Biological Psychology 152 (2020) 107869

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Commentary

A heartfelt response to Zimprich et al. (2020), and Ainley et al. (2020)'s commentaries: Acknowledging issues with the HCT would benefit interoception research

Olivier Corneille*, Olivier Desmedt[1], Giorgia Zamariola, Olivier Luminet[1], Pierre Maurage[1]

Psychological Science Research Institute, UCLouvain, Louvain-la-Neuve, Belgium

ARTICLE INFO

Keywords:
Interoception
Interoceptive accuracy
Construct validity
Methods

ABSTRACT

In two Commentaries, Zimprich et al. (2020), and Ainley et al. (2020) dispute conclusions raised by Zamariola et al. (2018) in a large sample study that questioned the validity of IAcc scores derived from the Heartbeat Counting Task (HCT). After clarifying the reliability of our procedures and the robustness of our main findings, we address the four points of contention discussed in the Commentaries. In doing so, we spell out why research using the HCT faces important interpretational issues, and we call for a theoretical clarification on the construct. In our conclusion, we provide recommendations for improving HCT research and research on interoception in general.

In two Commentaries on Zamariola et al. (2018), Zimprich, Nusser, and Pollatos (2020), and Ainley, Herbert, Tsakiris, Pollatos, and Schulz (2020), raise concerns about our original publication and provide constructive evaluations of our claims. Consistent with these authors, the purpose of the present Response is to contribute to advancing research on cardiac interoception. Let us say from the outset that we agree with some concerns raised in the Commentaries. We find it important, however, to qualify and communicate our disagreement with several others. We also explain why we wish a critical point we discussed in our original article were elaborated on in the Commentaries. As a first step, we clarify the reliability of our procedure and the robustness of our main findings. We then consider the four points of contention discussed in our original publication and in the two Commentaries. In doing so, we spell out why research using the Heartbeat Counting Task (HCT) faces important interpretational issues. In our conclusion, we provide recommendations for improving HCT research and research on interoception in general.

1. A clarification regarding the procedures

Ainley et al. (2020) suggest that the few odd data observed in our dataset could be due to a problematic recording procedure. We can report here that the procedure with the Polar Watch RS800CX included damping the electrodes with water and instructing participants not to raise the wrists from the electrodes. If the connection was interrupted at any time, the exercise was repeated. This procedure complies with current practice. The first author of our original article learned this procedure during a research stay at one of the current commentators' lab. A similar procedure was used recently by authors of the two Commentaries (Durlik & Tsakiris, 2015; Georgiou et al., 2015; Zamariola, Cardini, Mian, Serino, & Tsakiris, 2017).

Of importance too, our main findings have been replicated since then in two studies that used the Polar Watch recording procedure (our lab, UCLouvain, Belgium) and a Pulse oximeter (Geoffrey Bird's lab, University of Oxford, The United Kingdom), after excluding outliers (participants who had a heart rate and a mean difference of heart rate between intervals of more than three times the interquartile range). Specifically, we found in these new datasets (1) that 97%97 % (our lab, original instructions), 100 % (our lab, modified-instructions – see Section 3), and 96 % (GB's lab, modified instructions) of the participants underestimated their heartbeats, (2) that the correlation between reported and actual heartbeats was $r = .24$, $p = .008$, $N = 120$ (our lab, original instructions), $r = -.01$, $p = .943$, $N = 120$ (our lab, modified instructions), and $r = .21$, $p = .003$, $N = 194$ (GB's lab, modified instructions), (3) that interoceptive accuracy (IAcc) scores correlates with heart rate by $r = -.32$, $p < .001$, $N = 120$ (our lab), r

* Corresponding author at: UCLouvain-IPSY, 10 Place du Cardinal Mercier, B-1348, Louvain-la-Neuve, Belgium.
E-mail address: olivier.corneille@uclouvain.be (O. Corneille).
[1] Also at the Fund for Scientific Research – Belgium (FRS-FNRS).

https://doi.org/10.1016/j.biopsycho.2020.107869
Received 20 December 2019; Accepted 10 February 2020
0301-0511/

$= -.21$, $p = .024$, $N = 120$ (our lab, modified instructions), but $r = -.004$, $p = .96$, $N = 194$ (GB's lab, modified instructions), (4) that IAcc decreased at longer time intervals (see Section 5, "time intervals").[2]

Given the general stability of effects independently found across labs (Belgium and the UK) that relied on two different recording procedures (Polar Watch and Pulse Oximeter), with and without exclusion of outliers, and with original and modified instructions, it is implausible that the findings reported in our original article are due to procedural issues (this, in addition to the fact that we relied on a standard procedure).

