (Figure 4a). For each window, they calculated a normalized heartbeat mean, a normalized response mean (eq. 8), and a coefficient of variation (CV) for the responses (eq. 7) (Figure 4b). Given that high CV values reflect large mean-to-deviation variations, a CV $>$ 0.5 threshold was applied to exclude time windows with extreme response frequencies, under the assumption that such extreme variations represent task misunderstanding. Finally, the normalized response mean was subtracted from the normalized heartbeat mean for each window, and an average of the resultant vectors was computed to obtain the *MD* value (Figure 4c).

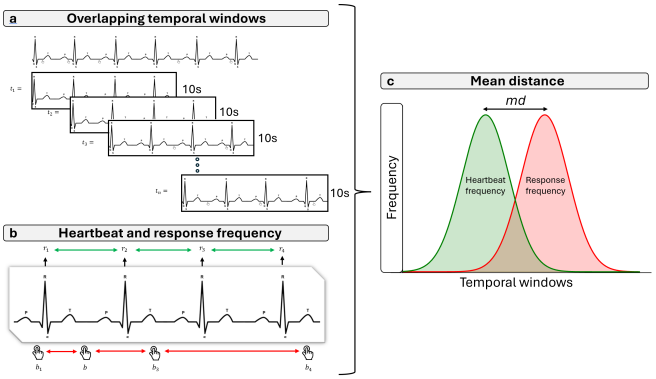

Fig. 4: **(a)** Ten second windows (t_n) starting from the first R peak stepping one R peak at a time. **(b)** Frequencies used for the coefficient of variation and *mean distance*. Heartbeat frequencies (green); Response frequencies (red). **(c)** *Mean distance* as a function of the distance between the mean heartbeat frequencies (green) and mean response frequencies (red)

Let $b_1, b_2, ..., b_n$, represent n observations of behavioral responses within a given time window t. The coefficient of variation is then calculated as

$$CV_{b,t} = \frac{\sigma_{b,t}}{\mu_{b,t}}. \tag{7}$$

Where $\sigma_{b,t}$ represents the standard deviation of the behavioral responses within time window t, and $\mu_{b,t}$ is the mean of the behavioral responses within time window t.

To account for the tapping amount, the normalized mean of $\Delta b = [b_{n-1} - b_1]$ is calculated for every t with CV lower than 0.5 as follows

$$NM_{b,t} = \begin{cases} \frac{n-1}{\Delta b}, & \text{if } CV_{b,t} < 0.5. \end{cases} \tag{8}$$

Next, let $r_1, r_2, ..., r_m$ represent m observations of r-peak latencies within a given time window t, we calculate $NM_{r,t}$ by replacing b by r in eq. 8.

Finally, let $z_1, z_2, ..., z_d$, represent d observations of the distance between the normalized means of tapping responses (NM_b) and r-peak latencies (NM_r), the distance is calculated as

$$\bar{z} = [NM_{b1} - NM_{r1}, NM_{b2} - NM_{r2}, ..., NM_{bd} - NM_{rd}]. \tag{9}$$

We then calculate *MD*

$$MD = \frac{1}{d} \sum_{i=1}^{d} \bar{z}_i. \tag{10}$$

To ensure consistency in heartbeat count across time windows, we segmented the data using three consecutive R-peaks rather than fixed-duration (e.g., 10-second) windows. This approach not only standardizes heartbeat occurrences per segment but also better aligns the temporal density of heartbeat and behavioral response events.

Mean distance leverages the temporal properties of tapping tasks to infer accuracy, effectively nullifying phase-related noise by operating within the domain of frequency distances.

In other words, *MD* reflects how consistently distances between responses and heartbeats align, whereas *d'* quantifies how individual responses and heartbeats align. Both measures incorporate sensitivity as an underlying factor and should therefore exhibit a strong negative (high *d'* = low *MD*) correlation. This sensitivity is what translates into *IAcc*.

Given the aforementioned OD limitations, our goal is to put our novel *d'* designs to the test by comparing how each SDT outcome maps to each other and to *d'* itself under the new underlying accuracy assumptions. Furthermore, we will assess the discriminative power of each *d'* metric by quantifying its correlation with *MD*, grounded in the aforementioned assumption of a shared sensitivity factor. We hypothesized that as the number of responses approached the actual number of heartbeats, *MD* would decrease and *d'* would increase, leading to a stronger negative correlation with *d'* in window-free designs (WD / PD). This could be attributed to an increase in hits due to the absence of temporal windows. Such a finding would suggest that both measures reflect the same underlying interoceptive phenomenon, captured in two distinct domains: temporal (*MD*) and spatiotemporal (*d'*). Since both are influenced by the same factors (misses, false alarms, correct rejections), we argue that they represent two complementary aspects of the same process.

IV. Methods

A. Participants

Our sample included n = 33 healthy control participants (Female = 20, Male = 12, Other = 1, 75.7% White, 24.3% Multiracial / Black / African American) after exclusion. Participant ages ranged from 8-52 years (mean = 26.6, std = 13.4). Study procedures were approved by the Vanderbilt Institutional Review Board, under protocol #191677. All individuals or their primary caregivers gave informed consent and/or assent for study participation in line with the Declaration of Helsinki. Exclusion criteria included factors that could limit the completion and interpretation of perceptual tasks; specifically, uncorrected sensory impairments (e.g., vision or hearing loss), atypical cardiac rhythms (e.g., arrhythmia or premature contractions) (n = 1), or genetic, psychiatric, or neurologic conditions. Participants who were unable to finish the task due to technical issues were also excluded (n = 3).

A shortened version of Abrevaya's Heartbeat Detection task [1], was administered, consisting of two exteroceptive blocks—tapping to a prerecorded heartbeat (80 bpm) as a control condition—and two interoceptive blocks, in which participants tapped in response to their own heartbeat. Heart activity was continuously recorded using electrocardiography (ECG), integrated as an external channel of a 128-channel EGI EEG system. The task was implemented in MATLAB R2024a using Psychtoolbox-3 [27] to ensure response temporal precision. Each block lasted two minutes to ensure sufficient trial numbers, with only the second block of each condition included in the analysis to account for task familiarization.

B. Data preprocessing

Data preprocessing was conducted using the EEGLAB toolbox v2022.1 [15] in MATLAB R2024a while statistical analysis was performed in MATLAB and RStudio. The ECG data for each subject was visually inspected to reject time frames in which the ECG channel's voltage deviated from its usual course. Once continuity of ECG was ensured, an automated version of HEPLAB [32] was used applying the Pan-Tompkins peak detection algorithm to mark R peaks under a 500ms R-R peak distance detection assumption. A second visual inspection was conducted, this time manually correcting for missing or wrongly marked R peaks. Finally, heartbeat events were renamed to indicate which task block (interoceptive or exteroceptive) they belonged to.

