A psychometric history of cardiac interoception

Micah G. Allen

Aarhus University

Preprint

Historical review
6 November 2026

Foreword

This review began with a question I had returned to for years: where did the delays in heartbeat discrimination tasks come from? Reading primary texts and following references had revealed a much richer debate than contemporary descriptions of the tasks usually convey. I wanted to see whether AI tools could help organize that accumulated material and make its history readable. The result is an experiment in AI-assisted scholarship, guided by my research experience and continuing engagement with the sources. Its value depends on the historical argument and the evidence supporting it; fluent prose is no substitute for either. A brief disclosure at the end describes how the tools were used.

Contents

1. Introduction 3
2. From bodily awareness to voluntary control 3
3. Making a heartbeat immediate or delayed 4
4. Letting the observer locate the heartbeat 6
   4.1. Timing as an individual difference 7
5. What counting asked of the observer 9
6. New delays, new purposes 10
   6.1. A different reference point 11
7. How the delays acquired a physiology 13
8. What the history leaves open 13

Disclosures 15

Appendix: scoring and sources 16

References 17

1. Introduction

A person may notice a forceful beat, estimate a rapid pulse, count bodily sensations or judge whether a tone coincides with them. Each performance bears on questions about emotion, bodily awareness and symptoms; each also asks something different of the observer. To describe someone as more aware of the heart does not tell us which of these performances would disclose that awareness. Much of the difficulty in measuring cardiac interoception arises when they are taken as interchangeable evidence of a single ability. The tasks were devised for particular questions, and their later applications have often outlived the debates that gave those questions experimental form.

Following those debates from autonomic awareness and biofeedback through heartbeat discrimination, individual timing estimates, emotion and metacognition, this review asks where the delays between a recorded heartbeat and an external signal came from, and how they acquired their present meanings. A procedure could persist while its purpose, physiological reference and criterion of success changed. Recovering those changes helps explain what the tasks actually measure and why choosing among them requires more than choosing a familiar name. The account follows this methodological lineage rather than attempting an exhaustive history of interoception. Its central concern is how a provisional answer to a timing problem became an inherited standard, and what the experiments that questioned it still have to teach us.

2. From bodily awareness to voluntary control

In 1958, George Mandler, Jean Mandler and Ellen Uviller introduced the Autonomic Perception Questionnaire to examine awareness of bodily activity during emotional states. Participants reported how often they noticed a beating heart, perspiration, altered breathing and other bodily changes; the investigators then asked whether these reports corresponded to autonomic events recorded in the laboratory. People who reported greater awareness showed greater physiological reactivity in some comparisons, but also tended to overestimate it. A report of awareness might therefore express the strength of the bodily response, the accuracy of its perception, or a disposition to describe oneself in bodily terms (Mandler et al. 1958, pp. 367-372).

These alternatives were already sufficient reason to seek behavioural evidence. Mandler and Kahn's 1960 paper described two attempts to train discrimination of changes in heart rate. In the first, a participant appeared to learn when his heart was accelerating or decelerating. Examination of his breathing changed the interpretation. He had adopted a regular respiratory manoeuvre, and the associated changes in heart rate allowed him to answer correctly. The response was controlled by breathing rather than by direct perception of the cardiac change. Their second attempt, using a disguised guessing task, failed to produce convincing discrimination (Mandler & Kahn 1960, pp. 21-25).

What had first appeared to be success was consequently more instructive than a straightforward failure might have been. Performance could be accurate even when the participant used information other than the sensation the task was intended to measure. The participant need not be mistaken or deceptive. The experiment might simply reward a strategy its designers had not intended to test.

Biofeedback made the distinction between reported awareness and perceptual ability consequential. As Whitehead and colleagues presented Brener's theory, feedback helped a learner recognize the sensory consequences of bodily changes, allowing the desired state to be reproduced voluntarily. They contrasted this with an operant account in which reinforcement could establish control without sensory awareness being necessary. A measure of perception was needed to test whether people who discriminated cardiac events also learned control more readily (Whitehead et al. 1977, pp. 371-372). Brener's original theoretical chapters remain unavailable here, so this account of the dispute follows Whitehead's formulation.

A report of frequently noticing the heart could not establish the ability required by that prediction. Investigators therefore sought a behavioural measure with which to examine the proposed relationship between awareness and control.

Roger Kleinman's 1970 dissertation, supervised by Brener, used button presses made in response to felt heartbeats. It examined the timing of presses relative to the electrocardiogram and treated detection as a skill that could be trained (Kleinman 1970, pp. 28-31). Brener and Jones's 1974 experiment instead asked participants to distinguish two sources of vibrations delivered to the wrist. On some trials the vibrations were triggered by the participant's heart; on others they came from a generator operating at approximately the same rate. Corrective feedback enabled the experimental group to improve. A control condition using recorded sequences addressed the possibility that participants merely distinguished an irregular heartbeat rhythm from a regular mechanical one (Brener & Jones 1974, pp. 764-765).

While listening to the vibrations, a participant could alter breathing and observe whether their rhythm changed with it. A change would suggest that the vibrations followed the live heart, and participants reported using such respiratory strategies (Brener & Jones 1974, pp. 765-766). Exploration of this kind could be informative in acquiring bodily control. In an experiment intended to isolate perception of the heartbeat, however, it left another route to the correct answer. The experimental interests were beginning to diverge.

Other laboratories approached the problem by asking participants to reproduce or count the cardiac rhythm. McFarland's 1975 procedure required button presses in time with the perceived heartbeat, but its score compared the total number of presses with the total number of beats. The temporal correspondence between an individual press and a heartbeat did not enter the formula. Dale and Anderson's 1978 procedure asked participants to count heartbeats during brief intervals without palpating the pulse (McFarland 1975, p. 403; Dale & Anderson 1978, p. 81).

A person pressing at the correct average rate could obtain a perfect count-based score even if no individual press bore a consistent temporal relation to a heartbeat. Silent counting according to an estimate of heart rate admits the same possibility. Counting and discrimination would become rival procedures partly because they responded differently to it: counting offered ease of administration, while discrimination sought evidence tied more closely to the relation between a particular heartbeat and an external event.