That a few extreme values were observed is better explained by our unusually large dataset. With larger samples come more extreme values. Surely, extreme values should be spotted, and analytic outcomes should be compared when including versus excluding these values. Unfortunately, there is currently no clear guideline regarding how this identification should be realized. Many criteria may be considered, among which: outliers in IAcc, mean heart rate (HR) across intervals, mean reported heartbeats (HB), HR variation across intervals, HR before task completion, as well as multivariate outliers. In the absence of clear exclusion criteria, we decided that our sample could accommodate a few extreme but realistic (i.e., representing acceptable variations in such large sample) values. This assumption proved reasonable since – as we just explained - the findings have been largely replicated since then in unrelated datasets that excluded extreme values and relied on two different recording procedures.

As to the eight participants (i.e., 1,38 %) who were not included in the analyses, the reason for removing them was duly reported in our original article: these participants were removed because of recording issues (this was mentioned in Section 3.1, "data exclusion"). Therefore, Ainley and colleagues mislead the reader when stating that we did "not state the criteria by which they excluded eight of their participants" (p. 8). We should have reported that 580 (not 572) participants were recruited, but this could have been easily figured out from the numbers reported (i.e., $8/(572 + 8) = 1.38$ %). These eight participants were excluded after inspection of the recorded data (i.e., they showed extremely high recorded heartbeats values), which complies with Ainley and colleagues' current recommendations.

Now that we have clarified the reliability of our procedures, the reliability of our data, and the robustness of our main findings, we turn to the four substantive points, namely (i) most IAcc scores reflect underestimations, (ii) there is a small association between actual and reported heartbeats at the sample level, (iii) IAcc scores correlate with actual heartbeats, and (iv) IAcc scores decrease at longer time intervals.

2. Underestimation of actual heartbeats in the HCT

There is one point where datasets, analyses, and interpretations converge: almost all observations in the HCT are underestimations. Zimprich and colleagues dispute our correlation between IAcc scores and the difference between actual and reported heartbeats (because the association between these two is not linear). This correlation and the scatterplot associated with it, however, were merely meant to illustrate our point of interest. That is, nearly all (i.e., 95 %) participants in the HCT underreport their heartbeats, and IAcc scores mostly capture by how much participants underreport their heartbeats, an effect that "deserves further investigation and theorizing." (Zimprich et al., p. 6).

Ainley and colleagues explain that overestimations are not relevant to interoceptive accuracy: "Were they to over-count, this would imply that they were hallucinating heartbeats that were never actually experienced." (p. 4). These authors are clear here (i) that they consider that the overreporting of internal experiences is of little interest for research on interoception and (ii) that "accuracy" can be equated with misses. We explain below why both views are problematic.

2.1. Overreporting is of practical and theoretical significance

Overreporting is relevant to clinical research and research on interoception. Patients with pathological health anxiety (e.g., Krautwurst, Gerlach, & Witthöft, 2016) and anxiety sensitivity (e.g., Petersen & Ritz, 2010), as well as people with high self-reported psychosomatic symptoms (e.g., Petersen, Van Staeyen, Vögele, von Leupoldt, & Van den Bergh, 2015) may overreport internal signals. Overreporting may be partly driven by perceiving an internal signal in the absence of interoceptive cues ("hallucinations") or by misinterpreting noisy/ambiguous internal information as a signal ("illusion") (Norton & Corbett, 2000). Studying overreporting is a valuable research endeavor. Even if overreporting were confined to the realm of "hallucinatory" phenomena, this research would remain worthwhile as patients may suffer from those.