C. Statistical analysis

To assess whether behavioral responses were modulated by interoceptive processes, each subject's response was temporally labeled to the R–R interval in which it occurred. Specifically, each response was normalized as a percentage of the corresponding R–R interval duration. These normalized values were then converted into vectors defined by both direction and magnitude, where direction corresponded to the R–R interval percentage (transformed into angular space), and magnitude reflected the number of responses within each interval bin (1% width).

Each response was thus converted to an angular value θ as

$$\theta_i = \frac{a_i}{100} \cdot 2\pi. \tag{11}$$

To compute the overall directional characteristics of the distribution, the cosine and sine components of each vector were summed for k bins

$$X = \sum_{i=1}^{k} j_i \cos(\theta_i), \quad Y = \sum_{i=1}^{k} j_i \sin(\theta_i), \tag{12}$$

where j_i denotes the magnitude in bin i. The mean vector length (MVL), representing the concentration of the distribution around the unit circle, was then calculated as

$$\text{MVL} = \frac{\sqrt{X^2 + Y^2}}{\sum_{i=1}^{k} j_i}. \tag{13}$$

Finally, the mean angle of the distribution (θ_{mean}) was determined using the two-argument arctangent function

$$\theta_{\text{mean}} = \arctan\left(\frac{Y}{X}\right). \tag{14}$$

This approach provides a robust quantification of both the directional bias and the consistency of temporal responses within the cardiac cycle.

To further validate the observed patterns, we conducted two stochastic simulations using the same analytical framework as applied to the empirical data. The first simulation drew samples from a standard uniform distribution over the open interval (0,1), while the second employed random numbers drawn from a lognormal distribution with a mean of zero and standard deviation of one ($\mu = 0, \sigma = 1$) reflecting bias towards the first R-peak. To ensure comparability across conditions, we generated an equivalent number of responses

per subject per R–R interval. For each simulated subject, the first response within each R–R interval was assigned either uniformly or according to the skewed distribution. Subsequent responses within the same interval were constrained to fall within the temporal boundaries of the initial response and the second R peak, maintaining temporal structure.

To statistically assess group-level differences, we employed a non-parametric permutation test with 10,000 iterations for each of the following comparisons: interoception versus skewed simulation, interoception versus uniform simulation, and interoception versus exteroception. In each iteration, group labels were shuffled and the mean vector length (MVL, eq. 13) was recomputed for both groups, generating a null distribution of MVL differences. The empirical MVL difference was then compared against this distribution, and a two-tailed p-value was estimated as follows

$$p = \frac{1}{n}\sum_{i=1}^{n} \Theta\left(|\text{perm}_{\text{diffs}}|_i - |\text{empirical}_{\text{diffs}}|\right), \qquad (15)$$

where n is the total amount of permutations and $\Theta(x)$ is the Heavyside function such that $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ if $x \leq 0$.

This approach allowed us to assess whether observed differences in directional concentration exceeded what would be expected by chance. Thus, we were able to assess the presence or absence of a response bias towards the R-peak.

After exploring the response distribution and contrasting them to random responses, a Wilcoxon signed-rank test was conducted for each rate to assess differences between the two designs (OD-WD)-hit, miss, false alarm, correct rejection, HitR, FAR. To evaluate the discriminatory power of the new method, we computed receiver operating characteristic (ROC) curves using the hit and false alarm rates [25]. The area under the curve (AUC) was calculated for each and a confusion matrix was plotted to visually assess changes in response classification.

Next, we compared the OD against the WD and PD, as well as the WD against the PD (eq. 13–15), using a Friedman's test, followed by post-hoc Wilcoxon signed-rank tests. The effect size d_c (eq. 16) for each variable pair was computed using Cohen's d formula for intra-group comparisons

$$d_c = \frac{M_d}{SD_d}, \qquad (16)$$

where M_d is the mean of the change scores (i.e., the mean difference between the two conditions), and SD_d is the standard deviation of the change scores

$$SD_d = \sqrt{SD_1^2 + SD_2^2 - 2 \cdot r \cdot SD_1 \cdot SD_2}, \qquad (17)$$

where SD_1 is the standard deviation of the first condition (OD), SD_2 is the standard deviation of the second condition (WD / PD), and item r is the Pearson correlation between the two conditions.

Effect sizes for each variable were then visually assessed using a forest plot. A Spearman rank correlation matrix was constructed to evaluate the relationships between SDT variables and *d'* in all three designs, with p-values corrected for multiple comparisons using the false discovery rate.

To further validate our method, we performed cross-domain correlations with *MD*, also used to evaluate interoceptive accuracy.

V. Results

1) Response distribution: Polar representations of both simulated and empirical block responses revealed distributions consistent with uniformity across all empirical conditions, with no dominant directional bias evident. Mean vector lengths, represented as directional vectors, further supported the absence of significant phase locking in all distributions except for the simulated skewed Gaussian condition, which exhibited a pronounced directional preference (Figure 5).

As anticipated, no statistically significant differences were observed between responses in the interoceptive block and those generated from a simulated uniform distribution ($MVL_1 = 0.0376, MVL_2 = 0.0322, p = 0.6418$), nor between interoceptive and exteroceptive conditions ($MVL_1 = 0.0376, MVL_2 = 0.0293, p = 0.4294$). In contrast, a highly significant divergence emerged when comparing the interoceptive responses with those derived from a simulated skewed Gaussian distribution ($MVL_1 = 0.0376, MVL_2 = 0.7417, p < 0.001$), consistent with the predicted sensitivity of the measure to distributional asymmetries (Figure 6).

These findings confirm our hypothesis that responses are temporally stochastic, thereby refuting the assumption that responses align temporally with R peaks. Additionally, albeit no differences were found between the exteroceptive and interoceptive blocks, we cannot rule out the possibility that interoceptive processes in the interoceptive block may be present but not be captured by temporal proximity. As no temporal proximity effect to the R-peak was observed, we tested our WD (Figure 3).

2) Design comparisons: Wilcoxon signed-rank tests were conducted on each measure—hit, miss, false alarm, correct rejection, HitR, and FAR. For plotting purposes, values were subsequently normalized between 0 and 1 based on the minimum and maximum of each variable. As expected, highly significant differences were observed across all comparisons ($W_{\text{hit}} = 0$, $W_{\text{fa}} = 561$, $W_{\text{cr}} = 0$, $W_{\text{hitr}} = 0$, $W_{\text{far}} = 561$, $Z_{\text{hit}} = -4.936$, $Z_{\text{fa}} = 5.014$, $Z_{\text{cr}} = -5.0124$, $n = 33$, $p < 0.001$; Figure 7a), with the exception of misses, which remained unchanged between designs due to the identical calculation method. The statistic W informs about the direction of the rank differences, where positive W values imply consistently higher values for the second variable and values closer to 0 imply either negative or random differences. These differences reflect significant shifts not only in our assumptions of interoceptive driven responses but also in how we interpret *d'* results from the new derived SDT outcomes.