3. Making a heartbeat immediate or delayed

At the Society for Psychophysiological Research meeting in Toronto in October 1975, Whitehead, Drescher and Blackwell presented a new test of heartbeat perception. Its abstract, published the following March, described a light flashing either 128 or 384 milliseconds after each R-wave. There were 200 trials of ten seconds each. After each trial, the participant judged whether the feedback had been immediate or delayed. No trial-by-trial information about accuracy was given, and sensitivity was assessed using signal detection theory (SPR abstracts 1976, p. 177).

When both kinds of trial followed the live heart, a respiratory manoeuvre that changed heart rate changed both signal trains. The participant could no longer distinguish the conditions merely by discovering which train responded to an action. Whitehead's design thus removed the contrast that had supported the respiratory solution and required a classification of timing relative to cardiac activity. It controlled one demonstrated route to success. Whether the resulting judgments depended on the intended sensation still required validation.

The abstract reported that the new score was essentially unrelated to the Autonomic Perception Questionnaire. In the full 1977 paper, with Peter Heiman added to the author list, discrimination also failed to show the predicted positive relationship with initial heart rate control or its acquisition

under feedback. A negative association in the first of three experiments did not replicate (Whitehead et al. 1977, pp. 371, 380-385). Making both trains cardiac-contingent, collecting many judgments and separating sensitivity from response preference had addressed identifiable weaknesses in earlier work. The physiological basis for the chosen delays was less fully developed.

In the 1975 abstract, 128 milliseconds represented immediate feedback because the pulse wave was said to require approximately 100 to 150 milliseconds to reach the neck. The 1977 paper referred to mechanical contraction at about 100 milliseconds after the R-wave and arrival of the pulse at the neck or wrist by about 150 milliseconds. Neither account cited a source for these estimates. The longer interval, 384 milliseconds, was described as delayed feedback without a corresponding physiological justification (SPR abstracts 1976, p. 177; Whitehead et al. 1977, p. 376). The R-wave provided a reproducible electrical reference; it did not mark the onset of a conscious sensation.

Whitehead and colleagues did not distinguish perception of cardiac contraction from perception of the pulse as pressure or sound. Their test required a discriminable event associated with the heartbeat; it did not identify the receptor or bodily site responsible for the experience (Whitehead et al. 1977, p. 376).

The equipment measured interbeat intervals in units of 16 milliseconds, providing a likely instrumental explanation for the unusual numbers. The paper explicitly attributed its unusual decimal values to binary timing. The two delays correspond to eight and twenty-four of those units. This makes the apparatus relevant to the numerical choices, but it does not establish why the investigators selected those particular multiples. The timing hardware explains the available increments more securely than it explains the experimental decision (Whitehead et al. 1977, p. 375).

Whatever led to those increments, the experimenter assigned the name *immediate* to a particular time. If the participant felt a heartbeat then, the task could work as intended; if the sensation occurred later, the relation between experience and the scoring rule changed. By ensuring that both signals followed the heart, the design had removed one ambiguity while accepting another. It now assumed when the signal should feel simultaneous with the heartbeat.

A participant who often answered *immediate* would produce more hits on immediate trials and more false alarms on delayed trials. Calculated from those two rates, d' was intended to distinguish discrimination from response preference. Signal detection theory thereby improved the analysis, but its interpretation still depended on the scoring definitions and model assumptions, including which condition was called the signal. The analysis could not establish independently that this condition coincided with the participant's sensation.

On each trial, the participant saw one signal train and classified it. Later descriptions sometimes call this a two-alternative forced-choice procedure, although the alternatives were the answers available to the observer: both trains were not presented within each trial for direct comparison. Keeping that difference visible tells the reader more about the experiment than the task label alone.

In Whitehead's first experiment, median d' was .33. For later experiments, the investigators selected participants from the extremes of the distribution: d' at or below .10, or at or above .75. These thresholds served recruitment into contrasting groups; they did not establish a general statistical criterion for who could perceive a heartbeat (Whitehead et al. 1977, pp. 381-382).

Brener, Liu and Ring referred in 1993 to "Whitehead's criterion" of d' greater than .75, and Brener and Ring's 2016 review again treated this value as a detector threshold (Brener et al. 1993, p. 663; Brener & Ring 2016, p. 3). In Whitehead's original experiment, however, the number marked entry into a selected group. The difference changes what can be inferred from the proportion exceeding it: a recruitment cutoff does not establish the prevalence of heartbeat perception. Detector proportions belong to specified procedures, samples and classification rules.

4. Letting the observer locate the heartbeat

The new procedure did not immediately settle the measurement question. Poor performance could mean that heartbeat sensations were weak or absent, that the comparison was difficult, that the participant adopted an ineffective strategy, or that the chosen delays did not suit the sensation. Different laboratories pursued different explanations.

William Clemens's 1979 study examined learning and reliability in the earlier Brener-Jones family of procedures. In addition to a generator, it used a replay of the participant's recent cardiac rhythm to create a comparison train with realistic variability. Performance before training was poorly consistent across days. On the final page, Clemens turned to Whitehead's solution. Making both trains follow the heart prevented the participant from identifying one through voluntary manipulation, but it meant that the supposedly incorrect train remained perfectly related to cardiac activity. Its designation as incorrect depended on an assumption about phase (Clemens 1979, pp. 334-338).

Edward Katkin's laboratory approached low performance as a problem of task difficulty. Katkin, Blascovich and Goldband's 1981 procedure placed the nominally synchronous tone 100 milliseconds after the R-wave. On the other trials, the delay increased by 30 milliseconds with each successive heartbeat, starting from a randomly chosen offset. A participant who felt the heart should therefore hear an increasingly conspicuous separation between the beat and the tone. The authors deliberately designed that drift to make the judgment easier (Katkin et al. 1981, p. 1097).

Training improved performance in the male participants in Katkin's sample, though not in the female participants. Replacing the live electrocardiogram with a recording reduced performance to chance, supporting dependence on information related to the current heart in that experiment (Katkin et al. 1981, pp. 1099-1100).

Davis and colleagues subsequently identified a difficulty within the drift itself. If its starting point was early, the first several supposedly asynchronous signals could fall close to the synchronous delay. Their 1986 comparison held the synchronous signal at 100 milliseconds and contrasted a fixed late signal, the Katkin procedure, and a drift beginning from a larger constant offset. Before feedback, performance was near chance in all three groups. With feedback, the condition beginning from the larger offset improved most clearly (Davis et al. 1986, pp. 77-79).