We also pointed out that participants who overcounted their heartbeats were characterized by high IAcc scores in our study. Hence, HCT performance does not distinguish between participants who are assumed to have excellent interoceptive abilities (i.e., those who underreport the less) and participants that Ainley and colleagues see as "hallucinators" (i.e., those who overreport). That HCT performance blends two different populations (according to Ainley and colleagues' analysis) questions the task validity. But does it? This leads to our second point, one we discussed in the original article, and that we regret did not attract more attention in either of the two Commentaries:

2.2. The level of underreporting should not be confused with "accuracy"

The extent to which individuals underreport (and, actually, overreport) their heartbeats in the HCT has not just to do with their interoceptive abilities. It also has to do with their response bias: the tendency to favor a response over another (see also Petersen et al., 2015). HCT performance may result in strikingly different IAcc scores for reasons that are entirely uninformative about participants' interoceptive abilities. In order to tease apart the contribution of response bias versus accuracy, one would generally need to implement control over the presence, frequency, or intensity of the signal. Without achieving this, "accuracy scores" cannot be interpreted as having much, little, or hardly anything to do with participants' ability to discriminate internal signals. It is fine to consider HCT performance as an indicator of underestimation. It is misleading, however, to equate underestimation with *accuracy*, i.e., the capacity to tease apart signal and noise.

In order to illustrate this point, imagine an eye test consisting of the detection of the letters "HB" on a computer screen (unlike HCT research, the signal intensity, i.e., eyes-to-screen distance, is held constant). There are 100 trials, all including the "HB" sequence appearing for 10 msec, sandwich-masked by visual noise. A trial is flashed every second, and two patients have to press the spacebar to enter their response on each trial (unlike HCT research, this limits a memory contribution to scores). Participant A answers "HB Present" 99 % of the time. Participant B answers "HB present" 1% of the time. In this case, A has only one miss (and thus receives an almost perfect "HB" accuracy score). B has 99 misses (and thus receives an almost null "HB" accuracy score).

In terms of accuracy, however, these two participants may be perfectly equal. In order to know, one would need to include trials where the "HB" sequence does *not* appear and look at how participants A and B respond to these trials (e.g., Kellen & Klauer, 2018). Should A and B respond similarly to trials where the signal is present and absent (i.e., if A and B answer 99 % and 1% of the time, respectively, that the signal is present no matter whether it is present or absent), "HB" detection

[2] The data for these unpublished analyses were collected by Desmedt, Corneille et al. (2019) and can be found at the following link: https://osf.io/7b39n/?view_only=522315d79a9944bf97623825cbbcc4b6.

scores (i.e., the number of misses) would widely vary between these participants. This, however, would only reflect different response biases, despite similar (i.e., null) levels of accuracy.

Similar to this thought experiment, participants' heart hardly stops beating while completing the HCT. Therefore, accuracy cannot be validly estimated. All we know is how large the underestimation is, and this underestimation is hardly interpretable in the absence of experimental intervention tackling its contributors.

3. Low sample-level correlation between reported and actual heartbeats

3.1. Is a sample-level correlation unwarranted for examining the HCT validity?

We found overall a low sample-level correlation between reported and actual heartbeats. We thought this was a surprising finding when the HCT is meant to assess participants' capacity to count their heartbeats. The two Commentaries argue that our conclusions are unwarranted because we relied on sample-level correlations. Zimprich and colleagues explain that: "(…) individual IAcc scores are not (and can not be) based on the association between actual and counted heartbeats in a sample of individuals." (p. 8). We agree with this principle. However, IAcc was not considered in this first analysis. Instead, we found it surprising that, in a *task* meant to capture how good people are at detecting their heartbeats, the sample-level correlation between reported and actual heartbeats is so low. If the correlation is low, this indicates that counted heartbeats are only loosely related to actual heartbeats in this task; in all likelihood, this is because the task is too challenging to complete at rest, thereby making room for the contribution of non-interoceptive processes (e.g., estimation strategies) to task performance. A sample-based correlation does not provide direct information on IAcc scores. However, if the sample-based correlation potentially suggests a lack of validity of the task, then it is unclear how scores derived from this task should be interpreted.

It is noteworthy that Schandry (i.e., the main promotor of the HCT) coined this very same sample-level correlation "cardioceptive sensibility" (Schandry, Bestler, & Montoya, 1993; see Fig. 5). Furthermore, in order to validate this approach to IAcc, these authors showed that this sample-based correlation increased at higher levels of physical intensity and with more tilted bodily postures. This is consistent with Ainley and colleagues' view that the dependency of cardiac interoception (here, according to Schandry et al. (1993), the sample-level correlation itself) on cardiac signal is "evidence in favor of the test's construct validity" (p.11).