To assess performance, we evaluated the ROC curves of both designs using the hit and false alarm rates outcomes. The WD demonstrated a significantly higher area under the curve (AUC = 0.994) compared to the OD design (AUC = 0.927; Figure 8a). This improvement was primarily driven by

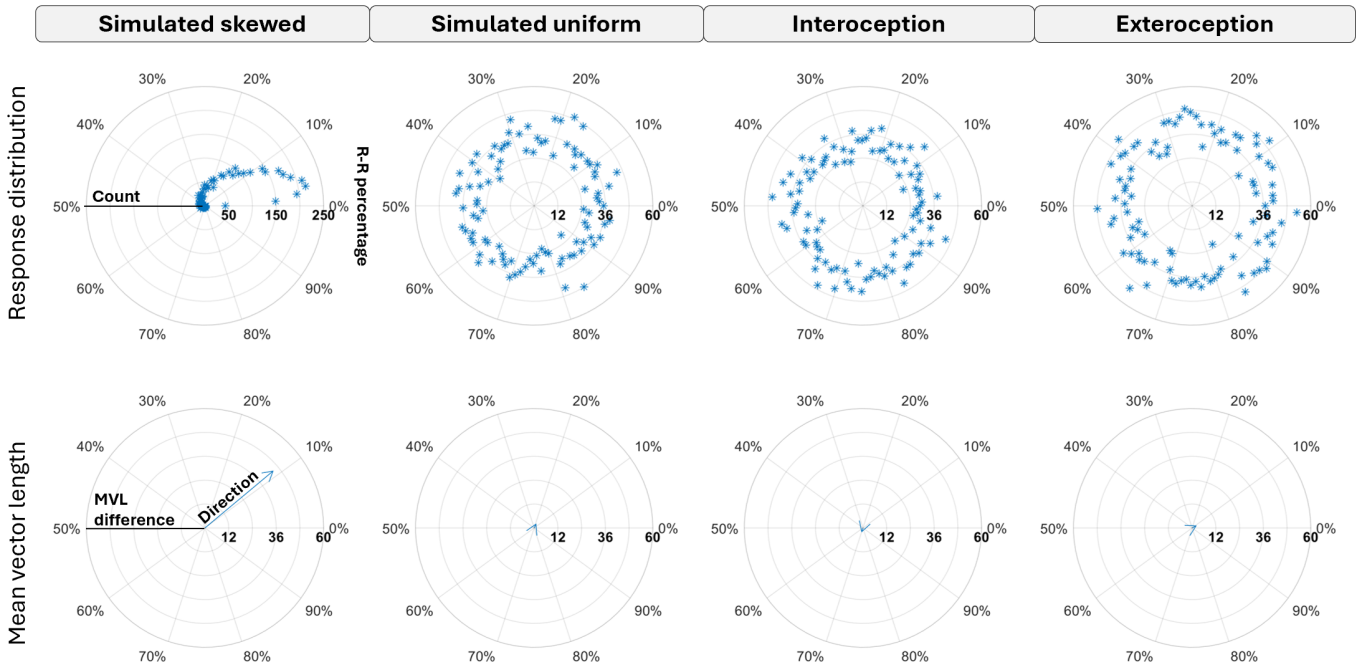

Fig. 5: **Top row**: Polar plot showing the R-R temporal correspondence of responses for each distribution. For the exteroceptive condition, the R peaks of the heartbeat audio were used. The radius indicates the proportion of responses within each percentage bin, while the angle represents the percentage at which the response was executed. **Bottom row**: Mean vector length and angle. The radius represents the mean vector length, which reflects the difference between quadrants of the response polar plot. The direction of the vector indicates the temporal bias of the responses.

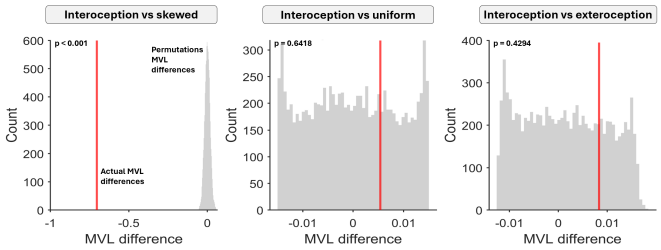

Fig. 6: **Non-parametric permutation analysis**: Distributions of mean vector length differences from simulated permuted responses between groups are shown in grey. The real mean vector length difference is indicated in red. The p-value represents the proportion of simulated results (in absolute values) that are as extreme or more extreme than the observed values from the real response distributions.

a reduction in false alarms, and consequently, a lower FAR, as confirmed by the confusion matrix (Figure 8b). A DeLong's test identified a significant difference across conditions ($p = 0.047$), likely driven by a ceiling effect in classification values. This effect arose from consistently low FAR in both designs, suggesting a constrained range that may have amplified the observed differences.

3) Statistical differences: A Friedman test was conducted to compare the three computed *d'* measures: OD, WD, and PD. The analysis revealed a significant effect across the three groups ($\chi^2 = 62.06$, df = 2, $p < 0.001$). To further investigate this effect, a post hoc analysis was performed.

Wilcoxon signed-rank tests confirmed highly significant differences between the groups ($OD-WD : W = 0, z = -5.01$; $OD - PD : W = 5, z = -4.92$; $WD - PD : W = 560, z = 4.99$; $p < 0.001$, Figure 7b). Effect size calculations indicated extremely large differences ($OD - WD : d = -2.52, SD = 0.12$; $OD - PD : d = -1.72, SD = 0.07$; $WD-PD : d = 2.42, SD = 0.11$, Figure 7c), suggesting that these results were driven by systematic mean differences rather than high variance. Given that the WD and PD, which share the same window-free approach, exhibit significant differences, we can infer that pooling alters the results. This also suggests that the distributions of the HitR and FAR differ.

To further explore how the OD and WD outcomes influence the OD and PD *d'* values, an FDR-corrected Spearman rank correlation matrix was conducted (Figure 9). Table 2 highlights significant Spearman correlations and contrasts SDT outcomes for the OD and WD.