An *asynchronous* train could begin near the presumed sensation and move away from it, while a sufficiently late signal could approach the next heartbeat. Its relation to the heart depended on the programmed delay, the tone's position within the train and heart rate together. Davis and colleagues expressly noted that, at a sufficiently rapid rate, a late tone might coincide with the next R-wave (Davis et al. 1986, p. 80). "In sync" and "out of sync" therefore left essential features of the temporal arrangement unstated.

Ross and Brener's 1981 comparison showed how training could also teach a way of approaching the task. Participants trained first on Brener-Jones tended to test the vibrations by altering breathing; those trained first on Whitehead tended to remain still and attend to a sensation. Many retained the initial strategy when the procedure changed (Ross & Brener 1981, pp. 64-67). A strategy useful for discovering bodily control could therefore support one discrimination more than another. The two scores organized behaviour differently despite their common description as cardiac awareness.

Jones, O'Leary and Pipkin tested the interchangeability directly in 1984. Their participants completed both procedures without feedback. Scores did not correlate meaningfully, and very few participants met the study's success criterion on both tasks. Without an independent standard of cardiac perception, the authors could not declare one measure correct and the other wrong (Jones et al. 1984, pp. 145-147). The disagreement between tests was itself evidence that a unitary ability could not simply be assumed from their names.

Even a consistently "wrong" answer might contain evidence of discrimination. Montgomery and Jones used the absolute difference between hit and false-alarm rates, allowing participants who reliably preferred the later signal to count as discriminators (Montgomery & Jones 1984, p. 461). The observer might distinguish the conditions while rejecting the experimenter's label. Such reverse scoring changes the statistical interpretation as well as the meaning of an error; the appendix records the distinction.

Clemens tested the timing assumption directly in 1984. Participants could distinguish flashes at either 0 or 100 milliseconds from flashes at 400 milliseconds, without a clear advantage for the 100-millisecond condition. In a further experiment, repeated trains at five delays between 0 and 400 milliseconds were judged more coincident with the heart at the earlier delays. When participants adjusted the signal themselves within the first 240 milliseconds, however, no single delay emerged as most heartlike (Clemens 1984, pp. 188-190).

An early portion of the cardiac cycle appeared suitable for the synchronous signal, yet the data did not identify 128 milliseconds as a privileged point. The adjustment experiment also introduced a different relationship between observer and apparatus. The observer could search for a match, rather than merely accept or reject the experimenter's choice.

Yates and colleagues obtained a different temporal profile in 1985. They presented a single flash on a designated heartbeat at one of six delays from 0 to 500 milliseconds. Participants judged whether it occurred at the same time as a heartbeat. The group gave the most simultaneity responses between 200 and 400 milliseconds, later than the region favoured in Clemens's experiments (Yates et al. 1985, pp. 562-566).

Repeatedly presenting a fixed set of values and examining how judgments vary across them is the method of constant stimuli. In this application, the proportion of simultaneity judgments was plotted against delay. Instead of a single proportion of correct answers, the experiment yielded a temporal response function whose location and shape could be examined.

Clemens and Yates had not made two otherwise equivalent estimates of the same quantity. Some experiments used trains and another used isolated flashes; their participants, instructions and experience also differed. Yates and colleagues further acknowledged that the warning signal's duration varied with the forthcoming delay and might supply an unintended cue (Yates et al. 1985, p. 566). Validation was still needed. Yet the results brought the timing problem into view: delays often judged simultaneous fell near the interval Whitehead's task designated as delayed.

Could low Whitehead scores simply reflect an inability to distinguish short temporal intervals? Kluvitse's 1987 thesis examined that possibility through comparisons involving external tones and lights. Jones and colleagues supplied a larger published test in 1988. Participants discriminated external stimulus pairs separated by 128 or 384 milliseconds with about 85 per cent accuracy, while averaging about 58 per cent on the heartbeat task. The abilities were only weakly related (Jones et al. 1988, pp. 550-553). Ordinary temporal discrimination could not account for most of the difficulty. The uncertain element was the cardiac reference against which the external signal had to be judged.

4.1. Timing as an individual difference

For Brener and Cecile Kluvitse, the variation in timing raised a practical difficulty with any fixed pair of delays. A participant who felt a heartbeat near the middle of the pair might find both signals comparably mismatched, and thus perform at chance. The same score could be obtained by someone who felt no usable sensation. A task that supplied only that pair would not distinguish the two cases (Kluvitse 1987, p. 155; Brener et al. 1993, p. 657).

The procedure published by Brener and Kluvitse in 1988 allowed participants to explore six R-wave-to-tone delays: 0, 100, 200, 300, 400 and 500 milliseconds. Each delay was assigned to a button,

with the assignments changed between trials. Participants could listen at a chosen delay, move to another and return as often as they wished before selecting the one that felt most simultaneous with the heart. There was no feedback declaring an objectively correct answer (Brener & Kluvitse 1988, pp. 556-557).

Among 32 participants, preferences clustered around 200 and 300 milliseconds, but individual modal choices extended from 100 to 400 milliseconds. None preferred 0 or 500 milliseconds most often (Brener & Kluvitse 1988, pp. 558-559). The regularity of the group was accompanied by a reason not to impose its preferred interval on every person.

Ring and Brener gave this variation a more explicit psychometric description in 1992. The median preferred interval located a participant's temporal match, while the interquartile range described the dispersion of repeated choices. A narrow distribution meant repeated selection within a restricted region; a broad one meant less precise or less differentiated responding. The properties could come apart: one participant might repeatedly select a late match while another selected early matches inconsistently (Ring & Brener 1992, pp. 538-542).

Because the tone itself takes time to be processed, its physical onset after the R-wave is not the same as the time at which its auditory experience coincides with a heartbeat sensation. Ring and Brener therefore made an assumption about auditory latency. Among sixteen participants classified as good perceivers, median tone delays ranged from 81 to 313 milliseconds; adding the assumed 50-millisecond latency yielded their estimated sensation range of 131 to 363 milliseconds (Ring & Brener 1992, p. 544). The latter values were inferred from the judgments and the latency assumption, rather than recorded as onsets of consciousness.