In sum, our overall correlational approach (i) is consistent with Schandry's approach to IAcc, and (ii) dovetails Ainley and colleagues' validation standards (which we disapprove of, as we explain later). It is remarkable that the analytic rationale we relied on remains unquestioned when used by Schandry and colleagues for supporting the validity of the HCT but is criticized when delivering outcomes questioning this task's validity.

3.2. What can we learn from the quintile analysis?

We divided our sample into quintiles of IAcc scores and expected the correlation between actual and reported heartbeats to increase linearly across IAcc quintiles. Instead, the correlation was highest at the median quintile. Zimprich and colleagues point out that correlations should be corrected for differences in variance across quintiles. These authors show that, in our sample, variance in reported heartbeats increases across quintiles, whereas variance in actual heartbeats decreases across quintiles. When implementing a variance correction, Zimprich and colleagues find no difference between quintiles, except for a slightly larger association between reported and actual heartbeats at the median (i.e., average IAcc scores) than at the fifth (i.e., best IAcc scores) quintile. Therefore, it is fair to concede that outcomes from the quintile analyses are "by far not as compelling as it might appear" (p. 10).

Ainley and colleagues additionally raise concerns about the conceptual value of our quintile analysis. These authors explain that (i) if the IAcc formula was different, (ii) if overcounting participants were excluded from the analyses, (iii) if our sample were broken down to semi-decile, then things would be different (i.e., an almost perfect correlation between reported and actual heartbeats would be found). We agree on this. However, (i) the IAcc formula is not different from what it is, (ii) the current IAcc formula is meant to include, not to exclude, participants who overcount, and (iii) our analysis applied to large quintiles, not to small semi-deciles. The simplification in the formula implemented by these authors, however, as well as their characterization of participants who overcount as mere outliers, highlights again that they consider that only undercounting is relevant to interoception. We have explained in the previous section why we find this conceptualization highly problematic.

Finally, the outcome of Ainley and colleagues' quintile analysis differed from ours. We found it important to identify the source of the discrepancy. We figured out that Ainley and colleagues sorted the participants after averaging the IAcc scores computed at the three time intervals [i.e., $(IAcc_{25} + IAcc_{35} + IAcc_{45})/3$], whereas we inadvertently sorted them based on another aggregation formula ([i.e. (1-ABS(Mean Actual BPM - Mean Reported BPM)/Mean Actual BPM)]. These two formulas return IAcc scores that are extremely close from each other (i.e., they correlate by .985) but using one formula or the other results in about 3,8% of the participants moving from one and occasionally two quintiles. In turn, this small change in the allocation of participants to quintiles drives a strikingly large difference in the outcome of the quintile analysis.[3]

This shows that correlations are extremely sensitive to slight changes in the data, *even* in relatively large samples (i.e., each quintile comprised 114–115 participants). Relatedly, this suggests that switching from the classic to a simplified version of the IAcc formula or removing participants who overcount from datasets, is likely to be impactful in IAcc research. Participants who switched quintiles often showed large variations of IAcc scores across time intervals. This indicates the value of a close inspection of individual data points, which was rightfully done by Ainley and colleagues. In turn, this calls for clear guidelines about exclusion criteria, as variations of IAcc scores across time intervals have to our knowledge never been used as an exclusion criterion in IAcc research. More generally, we believe that the IAcc formula used by Ainley and colleagues is the more relevant one. This highlights the value of transparent research practice: the public sharing of our data allowed for our initial conclusion to be easily challenged.

In sum, we agree that the conclusions from the quintile analysis are unwarranted. First, the two Commentaries are correct that interpreting individual scores from sample-level correlations is questionable. Second, only one small difference in the reported-actual correlation is observed across quintiles after variance correction.

3.3. Even though our sample-based approach is debatable, our conclusion proved correct

We suggested that the pattern of correlations between actual and reported heartbeats may indicate that the two are dissociated in HCT studies (i.e., that reported heartbeats poorly reflect felt heartbeats). Even though our demonstration may (i.e., overall analysis) and should (i.e., quintile analysis) be questioned, our conclusion proved correct. In a recent study that relied on semi-structured interviews, participants reported low awareness of their cardiac signals (Zamariola, Frost, Van

[3] The r codes for the comparative quintile analysis can be found at the following link: https://osf.io/q7hnp/?view_only=fdcce17a07c341eaa3b516a10dc06498.