The observed increase in the negative correlation between hits and misses (row 1, col (WD), r = -0.75, $p < 0.001$) in the WD suggests a stronger trade-off between true positives and false negatives, reinforcing the notion that window-free designs more effectively differentiates response tendencies. The more the participant hits, the less they miss. As more responses are considered hits, this correlation strengthened. Furthermore, the correlation between misses and hit rate also strengthened (row 8, col (WD), r = -0.98, $p < 0.001$), implying a greater influence of misses on overall detection performance. Similarly, the negative correlation between misses and false alarms strengthened (row 6, col (WD), r = -0.69, $p < 0.001$),

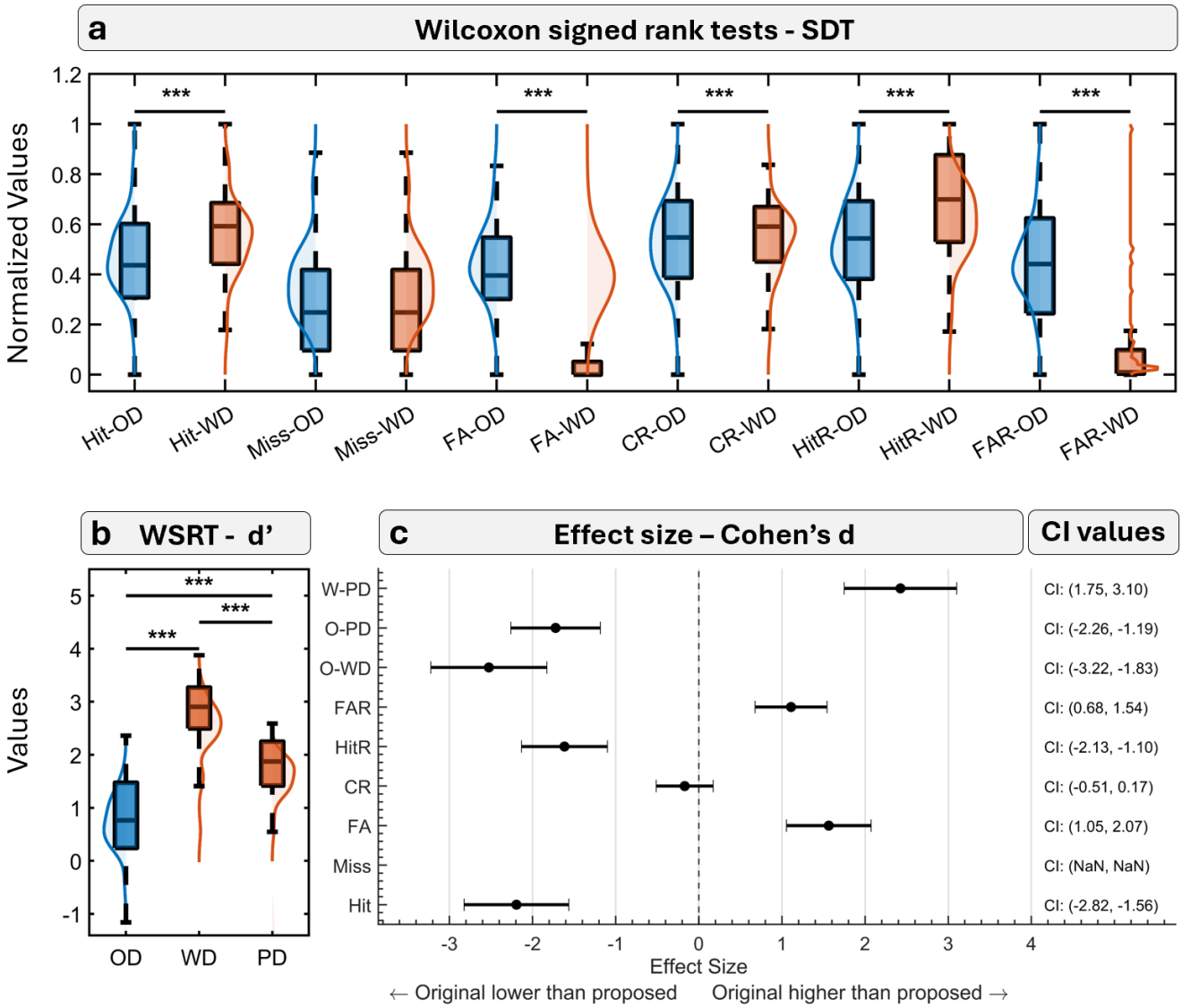

Fig. 7: Statistical significance: p < 0.001 (***). (a) Variables are labeled as either original design (OD) or window-free design (WD) and include, in order: hit, miss, false alarm, correct rejection, hit rate, and false alarm rate. Each pair represents a Wilcoxon signed-rank test comparing the same variable across the two designs (OD vs. WD). For density plot visualization, all values were normalized to a [0,1] range using each variable's maximum value as the upper bound. (b) Wilcoxon signed-rank test comparisons of *d'* values across three designs: OD, WD, and PD (window-free with pooled z-score). (c) Cohen's d effect sizes with confidence intervals (CIs). *d'* comparisons are abbreviated as follows: WD-PD, OD-PD, OD-WD. Other variables follow the same abbreviations as in (a). Misses show no effect size, as values were identical across conditions.

indicating that participants who miss more (presumably tapping less) are also less likely to produce false alarms (which require more frequent responses).

Meanwhile, the correlation between false alarms and hit rate became significant (row 9, col (WD), r = 0.66, p < 0.001), reflecting the increased variability in false alarms when hit rates are high, a pattern previously hidden in the OD, where hits were classified as false alarms due to the fixed windows. This result is linked to the correlation between hits and false alarms also becoming significant (row 2, col (WD), r = 0.49, p = 0.004), as hits are now necessary for false alarms to occur. Thus an increase in false alarms imply higher hit rate, as hits and hit rates are highly correlated (row 4, col (WD), r = 0.85, p < 0.001).

While the near-perfect correlation between hits and correct rejections persisted (row 3, col (OD), r = 0.98, p < 0.001; col (WD), r = 0.98, p < 0.001), the relationship between correct rejections and false alarm rates weakened (row 12, col (OD), r = -0.56, p < 0.001; col (WD), r = -0.14, p = 0.427). This may result from an interaction between correct rejections relationship with hits and hits' relationship with false alarms. A similar pattern is observed in row 5 (col (OD) r = -0.48, p < 0.005; col (WD), r = 0.01, p = 0.952), where hits and false alarm rates do not exhibit a clear relationship.

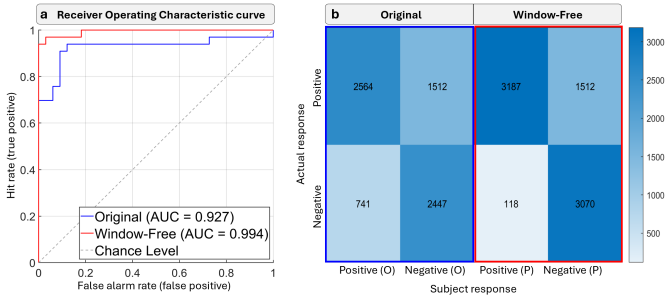

Fig. 8: **(a)** Receiver Operating Characteristic (ROC) curves for the original design (blue) and window-free design (red), showing high area under the curve (AUC) values for both conditions. **(b)** Confusion matrix displaying the four outcomes of the tapping task using both the original and window-free design. In each 2×2 quadrant we can find: hit (top left), miss (top right), false alarm (bottom left), and correct rejection (bottom right).