Participants also located sensations in different parts of the body. The small groups reporting the hand or finger had later median intervals than those reporting the head or neck. Ring and Brener considered pulse transmission as one explanation, while recognizing that it did not fully account for the pattern (Ring & Brener 1992, pp. 542-544). Bodily location had become a question for measurement rather than an anatomical premise of the task.

Where earlier procedures had often treated variation in timing as a cause of error, adjustment made it part of the phenomenon. The delay was becoming an outcome of the experiment. That changed the problem from selecting a universally correct signal to establishing what an individual's pattern of judgments represented.

Brener, Liu and Ring's 1993 method of constant stimuli retained the six delays but reduced the demands of active exploration. Each trial presented a train of ten tones at one delay, after which the participant judged whether the tones were simultaneous with the heartbeat. Repetition across all six delays produced a distribution of responses from which location and precision could be estimated.

In 24 participants, thirteen met the detector criterion on constant stimuli, twelve on adjustment and eight on a Whitehead variant. Dispersion estimates from the two multi-interval methods correlated .72; within constant stimuli, the split-half correlation was .86 for temporal location and .60 for dispersion (Brener et al. 1993, pp. 659-664, Table 4). These were different kinds of agreement, between procedures and within one session. They supported examining more than a pair of delays without establishing the long-term stability of every resulting estimate.

Although the responses tended to peak around 200 milliseconds, the authors argued against turning that average into a synchronous interval prescribed for everyone. Individual variation was part of the reason for collecting responses across delays. Substituting a better group average for the original delay would leave the disputed assumption in place (Brener et al. 1993, p. 664).

In the two-delay variant used for their comparison with Whitehead, whichever interval a participant more often judged simultaneous was designated the signal. Some participants preferred the later interval, and these participants tended to have later estimates on the multi-interval tasks. The sensitivity score

consequently measured separation between the two conditions without insisting that the earlier one receive the synchronous label (Brener et al. 1993, pp. 661, 663).

Once the correct delay was no longer specified independently, what would distinguish a perceptual response function from an arbitrary preference? The initial detector criteria tested whether responses differed across intervals. A statistically non-uniform pattern provides evidence of differentiated responding, but its origin still has to be established.

Schneider, Ring and Katkin addressed the origin in 1998 by replacing live cardiac timing with a recording. In the second session, six of twenty participants receiving live signals met the detector criterion, compared with one of eighteen receiving recorded signals. The result supported dependence on the relationship between the tones and current cardiac activity (Schneider et al. 1998, p. 88). As with the earlier replay controls, the evidence concerned the information required to perform the task; it did not isolate the anatomical source of that information.

Even a nearly uniform distribution over the available delays has a median near the centre of the range. An experiment can therefore yield a plausible temporal estimate without evidence of discrimination. Schneider and colleagues tested this problem in a separate experiment by moving all delays 100 milliseconds later, to a range of 100 to 600 milliseconds. Non-detectors had a median of about 344 milliseconds, near the new centre, while detectors had an earlier median of about 284 milliseconds (Schneider et al. 1998, pp. 88-89).

Because the shifted range was tested in a separate sample, this was not evidence of stable individual timing across ranges. It showed why a plausible median was insufficient: the distribution needed to differ from a nearly uniform one, and its relation to the live heart needed independent support.

When Brener, Ring and Liu compared one, five and ten tones per trial in 1994, detector proportions were 13, 43 and 47 per cent, respectively. In those conditions, five tones offered much of the benefit of ten, whereas a single tone often supplied too little information (Brener et al. 1994, pp. 309-311). The finding concerned a particular constant-stimuli procedure. It did not establish a universal minimum train length for every heartbeat task.

5. What counting asked of the observer

While these experiments tried to separate preferred timing from arbitrary responding, counting offered a parallel answer to a different practical demand. It required no carefully timed comparison stimulus and could be incorporated readily into studies of emotion and clinical symptoms. The procedure associated with Schandry's 1981 paper became particularly influential. In the version described by Knoll and Hodapp, participants counted during intervals of 25, 35 and 45 seconds, and the score expressed the discrepancy between reported and recorded beats relative to the recorded total (Knoll & Hodapp 1992, pp. 219-220). Schandry's original paper is considered here through these later descriptions.

A count close to the recorded total could be reached through perception, an accurate estimate of rate, or both. The score did not reveal which process had produced it. Unlike a binary discrimination task, counting also supplied no single guessing level against which to judge performance: an error had to be interpreted in relation to the participant's likely knowledge and strategy (Knoll & Hodapp 1992, pp. 220, 222).

For the clinical question pursued in their 1992 study of panic disorder, Ehlers and Breuer questioned the use of discrimination tasks on the grounds that untrained participants often performed near chance and that training changed the ability under examination. For investigators interested in spontaneous bodily awareness, a lengthy programme that taught participants to solve a temporal comparison could be poorly matched to the clinical question. Their choice of counting followed from that concern, and

their patients with panic disorder showed more accurate counting than controls (Ehlers & Breuer 1992, pp. 374-377).

Ehlers and Breuer themselves considered habitual attention and practice as alternatives to superior sensory capacity. They regarded these as clinically relevant possibilities, while distinguishing noticing a sensation from interpreting it as threatening. Their counting result did not establish that interpretation as a mechanism of panic or settle whether heightened awareness preceded the disorder. They called for replication with other tasks and questioned generalization from quiet laboratory rest to everyday life (Ehlers & Breuer 1992, pp. 380-381).

What an adequate test should preserve depended on the investigator's question. Discrimination experiments tried to exclude estimates and indirect routes to a correct answer; clinical investigators also needed untrained participants to express differences in their ordinary bodily attention. That attention might compete with sensory sensitivity as an explanation of a score, or itself be the object of a study of symptom monitoring. Increasing experimental control could not decide which question should be asked.

Knoll and Hodapp compared counting with a discrimination procedure in 1992. Their discrimination task used Störmer's proportional timing: signals occurred at one quarter or three quarters of the preceding interbeat interval (Knoll & Hodapp 1992, p. 219). This differed from a fixed delay in milliseconds. The temporal separation widened when the preceding interval was longer and narrowed when it was shorter. Because the next interval had not yet occurred, proportional scheduling predicted its duration from a previous beat; it did not place the signal at a known fraction of the forthcoming cycle.