Oost, Corneille, & Luminet, 2019). Critical readers may object that self-reports do not provide conclusive evidence, and they would be right: an experimental study makes for a more compelling test.

Desmedt, Luminet, and Corneille (2018) precisely used this approach in another recent experiment where we manipulated counting instructions. We reasoned that, if participants truly rely on interoception when completing the task, then instructing them to rely on their cardiac sensations only (thereby, reducing the potential use of non-interoceptive strategies) should leave IAcc scores unaffected. Instead, IAcc scores were cut by half under modified instruction conditions. Consistent with our original conclusion, and consistent with participants' self-report on the way they complete the HCT under usual instructions, this indicates that IAcc scores largely capture non-interoceptive variance. This finding, among many others (Phillips, Jones, Rieger, & Snell, 1999; Ring & Brener, 1996, 2018; Ring, Brener, Knapp, & Mailloux, 2015; Windmann, Schonecke, Fröhlig, & Maldener, 1999), raises important questions on how to interpret IAcc scores in the vast majority of HCT studies that relied on non-modified instructions. We surmise these scores massively reflect response bias and estimation strategies rather than genuine interoceptive abilities.

3.3.1. Interim conclusions

IAcc scores derived from the HCT suffer from two critical interpretational issues under usual test conditions. Firstly, these scores do not allow teasing apart accuracy and response bias. Secondly, they are largely contaminated by estimation (i.e., non-interoceptive) processes.

4. Correlation between heart rate and IAcc

Another point of disagreement concerns the correlation between heart rate and IAcc. It is explained in the Commentaries that this relation may reflect a mathematical artifact (i.e., a ratio of two variables will be correlated with one of its components). Whether it is legitimate or not to correlate a ratio variable with its component, however, is far from consensual. Many authors reject the view that ratio variables create "spurious" relations (e.g., Firebaugh & Gibbs, 1985; Kasarda & Nolan, 1979; MacMillan & Daft, 1980).

Even more important, that this correlation may be mathematically constrained does not make the problem vanish at all. On the very contrary, it makes it even worse: better IAcc scorers generally have a slower beating heart; therefore, when an association is found between IAcc and a third variable, it may be interpreted in terms of cardiac condition. For instance, Schandry et al. (1993) noted that: "Habitually good heartbeat perceivers demonstrate a lower heart rate (Schandry, 1981). Because at constant cardiac output a lower heart rate is accompanied by a greater stroke volume, this finding could also be explained as a consequence of a higher stroke volume in good heartbeat perceivers (cf. Dale and Anderson, 1978)." (Schandry et al., 1993, p.468). A .62 correlation was reported between heart stroke and cardiac perception by Bestler, Schandry, Weitkunat, and Alt (1990).

Ainley and colleagues explain that it is desirable that IAcc depends on the cardiac signal. As discussed above, they state that this dependency is "evidence in favor of the test's construct validity" (p.11). Surely, interoceptive accuracy is anchored in physiology. A critical question, however, is whether it is acceptable that a measure of accuracy is *confounded* with characteristics of the to-be-detected signal (e.g., signal intensity) or characteristics of the organ (e.g., its size) that supports it. Consider variations in signal intensity. Going back to the visual acuity example, an ophthalmologist uses a Chart Test with letters for measuring visual accuracy among patients. The eyes-to-chart distance, however, varies across patients (the signal is stronger for some patients than others). The doctor concludes that visual acuity is better for patients sitting two meters closer to the chart. We believe many would question the ophthalmologist's assessment. Ainley and colleagues apparently would not. For these authors, inter-individual variation in cardiac detection abilities can be confounded with interindividual variation in the to-be-detected signal. That is, if a participant completing an easy test performs better than a participant completing a difficult test, one may nevertheless state that the former has better abilities than the latter.

Again, an important question here is what we mean by interoceptive "accuracy." Should this construct refer to individuals' ability to discriminate between internal signal and noise, or should it merely refer to their performance at the HCT task, no matter what factors, psychological or physical (e.g., heart rate), interoceptive or non-interoceptive (e.g., heart rate knowledge), or interoceptive but related either to abilities or to response criteria, contribute to task performance? This is a key issue to be clarified in interoception research, not just for conceptual reasons but also for practical ones. In our eye test example, the doctor may prescribe the wrong lenses if concluding that patients are more myopic when (because) they performed the test five meters rather than three meters away from the chart.