		Statistic (OD)	*p* (OD)	Statistic (WD)	*p* (WD)
Hit	Miss	-0.66	< 0.001	-0.75	< 0.001
Hit	False alarm	0.19	0.294	0.49	**0.004
Hit	Correct rejection	0.98	< 0.001	0.98	< 0.001
Hit	Hit rate	0.92	< 0.001	0.85	< 0.001
Hit	False alarm rate	-0.48	**0.005	0.01	0.952
Miss	False alarm	-0.6	< 0.001	-0.69	< 0.001
Miss	Correct rejection	-0.6	< 0.001	-0.65	< 0.001
Miss	Hit rate	-0.87	< 0.001	-0.98	< 0.001
False alarm	Hit rate	0.31	0.079	0.66	< 0.001
False alarm	False alarm rate	0.71	< 0.001	0.82	< 0.001
Correct rejection	Hit rate	0.89	< 0.001	0.78	< 0.001
Correct rejection	False alarm rate	-0.56	< 0.001	-0.14	0.427

TABLE II: Significant Spearman Rank correlations - SDT outcomes ** = p < 0.01; *** = p < 0.001. (OD): original design, time window design; (WD) window-free design. Abbreviations: correct rejections (Cr); false alarms (FA); hit rate (HitR); false alarm rate (FAR)

As correct rejections increase with hits while false alarms also increase, the calculation of false alarm rates (eq. 1) remains stable, since both factors shift in parallel. Consequently, the correlation between hits and false alarm rates is likely driven by variability rather than a direct association. The low number of false alarms (Figure 8b) in the new design may contribute to these spurious correlations, particularly in cases where false alarms are absent (i.e., when FA = 0).

Table 3, on the other hand, shows how all three versions of *d'* correlate to each SDT outcome.

	r (OD)	*p* (OD)	*r* (WD)	p (WD)	*r* (PD)	*p* (PD)
Hit	0.95	< 0.001	0.69	< 0.001	0.80	< 0.001
Miss	-0.64	< 0.001	-0.61	< 0.001	-0.86	< 0.001
False alarm	0.01	0.971	-0.03	0.885	0.39	*0.026
Correct rejection	0.95	< 0.001	0.75	< 0.001	0.79	< 0.001
Hit rate	0.91	< 0.001	0.68	< 0.001	0.9	< 0.001
False alarm rate	-0.6	< 0.001	-0.44	0.27	0	0.988

TABLE III: Spearman Rank Correlation table - *d'*-SDT outcomes * = p < 0.05; ** = p < 0.01; p < 0.001. (OD) original design *d'*; (WD) window-free design *d'*; (PD) window-free design with pooled z-score *d'*.

Due to a stronger relationship between misses and hits (Table 2, row 1, col (WD), r = -0.75, p < 0.001), making misses more relevant for hit rates (Table 2, row 8, col (OD), r = -0.87, p < 0.001; col (WD), r = -0.98, p < 0.001), misses now have an enhanced negative impact on *d'* (Table 3, row 2, col (OD), r = -0.61, p < 0,001; col (PD), r = -0.86, p < 0.001) resulting in a reduced influence of hits (Table 3, row 1, col (OD), r= 0.95, p < 0.001; col (PD) r = 0.80, p < 0.001). As a result, misses and hits reflect *d'* within the PD in a more balanced way compared to the OD, as pooling z-scores redistributes their influence. Because correct rejections are symmetrically linked to hits, this pattern of reduced influence repeats itself with the relationship of correct rejections and *d'* (Table 3, row 4, (WD) r = 0.95, p < 0.001; (PD) r = 0.79 , p < 0.001).

False alarm rate and hit rate distributions are tied together within the PD, having both a strong correlation with false alarms and hits respectively. This creates a strong correlation between false alarms and hits (Table 2, row 2, col (WD), r = 0.49, p = 0.004) which in turn reflects a positive correlation between *d'* and false alarms (Table 3, row 3, col (WD), r = -0.03, p = 0.885; col (PD), r = 0.39, p = 0.026) when false alarm occurrences are low.

Although the hit rate relationship to *d'* decreases with the design change (Table 3, row 5, col (WD), r = 0.68, p < 0.001), it becomes highly correlated again with *d'* when pooling the rates (Table 3, row 5, col (OD), r = 0.91, p < 0.001; col (PD), r = 0.90, p < 0.001). This decrease in correlation when changing the design, despite using the same *d'* formula, is likely due to the INCDF transformation failing to adequately normalize the rate distributions under the new design. In contrast, the PD combines both hit and false alarm rates under the same framework, revealing a more monotonic relationship between them. This approach results in higher *d'* values associated with the hit rates and lower *d'* values associated with the false alarm rate, although the latter is not clearly observed due to the low frequency of false alarms (Table 3, row 6, col (OD), r = -0.6, p < 0.001; col (PD), r = 0, p = 0.988).

4) Cross-interoceptive correlations: Separate Spearman Rank correlations were conducted to evaluate the relationship between the three *d'* outputs and *MD* (Figure 10).

	r (OD)	*p* (OD)	*r* (WD)	*r* (WD)	*r* (PD)	*p* (PD)
Mean Distance	-0.59	< 0.001	-0.78	< 0.001	-0.93	< 0.001

TABLE IV: Spearman Rank Correlation table - *d'-mean distance* (OD) original design *d'*; (WD) window-free design *d'*; (PD) window-free design with pooled z-score *d'*.

Every design showed strong correlations with *MD* (Table 4); however, we observed an increase in correlation in the order of OD > WD > PD (OD: r = -0.59, p < 0.001; WD: r = -0.78, p < 0.001; PD: r = -0.93, p < 0.001), where the more negative the correlation, the stronger their relationship was. As expected, the new design amplified the relationship between *MD* and *d'*, which can be primarily attributed to the influence of misses on *MD*. Misses are critical to *MD*, as they capture the largest shifts in response timing. As misses become more influential in *d'* with the new design (Table 2), the correlation between *d'* and *MD* strengthens. The PD stabilizes