The two measures showed a rank correlation of .59, but their agreement was uneven. Strong performers often did well on both; individual classifications in the middle of the distributions were less consistent. The authors concluded that counting could contain a perceptual contribution while still failing to reveal whether a particular person's score arose from perception or estimation (Knoll & Hodapp 1992, pp. 220-222).

Schandry, Bestler and Montoya took another approach in 1993. By varying posture and exercise, they examined whether counting performance changed with cardiac mechanics. Across their experimental conditions, accuracy was associated with measures related to the movement of blood, particularly its estimated momentum. They interpreted this as evidence for a sensory contribution arising in or near the heart (Schandry et al. 1993, pp. 467-474).

Counting could thus respond to bodily conditions without revealing whether an individual counted through perception or estimation. Discrimination offered no uncontested standard against which to settle the issue, since its scores depended on timing and task demands of their own. The disagreement reflected different experimental purposes as much as competing claims to measure the same thing.

6. New delays, new purposes

Although the multi-interval findings challenged the use of a universal synchronous delay, they also suggested how a fixed pair might be improved. A laboratory could move the early signal into the region between 200 and 300 milliseconds that many participants preferred and place the late signal outside it. The convenience of binary discrimination would remain, now informed by the very findings that had challenged its original timing.

Eichler and Katkin's 1994 study made that reasoning explicit. Their fixed-delay condition used tones at 200 and 500 milliseconds after the R-wave. The early interval was justified through the accumulating timing studies; the late interval had not been a modal preference in Brener and Kluvitse's adjustment experiment. They compared this pair with a version of the drifting-delay task that also began its synchronous condition at 200 milliseconds (Eichler & Katkin 1994, pp. 230-231, 233).

With the fixed pair, about 40 per cent met the study's detector criterion in that condition, compared with 26 per cent in the drifting condition. The difference arose particularly from the ability to identify the asynchronous trains. The drifting signals could remain too close to the likely sensation during the early portion of a trial (Eichler & Katkin 1994, pp. 231-233). The fixed-delay tradition thus incorporated the results of its critics, even though it retained the assumption they questioned. The numbers changed; the decision to assign a common synchronous interval remained.

The detector rule also admitted consistent reverse discrimination: in forty trials, either at least 26 correct answers or 14 or fewer could qualify. The paper explicitly acknowledged the contribution of both tails to its error probability (Eichler & Katkin 1994, p. 231). Even with the same pair of delays, the classification depended on what the investigator allowed a successful response to look like.

Wiens and Palmer's 2001 paper offered a more developed defence of selected interval pairs. It also questioned the usual classification of multi-interval performance. A chi-square test can detect any sufficiently strong dependence of responses on delay. But if sensations generally occur between the earliest and latest delays, a detector might be expected to produce a particular pattern: fewer simultaneity judgments at the ends and more in the middle.

Their quadratic contrast tested a broad inverted-U pattern, with negative weights at the earliest and latest delays and the largest positive weights at 200 and 300 milliseconds. Some resulting scores related to gender and affect intensity where the general chi-square score did not (Wiens & Palmer 2001, pp. 168-174). These associations offered support within a theoretical account, though neither outcome was an independent standard of heartbeat perception.

The chi-square criterion admitted unusual response patterns as evidence of discrimination, whereas the quadratic criterion favoured the pattern predicted by its authors' account of heartbeat timing. Brener and Ring later objected that sensations at more than one bodily site could yield more than one peak, poorly represented by a single broad quadratic contrast (Brener & Ring 2016, pp. 5-6). The statistical choice thus carried a substantive judgment about how much variation a perceptual response could show.

Wiens and Palmer also analysed the 200/500 and 0/200-millisecond subsets of the six-delay data. Those analyses offered evidence relevant to using two selected delays. Their experiment did not directly compare the original 128/384 pair with 200/500, and it did not test a 250/550 pair. These distinctions become important in later papers that cited it for the superiority or precise timing of a particular implementation (Wiens & Palmer 2001, pp. 165-168, 171-175).

6.1. A different reference point

The move into neuroimaging gave the task a wider explanatory role. Critchley and colleagues' 2004 experiment asked participants to judge heartbeat-related tones while undergoing functional magnetic resonance imaging. A comparison condition required judgments about the tones' pitch. Heartbeat performance was associated with activity and grey matter volume in the right anterior insula, helping connect individual differences in cardiac task performance with a proposed neural basis of bodily awareness (Critchley et al. 2004, pp. 189-191).

During imaging, timing came from a pulse oximeter on the finger. Tones began at detection of the pulse or 500 milliseconds later (Critchley et al. 2004, p. 194). "Immediate" now meant immediate relative to a peripheral waveform, rather than to the electrical event recorded by the electrocardiogram.

To compare a finger-pulse delay with an R-wave delay, one must account for the events between them. Ventricular electrical activity is followed by mechanical events, pulse propagation and detection of a particular feature of the peripheral waveform. The recording system contributes to that detection. Until the triggering event is specified, the same numerical delay can refer to different portions of the cardiac sequence.

Garfinkel and colleagues made the conversion explicit in 2015. Tones began at the rising edge of the finger pulse or 300 milliseconds later. Assuming an average interval of about 250 milliseconds from the R-wave to the finger pulse, they described the conditions as approximately 250 and 550 milliseconds after the R-wave (Garfinkel et al. 2015, p. 67). These coordinates depended on an average taken from earlier work, not an individual ECG-to-pulse interval measured in the experiment.

They cited Wiens and Palmer for the claim that 250 and 550 milliseconds corresponded to maximum and minimum synchrony judgments. The earlier experiment supported the relevance of an intermediate region and a later comparison, but had not tested that exact pair. The citation had made the empirical basis more specific than its source allowed. Related procedures then appeared in studies of autism and across cardiac and respiratory tasks (Garfinkel et al. 2016b, manuscript p. 19; Garfinkel et al. 2016a, p. 3).

A further contribution of Garfinkel and colleagues' study was to distinguish performance from the participant's knowledge of that performance. Their framework separated *interoceptive accuracy*, assessed through behavioural tasks; *interoceptive sensibility*, assessed through reports of bodily awareness or perceived aptitude; and *interoceptive awareness*, operationalized as the relationship between confidence and accuracy (Garfinkel et al. 2015, pp. 66-68). Under this definition, *awareness* had a more specific meaning than in much of the earlier literature. A person might perform well yet have little sense of which answers were correct; another might distinguish successful judgments from guesses despite modest average performance. Their mean confidence would not capture that difference.

Although confidence ratings had been collected earlier, Garfinkel and colleagues made their relationship to correctness an object of study, using a receiver operating characteristic analysis to distinguish confidence in correct and incorrect decisions (Katkin et al. 1981, pp. 1098-1099; Garfinkel et al. 2015, pp. 67-68). That analysis inherited the first-order task's timing assumptions. It measured how confidence distinguished answers scored as correct from those scored as incorrect, rather than providing an independent standard for when the heartbeat was felt. The 2015 protocol used fifteen discrimination trials of ten tones each. That made it brief, but it left few observations from which to estimate either performance or the relationship between confidence and correctness. Garfinkel and colleagues acknowledged limited sensitivity for metacognitive assessment (Garfinkel et al. 2015, pp. 67, 71).

Using 100-trial records from 174 participants on a 200/500-millisecond task, Kleckner and colleagues compared shorter estimates with the longer record and simulated consequences for detecting individual-difference associations. They recommended at least forty trials, with forty to sixty needed for their reliability benchmark (Kleckner et al. 2015, pp. 1432, 1435-1439). Agreement with a longer record was not independent retest stability, and a benchmark for accuracy did not determine how many observations metacognition required.

A longer task could support a more precise score, but occupied time needed for other measurements. Shortening it widened its practical uses while introducing noise that could weaken the associations those uses sought to explain. Between Whitehead's 200 trials and fifteen-trial applications, the procedure's place in a research programme had changed its measurement properties. Continuity of the task name did not ensure continuity of the estimate.

By this point, both the answer demanded of the observer and the event used to time the signal had changed. Yet another continuity remained: accounts of the delays often appealed to a physiological rationale inherited from the original method. Following that rationale backwards reveals a second history, running through the same experiments.

7. How the delays acquired a physiology

Whitehead and colleagues offered a brief estimate of contraction and pulse arrival, leaving the sensory source unspecified. Later accounts attached more definite physiology to the same interval. Clemens associated it with the end of ventricular contraction and the carotid sinus; Davis and colleagues referred to aortic baroreceptors; Brener and Kluvitse described vascular receptors and added that the later interval fell after discriminable cardiac events had ended. Ring and Brener referred to mechanoreceptive sources in the neck and arm, while Brener and Ring's later review located relevant receptors in and near the heart (Clemens 1984, p. 187; Davis et al. 1986, p. 76; Brener & Kluvitse 1988, p. 554; Ring & Brener 1992, pp. 535-536; Brener & Ring 2016, p. 3).

Following these descriptions backwards leads to a less specific original claim. They nominate different sites, sometimes move from a mechanical event to a sensation, and supply a rationale for the late interval that Whitehead's paper did not give. Their citation relations also differ. Davis cited McDonald for the aortic-pressure-wave account and acknowledged that the sensory pathways remained unidentified (Davis et al. 1986, p. 76). That separately sourced interpretation should not be treated as though it were simply a false quotation of Whitehead. Nor do the discrepancies establish that each author copied the preceding one. They show a familiar instrument acquiring explanations that require evidence of their own.

The distinction separates two questions that the later literature can easily bring together: why a delay was originally selected, and what might make it physiologically useful. A plausible account of the latter cannot establish the former. The same distinction applies to the later 250/550-millisecond pair. A broad region of preferred timing may motivate that choice without an earlier experiment having established those exact values as an optimum.

Research on cardiac-cycle effects supplied a more developed physiological argument. In discussing cutaneous detection, Edwards and colleagues combined estimates of pulse arrival with the time course of arterial baroreceptor discharge. They proposed integrated output extending approximately from 90 to 390 milliseconds after the R-wave, greatest around 250 milliseconds (Edwards et al. 2009, p. 255). The derivation combined earlier cardiovascular measurements, a physiology text and animal recordings. It was a synthesis for interpreting an effect of the cardiac cycle on another sensory modality, rather than a measurement of when an individual human experienced a heartbeat.

The pulse reaching a receptor, the receptor discharging, neural activity altering processing and a person judging a tone simultaneous with a sensation are related events, but not interchangeable timestamps. Evidence about one can constrain explanations of another. It cannot simply be carried across as a measurement of conscious timing. Garfinkel and colleagues' fear experiment, for example, presented stimuli at approximately 300 milliseconds after the R-wave to examine a selected cardiac phase; it did not estimate subjective simultaneity (Garfinkel et al. 2014, p. 6574). More recent scrutiny of inconsistent cardiac-phase labels shows that the problem extends beyond heartbeat discrimination itself (Caparco et al. 2025, pp. 2-3, 7).

Physiology therefore helps explain why parts of the cardiac cycle may matter without settling the task's original assumption. From the first conjecture about pulse arrival to the later receptor accounts, the missing link remains the relation between a bodily event and the judgment by which it is said to have been perceived. The experiments that allowed observers to locate their own sensations made that link a question rather than an answer built into the apparatus.

8. What the history leaves open

A question about two delays has led through several changes in what investigators meant by cardiac awareness. Whitehead's task served a dispute about learning and control. Its critics asked whether the

stipulated timing could recognize the variation among observers. Later studies used related judgments to explain emotion, brain activity and knowledge of one's own performance. The procedure remained recognizable while its purpose, reference event and criterion of success changed. A history of those changes is therefore part of the meaning of the measurement.

The early dispute did not produce a universally satisfactory delay. It showed why that was unlikely to be the only problem worth solving. A person could exploit a cardiac-related cue, prefer an interval labelled incorrect, or produce a plausible median without a differentiated response. More elaborate scoring did not remove those possibilities. Conversely, failure under a particular timing rule did not establish an absence of cardiac experience. The enduring question is what information the observer used to make the judgment, and what the design permits us to infer from it.