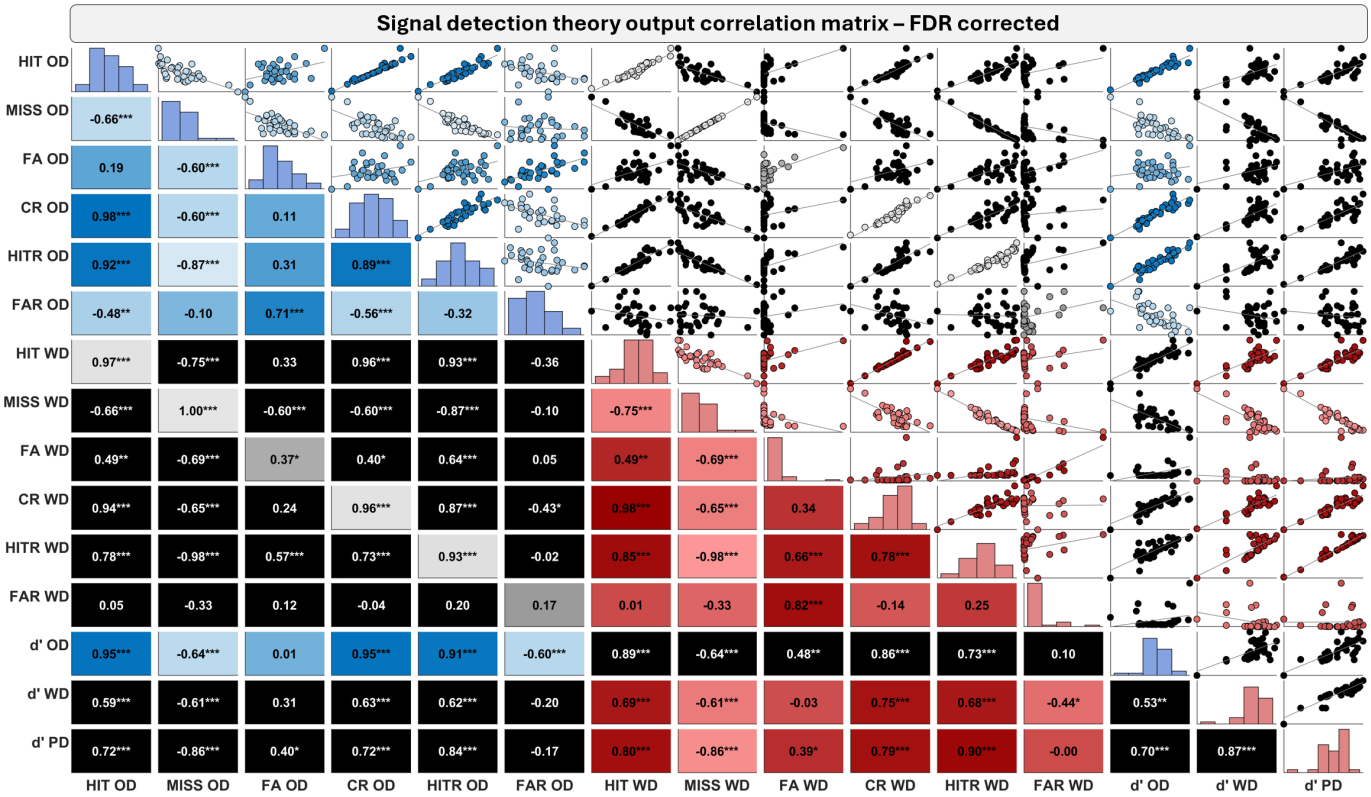

Fig. 9: * = p < 0.05; ** = p < 0.01; *** = p < 0.001. Labels indicate either (OD) = Original design, (WD) = window-free design, or (PD) window-free design with pooled z-score. Blue represents correlations between two OD variables, red represents correlations between two WD variables, grey represents correlations between corresponding SDT outcomes from OD and WD, and black represents correlations between non-corresponding OD and WD variables. Variables are in order: hit; miss; false alarm (FA); correct rejection (CR); hit rate (HitR); false alarm rate (FAR); d' original (d' OD); d' window-free design (d' WD); d' window-free design with pooled z-score (d' PD). All correlations were corrected for multiple comparisons using the false discovery rate (FDR).

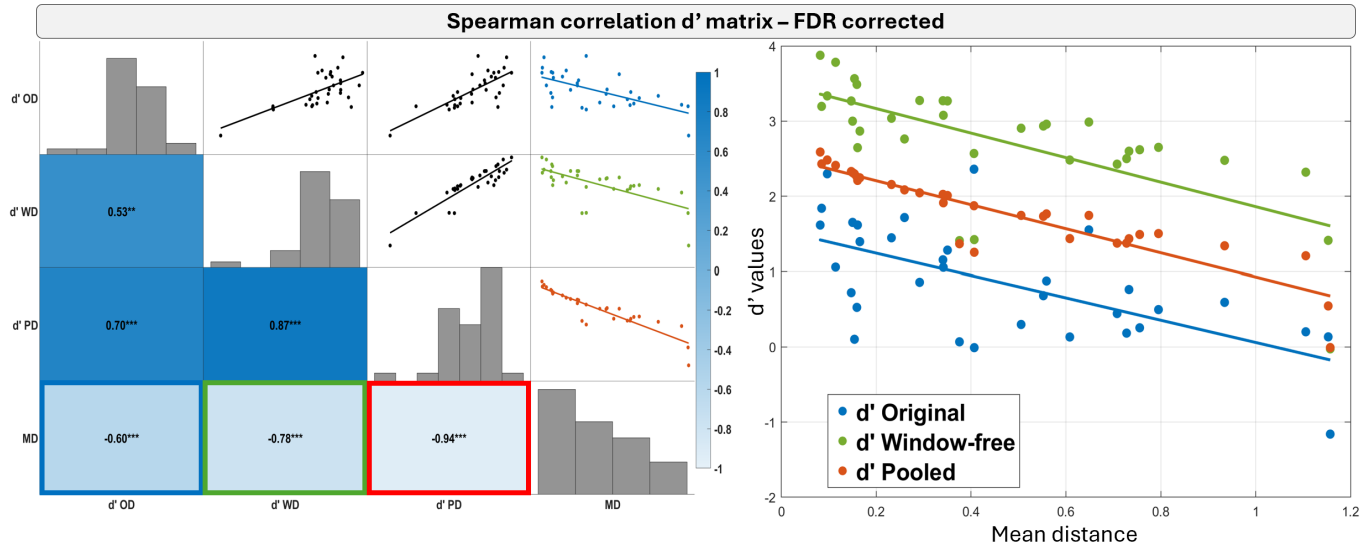

Fig. 10: **Spearman rank correlation matrix and scatter plot with regression and residuals - *d'-mean distance***. *d'* Original design (blue); *d'* Window free design (green); *d'* Window free design with pooled z-score (red).

this relationship, avoiding the additional variability introduced by the INCDF-transformed rates, which further dampened the original correlation.

Finally, we assessed the residual dispersion of each corre-

lation using the Fligner–Killeen test [18], a non-parametric method for evaluating the homogeneity of variances across multiple groups. Residuals were compared pairwise across models (Original vs. Design, Original vs. Pooled, and Design vs. Pooled). All three comparisons revealed significantly different dispersions (all $p < 0.001$), with the dispersion magnitude following the order: Original < Design < Pooled. These results are consistent with the observed differences in alignment to the regression trend (Figure 10).

VI. Discussion

Our findings highlight critical challenges in using *d'* as a measure of interoceptive accuracy in heartbeat detection tasks. While *d'* is theoretically well-suited for assessing sensitivity, its application becomes problematic when task designs lack a clear distinction between signal and noise trials. Specifically, we showed that responses in interoceptive tapping tasks do not necessarily align with R-peaks, challenging the traditional assumption that these responses follow a predictable temporal alignment. Furthermore, we identified inherent biases in the established heartbeat detection task *d'* version [34] as it relates to heart rate and heart rate variability, which could affect the accuracy of interoceptive assessments.

To address these issues, we proposed an alternative window-free design, which removes the reliance on R-peak alignment. This design offers a more flexible and ecologically valid approach to measuring interoception, better reflecting real-world scenarios by assuming response variability from both motor lag and early tapping due to heartbeat prediction [4], [37]. Additionally, we introduced a pooled z-score method to overcome the limitations inherent when using INCDF in *d'* z-score calculations for tapping tasks [34]. This method enhances the reliability of the measure by accounting for variations in the response distributions. The aggregated distribution avoids adding bias from differences in individual false alarm and hit rate distributions when calculating the z-values later used in *d'*.

Our proposed design revealed a stronger influence of misses and false alarms on *d'*, which we argue are a reflection of our pooled z-score design. This approach distinguishes poor perceivers—individuals with low interoceptive accuracy and inconsistent responses—from good perceivers, who are assumed to respond more consistently [24], [41]. In our healthy control dataset, false alarms were rare, with many instances in which subjects had no false alarms at all. This pattern, however, is also a promising aspect of our new design, as it suggests that healthy people are following a normal heartbeat response rhythm with few instances of more than one response per inter-beat interval. When evaluating interoception in clinical groups against a control group, we can expect a high control *d'* as our baseline.

Furthermore, the strong negative relationship between false alarms and misses penalizes *d'* equally, making the interpretation of interoceptive accuracy valid for both under-responders and over-responders. Because of this consideration, the new approach could differentiate between individuals with higher interoceptive accuracy and those with more frequent false alarms or misses.

Notably, when combined with our revised task design, the pooled z-values design exhibited a near-perfect correlation with mean distance, a widely used measure of interoceptive accuracy based on tapping frequency. Although it may be tempting to conclude that this finding suggests that both metrics are measuring the same thing, it underscores the importance of understanding how these two measures capture distinct spatiotemporal aspects of interoception. For instance, two responses occurring within an R-R interval, positioned near the first and second R peaks, would result in a near-zero MD value, indicating a perfect match. However, d' would classify the second response as a false alarm, penalizing the outcome. This distinction highlights that d' captures spatial precision, which mean distance overlooks, while mean distance retains temporal information that d' obscures. The strong correlation between these metrics suggests they may reflect a common underlying sensitivity process, offering new insights into the structure of interoceptive ability. We attribute this improved relationship to the enhanced consideration of misses and false alarms, which are more effectively captured by our new design. Additionally, the relationship appears to be influenced by z-score pooling, as evidenced by a reduction in inter-subject variability when pooled means and standard deviations are used to compute z-values. This increase in homogeneity holds promise for clinical group analysis, as deviations from this tight correlation could offer valuable insights into interoceptive deficits.

Our proposed methodology enhances the reliability and comparability of interoceptive accuracy measures, facilitating more accurate cross-modal investigations. By better aligning temporal and spatial metrics, this approach lays the groundwork for more refined assessments of interoceptive accuracy. Future studies should explore its applicability across different interoceptive paradigms and populations, further validating its robustness in real-world settings. Indeed, obtaining accurate measures of interoception is essential to enhance our understanding of interoception in health and disease. We hypothesize that, in clinical groups with impaired interoception, there would be a more pronounced decrease in d', accompanied by an increase in instances of false alarms and misses. We also expect our d' approach to have a stronger correlation with the Heart Evoked Potential and other interoceptive measurements than the original d' approach. Although, this remains to be tested. Furthermore, we believe that new metrics combining mean distance and *d'* could offer valuable insights into interoceptive processes. For example, one could measure mean distance only in hit trials within *d'*, creating a "hit distance" approach. This method might help isolate the influence of good interoceptive processes, potentially revealing important distinctions between individuals with strong interoceptive accuracy and those whose performance is obscured by false alarms.

In conclusion, these findings suggest that our pooled z-score design introduces a more structured relationship between Signal Detection Theory outcomes with stronger correlation to other interoceptive based measures. By avoiding assumptions of temporal proximity to R-peaks and comparable rate distributions, this restructuring enhances the interpretability of signal detection outcomes, enhancing our understanding

on interoceptive accuracy within heartbeat detection tasks. Moreover, the independence of our metric from both heart rate and heart rate variability presents a promising advance for future investigations of interoception in clinical populations.

REFERENCES

[1] S. Abrevaya, S. Fittipaldi, A. M. García, M. Dottori, H. Santamaria-Garcia, A. Birba, A. Yoris, M. K. Hildebrandt, P. Salamone, A. De la Fuente, et al. At the heart of neurological dimensionality: Cross-nosological and multimodal cardiac interoceptive deficits. *Biopsychosocial Science and Medicine*, 82(9):850–861, 2020.

[2] M. W. Agelink, C. Boz, H. Ullrich, and J. Andrich. Relationship between major depression and heart rate variability.: Clinical consequences and implications for antidepressive treatment. *Psychiatry research*, 113(1-2):139–149, 2002.

[3] M. Ardizzi, M. Ambrosecchia, L. Buratta, F. Ferri, M. Peciccia, S. Donnari, C. Mazzeschi, and V. Gallese. Interoception and positive symptoms in schizophrenia. *Frontiers in human neuroscience*, 10:379, 2016.

[4] L. F. Barrett and W. K. Simmons. Interoceptive predictions in the brain. *Nature reviews neuroscience*, 16(7):419–429, 2015.

[5] G. G. Berntson and S. S. Khalsa. Neural circuits of interoception. *Trends in neurosciences*, 44(1):17–28, 2021.

[6] J. Brener and C. Ring. Towards a psychophysics of interoceptive processes: the measurement of heartbeat detection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160015, 2016.

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

[8] E. Ceunen, J. W. Vlaeyen, and I. Van Diest. On the origin of interoception. *Frontiers in psychology*, 7:743, 2016.

[9] W. G. Chen, D. Schloesser, A. M. Arensdorf, J. M. Simmons, C. Cui, R. Valentino, J. W. Gnadt, L. Nielsen, C. S. Hillaire-Clarke, V. Spruance, et al. The emerging science of interoception: sensing, integrating, interpreting, and regulating signals within the self. *Trends in neurosciences*, 44(1):3–16, 2021.

[10] Y.-C. Cheng, Y.-C. Huang, and W.-L. Huang. Heart rate variability in individuals with autism spectrum disorders: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, 118:463–471, 2020.