Recent methods continue to separate questions once compressed into an accuracy score. Heart rate discrimination asks whether an external rhythm is faster or slower than the perceived heart rate, allowing the location and steepness of a response function to be estimated separately. Those properties need not have the same stability: in the original study, retest correlations were .51 for threshold and .10 for slope (Legrand et al. 2022, p. 9). The task makes an estimate of rate available for study; it does not locate a felt beat within its cycle or identify, by itself, the sensory and cognitive contributions to that estimate.

The phase adjustment task (PAT) returns to the observer's search for a match, now with contemporary recording and analysis methods (Plans et al. 2021). Recent versions distinguish selections consistent in absolute delay from those consistent in cardiac phase, and an ECG implementation changes the triggering reference once again (Palmer et al. 2026, section 2.1.6; Hodgetts et al. 2026, sections 1 and 3.4). Whether consistency is used to classify a participant or interpreted continuously remains consequential. These developments extend the older measurement questions rather than make them obsolete. Likewise, Cardioception's developmental heartbeat synchrony discrimination task samples delays adaptively, but its assumed response shape and stopping criterion require validation; its documentation does not yet report a participant validation study (Cardioception 2026).

The next explanatory step, I suggest, is to connect these judgments to physiological observations through models of how the responses are generated. The history gives a concrete reason for doing so. A broad synchrony curve might arise because the bodily signal varies in timing, because the comparison with the tone varies, or because both do. Fitting the curve cannot allocate that variation by itself. Recording the ECG, peripheral pulse and delivered stimulus onset can constrain some explanations; manipulations of the comparison or of information available to the observer can test others. A useful generative model would make these alternatives explicit and show which observations distinguish them. More parameters would be warranted only where the experiment supplies the information to estimate them.

That information cannot be assumed from a small trial count. One adjustment may involve many beats and repeated exploration, whereas one fixed-delay judgment may follow a single brief train. The PAT argument for fifteen to twenty trials concerns identifying sufficiently consistent choices against a specified chance process, not every question about stable individual differences or sensory precision (Hodgetts et al. 2026, introduction). Kleckner's recommendation concerned a different score under a different procedure. Adaptive sampling can improve efficiency, but locating a peak does not establish that its width, its physiological source and its stability have also been measured. Very brief tasks remain inadequate for claims that outrun the observations they provide.

None of this requires a single method to answer every question about the heart. It requires the question to remain visible as the method travels. Counting, rate judgments, temporal matches and confidence reports can each be informative when the inference is suited to the response. Their usefulness is weakened when a familiar name allows differences in purpose or implementation to disappear.

The original delays were practical choices, explained by a physiological conjecture rather than an established law of sensation. Their subsequent history is instructive because so much experimental reasoning survives beneath numbers that now look conventional. Recovering that reasoning lets us retain the ingenuity of the methods without inheriting their assumptions unexamined. The point is not to find the one delay the field should always have used. It is to understand what each choice enabled observers to tell us about their hearts, and what it left unknown.

Disclosures

Micah G. Allen is a coauthor of the original heart rate discrimination paper and the creator and maintainer of the Cardioception toolkit and its heartbeat synchrony discrimination task. He developed Cardioception's port of the phase adjustment task, but did not originate PAT. The discussion of these methods reflects those research interests.

This experimental AI-assisted review was directed by the author's research questions and years of engagement with the literature. Claude Code assisted with finding and organizing sources, summarizing articles, extracting facts, constructing tables and drafting a historical timeline and initial manuscript. Codex assisted with prose revision, editorial organization, targeted source checks and LaTeX preparation. AI-generated summaries and interpretations require verification against the primary texts. The author retains responsibility for the argument and for the accuracy of the final manuscript.

Appendix: scoring and sources

The detector label depends on the response structure and scoring rule. In Whitehead's single-train classification, hit and false-alarm rates describe how often the same answer is given to the two conditions. Signed sensitivity separates those rates under its model, while criterion describes response preference. Selecting the signal label from the responses, or replacing a signed difference with its absolute value, changes the null distribution: positive unsigned values can arise from sampling variation. A threshold allowing reverse discrimination must account for that possibility. This is why a detector proportion cannot be interpreted without its rule, trial count and handling of extreme response rates.

The historical scales also differ. They include d', A', hit rate minus false-alarm rate and the transformed area measure 2 arcsin $\sqrt{P(A)}$. For the latter, an area of .5 yields $\pi/2$, approximately 1.57; the chance value for A' is .5 (Katkin et al. 1981, p. 1099; Jones et al. 1984, p. 145; Davis et al. 1986, p. 78). Adjustment tests may compare selected delays with a uniform distribution, whereas the 1993 constant-stimuli test crossed delay with simultaneous versus nonsimultaneous responses. Wiens and Palmer's quadratic contrast instead used weights of -5, 1, 4, 4, 1 and -5 to test a broad central peak (Brener et al. 1993, p. 661; Wiens & Palmer 2001, p. 168). These procedures do not make identical claims about the shape of perceptual responding.

This review is a narrative reconstruction from accumulated reading, citation tracing and targeted primary-source checks, not an exhaustive systematic search. It concentrates on the timing debates of the 1970s through the 1990s, their later applications and selected developments available by 6 November 2026. Explicit citation links support claims of attribution or adoption; other similarities between methods are comparisons made here, not evidence of personal influence.

Brener's founding theoretical chapters, Schandry's original counting paper and the original Störmer report remain unavailable in the present collection. Their accounts are therefore mediated through the sources identified in the text. Cardioception documentation is cited at the recorded commit rather than as an unversioned description of a changing task.

References

Brener J, Jones JM. 1974. Interoceptive discrimination in intact humans: detection of cardiac activity. *Physiology and Behavior* 13:763–767

Brener J, Kluvitse C. 1988. Heartbeat detection: judgments of the simultaneity of external stimuli and heartbeats. *Psychophysiology* 25:554–561

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* 30:657–665

Brener J, Ring C. 2016. Towards a psychophysics of interoceptive processes: the measurement of heartbeat detection. *Philosophical Transactions of the Royal Society B* 371:20160015

Brener J, Ring C, Liu X. 1994. Effects of data limitations on heartbeat detection in the method of constant stimuli. *Psychophysiology* 31:309–312

Caparco A, Lopez-Martin G, Galvez-Pol A. 2025. Phase confusion: how inconsistent cardiac labeling obscures interoception research. *Biological Psychology* 199:109078

Cardioception. 2026. Development documentation: Measuring cardiac interoception; Phase Adjustment Task; Heartbeat Synchrony Discrimination task. `https://github.com/embodied-computation-group/Cardioception/tree/f28491c58ae22b4aec418c4725b15e9882c7021b/docs/source/tasks`. Commit f28491c58ae22b4aec418c4725b15e9882c7021b; consulted 6 November 2026

Clemens WJ. 1979. Assessment, learning, and retention of heart beat discrimination. *Psychophysiology* 16:333–338

Clemens WJ. 1984. Temporal arrangement of signals in heartbeat discrimination procedures. *Psychophysiology* 21:187–190

Critchley HD, Wiens S, Rotshtein P, Öhman A, Dolan RJ. 2004. Neural systems supporting interoceptive awareness. *Nature Neuroscience* 7:189–195

Dale A, Anderson D. 1978. Information variables in voluntary control and classical conditioning of heart rate: field dependence and heart-rate perception. *Perceptual and Motor Skills* 47:79–85

Davis MR, Langer AW, Sutterer JR, Gelling PD, Marlin M. 1986. Relative discriminability of heartbeat-contingent stimuli under three procedures for assessing cardiac perception. *Psychophysiology* 23:76–81

Edwards L, Ring C, McIntyre D, Winer JB, Martin U. 2009. Sensory detection thresholds are modulated across the cardiac cycle: evidence that cutaneous sensibility is greatest for systolic stimulation. *Psychophysiology* 46:252–256

Ehlers A, Breuer P. 1992. Increased cardiac awareness in panic disorder. *Journal of Abnormal Psychology* 101:371–382

Eichler S, Katkin ES. 1994. The relationship between cardiovascular reactivity and heartbeat detection. *Psychophysiology* 31:229–234

Garfinkel SN, Manassei MF, Hamilton-Fletcher G, In den Bosch Y, Critchley HD, Engels M. 2016a. Interoceptive dimensions across cardiac and respiratory axes. *Philosophical Transactions of the Royal Society B* 371:20160014

Garfinkel SN, Minati L, Gray MA, Seth AK, Dolan RJ, Critchley HD. 2014. Fear from the heart: sensitivity to fear stimuli depends on individual heartbeats. *Journal of Neuroscience* 34:6573–6582

Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD. 2015. Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology* 104:65–74

Garfinkel SN, Tiley C, O'Keeffe S, Harrison NA, Seth AK, Critchley HD. 2016b. Discrepancies between dimensions of interoception in autism: implications for emotion and anxiety. *Biological Psychology* 114:117–126

Hodgetts J, Gallicchio G, McKiernan S, Ring C, Murphy J, Cooke A. 2026. The electrocardiogram-based phase adjustment task (PAT-ECG): An open-access laboratory test of cardiac interoceptive accuracy. *Psychophysiology* 63:e70415

Jones GE, Collins SW, Dabkowski EA, Jones KR. 1988. The discriminability of temporal patterns used in the Whitehead discrimination procedure. *Psychophysiology* 25:547–553

Jones GE, O'Leary RT, Pipkin BL. 1984. Comparison of the Brener-Jones and Whitehead procedures for assessing cardiac awareness. *Psychophysiology* 21:143–148

Katkin ES, Blascovich J, Goldband S. 1981. Empirical assessment of visceral self-perception: individual and sex differences in the acquisition of heartbeat discrimination. *Journal of Personality and Social Psychology* 40:1095–1101

Kleckner IR, Wormwood JB, Simmons WK, Barrett LF, Quigley KS. 2015. Methodological recommendations for a heartbeat detection-based measure of interoceptive sensitivity. *Psychophysiology* 52:1432–1440

Kleinman RA. 1970. The development of voluntary cardiovascular control. Ph.D. thesis, University of Tennessee

Kluvitse CD. 1987. Discrimination of cardiac activity. Ph.D. thesis, University of Hull

Knoll JF, Hodapp V. 1992. A comparison between two methods for assessing heartbeat perception. *Psychophysiology* 29:218–222

Legrand N, Nikolova N, Correa C, Brændholt M, Stuckert A, et al. 2022. The heart rate discrimination task: A psychophysical method to estimate the accuracy and precision of interoceptive beliefs. *Biological Psychology* 168:108239

Mandler G, Kahn M. 1960. Discrimination of changes in heart rate: two unsuccessful attempts. *Journal of the Experimental Analysis of Behavior* 3:21–25

Mandler G, Mandler JM, Uviller ET. 1958. Autonomic feedback: the perception of autonomic activity. *Journal of Abnormal and Social Psychology* 56:367–373

McFarland RA. 1975. Heart rate perception and heart rate control. *Psychophysiology* 12:402–405

Montgomery WA, Jones GE. 1984. Laterality, emotionality, and heartbeat perception. *Psychophysiology* 21:459–465

Palmer R, Murphy J, Bird JM, Donaghy R, Piercy T, et al. 2026. Refinements of the phase adjustment task (PAT 2.0). *Frontiers in Psychology* 17:1677186

Plans D, Ponzo S, Morelli D, Cairo M, Ring C, et al. 2021. Measuring interoception: The phase adjustment task. *Biological Psychology* 165:108171

Ring C, Brener J. 1992. The temporal locations of heartbeat sensations. *Psychophysiology* 29:535–545

Ross A, Brener J. 1981. Two procedures for training cardiac discrimination: a comparison of solution strategies and their relationship to heart rate control. *Psychophysiology* 18:62–70

Schandry R, Bestler M, Montoya P. 1993. On the relation between cardiodynamics and heartbeat perception. *Psychophysiology* 30:467–474

Schneider TR, Ring C, Katkin ES. 1998. A test of the validity of the method of constant stimuli as an index of heartbeat detection. *Psychophysiology* 35:86–89

SPR abstracts. 1976. Abstracts of papers presented at the Fifteenth Annual Meeting of the Society for Psychophysiological Research. *Psychophysiology* 13(2):160–187. Meeting held in Toronto, 16–19 October 1975

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

Wiens S, Palmer SN. 2001. Quadratic trend analysis and heartbeat detection. *Biological Psychology* 58:159–175

Yates AJ, Jones KE, Marie GV, Hogben JH. 1985. Detection of the heartbeat and events in the cardiac cycle. *Psychophysiology* 22:561–567