[11] M.-P. Coll, H. Hobson, G. Bird, and J. Murphy. Systematic review and meta-analysis of the relationship between the heartbeat-evoked potential and interoception. *Neuroscience & Biobehavioral Reviews*, 122:190–200, 2021.

[12] B. Couto, F. Adolfi, M. Velasquez, M. Mesow, J. Feinstein, A. Canales-Johnson, E. Mikulan, D. Martínez-Pernía, T. Bekinschtein, M. Sigman, et al. Heart evoked potential triggers brain responses to natural affective scenes: a preliminary study. *Autonomic Neuroscience*, 193:132–137, 2015.

[13] A. Dale and D. Anderson. 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, 1978.

[14] A. de la Fuente, L. Sedeño, S. S. Vignaga, C. Ellmann, S. Sonzogni, L. Belluscio, I. García-Cordero, E. Castagnaro, M. Boano, M. Cetkovich, et al. Multimodal neurocognitive markers of interoceptive tuning in smoked cocaine. *Neuropsychopharmacology*, 44(8):1425–1434, 2019.

[15] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[16] O. Desmedt, O. Luminet, and O. Corneille. The heartbeat counting task largely involves non-interoceptive processes: Evidence from both the original and an adapted counting task. *Biological psychology*, 138:185–188, 2018.

[17] O. Desmedt, O. Luminet, M. Walentynowicz, and O. Corneille. The new measures of interoceptive accuracy: a systematic review and assessment. *Neuroscience & Biobehavioral Reviews*, page 105388, 2023.

[18] M. A. Fligner and T. J. Killeen. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, 71(353):210–213, 1976.

[19] S. N. Garfinkel, A. K. Seth, A. B. Barrett, K. Suzuki, and H. D. Critchley. Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological psychology*, 104:65–74, 2015.

[20] D. M. Green, J. A. Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[21] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl. Heart rate variability as indicator of clinical state in depression. *Frontiers in psychiatry*, 9:735, 2019.

[22] J. L. Hazelton, G. Della Bella, P. Barttfeld, M. Dottori, R. Gonzalez-Gomez, J. Migeot, S. Moguilner, A. Legaz, H. Hernandez, P. Prado, et al. Altered spatiotemporal brain dynamics of interoception in behavioural-variant frontotemporal dementia. *EBioMedicine*, 113, 2025.

[23] J. L. Hazelton, S. Fittipaldi, M. Fraile-Vazquez, M. Sourty, A. Legaz, A. L. Hudson, I. G. Cordero, P. C. Salamone, A. Yoris, A. Ibañez, et al. Thinking versus feeling: How interoception and cognition influence emotion recognition in behavioural-variant frontotemporal dementia, alzheimer's disease, and parkinson's disease. *Cortex*, 163:66–79, 2023.

[24] B. M. Herbert and O. Pollatos. The body in the mind: on the relationship between interoception and embodiment. *Topics in cognitive science*, 4(4):692–704, 2012.

[25] Z. H. Hoo, J. Candlish, and D. Teare. What is an roc curve?, 2017.

[26] Y. Imahori, D. L. Vetrano, X. Xia, G. Grande, P. Ljungman, L. Fratiglioni, and C. Qiu. Association of resting heart rate with cognitive decline and dementia in older adults: A population-based cohort study. *Alzheimer's & Dementia*, 18(10):1779–1787, 2022.

[27] M. Kleiner, D. Brainard, and D. Pelli. What's new in psychtoolbox-3? 2007.

[28] D. N. Levine. Sherrington's "the integrative action of the nervous system": A centennial appraisal. *Journal of the neurological sciences*, 253(1-2):1–6, 2007.

[29] C. D. B. Luft and J. Bhattacharya. Aroused with heart: Modulation of heartbeat evoked potential by arousal induction and its oscillatory correlates. *Scientific reports*, 5(1):15717, 2015.

[30] A. C. Marshall, A. Gentsch, and S. Schütz-Bosbach. The interaction between interoceptive and action states within a framework of predictive coding. *Frontiers in Psychology*, 9:180, 2018.

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

[32] P. Perakakis. Heplab: a matlab graphical interface for the preprocessing of the heartbeat-evoked potential. *Zenodo*, 2019.

[33] D. Plans, S. Ponzo, D. Morelli, M. Cairo, C. Ring, C. T. Keating, A. Cunningham, C. Catmur, J. Murphy, and G. Bird. Measuring interoception: The phase adjustment task. *Biological Psychology*, 165:108171, 2021.

[34] P. C. Salamone, A. Legaz, L. Sedeño, S. Moguilner, M. Fraile-Vazquez, C. G. Campo, S. Fittipaldi, A. Yoris, M. Miranda, A. Birba, et al. Interoception primes emotional processing: multimodal evidence from neurodegeneration. *Journal of Neuroscience*, 41(19):4276–4292, 2021.

[35] R. Salomon, R. Ronchi, J. Dönz, J. Bello-Ruiz, B. Herbelin, R. Martet, N. Faivre, K. Schaller, and O. Blanke. The insula mediates access to awareness of visual stimuli presented synchronously to the heartbeat. *Journal of Neuroscience*, 36(18):5115–5127, 2016.

[36] K. B. Schauder, L. E. Mash, L. K. Bryant, and C. J. Cascio. Interoceptive ability and body awareness in autism spectrum disorder. *Journal of experimental child psychology*, 131:193–200, 2015.

[37] A. K. Seth, K. Suzuki, and H. D. Critchley. An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2:395, 2012.

[38] S. S. Shergill, P. M. Bays, C. D. Frith, and D. M. Wolpert. Two eyes for an eye: the neuroscience of force escalation. *Science*, 301(5630):187–187, 2003.

[39] H. Stanislaw and N. Todorov. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149, 1999.

[40] J. A. Swets. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press, 2014.

[41] Y. Terasawa, Y. Moriguchi, S. Tochizawa, and S. Umeda. Interoceptive sensitivity predicts sensitivity to the emotions of others. *Cognition and Emotion*, 28(8):1435–1448, 2014.

[42] R. Thapa, G. A. Alvares, T. A. Zaidi, E. E. Thomas, I. B. Hickie, S. H. Park, and A. J. Guastella. Reduced heart rate variability in adults with autism spectrum disorder. *Autism Research*, 12(6):922–930, 2019.

[43] Z. J. Williams, E. Suzman, S. L. Bordman, J. E. Markfeld, S. M. Kaiser, K. A. Dunham, A. R. Zoltowski, M. D. Failla, C. J. Cascio, and T. G. Woynaroski. Characterizing interoceptive differences in autism: a systematic review and meta-analysis of case–control studies. *Journal of autism and developmental disorders*, 53(3):947–962, 2023.

[44] G. Zamariola, P. Maurage, O. Luminet, and O. Corneille. Interoceptive accuracy scores from the heartbeat counting task are problematic: Evidence from simple bivariate correlations. *Biological psychology*, 137:12–17, 2018.