So far, and consistent with an ability view on interoception, interventions (e.g., mindfulness; Bornemann & Singer, 2017) have largely targeted the perceptual component of interoception by trying to train people's capacity to better detect their internal bodily signals. Under a broader conceptualization of IAcc, however, one may as well consider physical interventions (e.g., enhancing signal strength by removing adiposity) or health literacy interventions (e.g., helping people make better informed heart rate guesses) as valid ways to improve a patient's interoceptive abilities. One may then target the most efficient or less intrusive mean to improve task performance (e.g., educating patients on heart rates, or boosting their confidence in their reports to mechanically reduce their underreporting).

It should be clear from the current discussion that specifying the construct and finding ways to tease apart various contributors to task performance is tremendously important for both theoretical and practical reasons. If only one message is retained from the current exchange, we hope it is this one.

5. IAcc scores decrease with longer time intervals

We found smaller IAcc scores at longer time intervals. Following common practice in HCT research (e.g., Durlik & Tsakiris, 2015), the order of time intervals was semi-randomized (not fully randomized, as we reported in our study and as Ainley and colleagues correctly anticipated). There was, in Zamariola et al. (2018), a slight variation in heart rate across intervals (i.e., heart rate decreased from 78,9 to 77,1 to 76,5, across the 35 s., then 25 s., then 45 s. intervals). However, the time interval effect was mainly driven by a lowering proportion of reported heartbeats at longer time intervals (with reported heartbeats -/min - decreasing from 51,5 to 50,5 to 46 across the three intervals).

Critically, the dependency of IAcc scores to time interval has been replicated[4] since then in our lab with a semi-randomized order as well as in GB's lab with a fully randomized order; again, after excluding outliers and by relying on two different recording procedures. Therefore, the interval effect cannot be due to an order effect. In GB's lab ($N = 194$), HCT performance decreased as time interval increased ($M_{25/28} = .49$, $M_{35/38} = .46$, $M_{45/48} = .46$, $M_{100/103} = .44$). A significant mean difference was found between 25/28 s and 45/48 s intervals ($t(193) = 1.99$, $p = .05$) as well as between 25/28 s and 100/103 s intervals ($t(193) = 3.24$, $p = .001$). All other differences were non-significant ($t < 1.73$, $p > .05$). In our lab ($N = 120$), HCT performance also decreased as time interval increased ($M_{25} = .63$, $M_{35} = .61$, $M_{45} = .59$), under original instructions. A significant difference was found between

[4] The data for these unpublished analyses were collected by Desmedt, Corneille et al. (2019) and can be found at the following link: https://osf.io/7b39n/?view_only=522315d79a9944bf97623825cbbcc4b6. Slight differences in results can be found when applying non-parametric analyses (i.e., Wilcoxon Sign Test), but conclusions are generally unchanged.

25 s and 45 s intervals ($t(119) = 4.33$, $p < .001$). All other differences were non-significant ($t < 1.82$, $p > .05$). Under adapted instructions (which prompt participants to avoid guessing), the same pattern was observed ($M_{25} = .32$, $M_{35} = .30$, $M_{45} = .27$), with two out of three differences being statistically significant (25–35s: $t(119) = 1.08$, $p = .29$; 25–45s: $t(119) = -3.67$, $p < .001$; 35–45 s: $t(119) = -2.08$, $p = .04$). Importantly too, although, in our lab, the order of intervals was different between original (i.e., 35sec., 25sec., and 45sec.) and adapted instructions (i.e., 45sec., 25sec., 35sec.), the same pattern (i.e., lower IAcc scores at longer intervals) was observed in both conditions. Ainley and colleagues find no time interval effect in their dataset. This is all fine, but it does not contradict our original and renewed recommendation that researchers check if problems found in our original dataset apply to theirs (Zamariola et al., 2019a, p. 16).

More generally, Zimprich and colleagues note that the effect of time intervals on IAcc is small under the time intervals currently used, and that IAcc scores are highly correlated across various time intervals. We agree with this analysis, as well as with Zimprich and colleagues' recommendation to set constant time intervals across studies to facilitate comparisons across them.

6. Conclusions

6.1. HCT research is less robust and less interpretable than one may think

Ainley and colleagues urged us to correct our "erroneous conclusions" that "have the potential to do disservice" to interoception research. Apart from the quintile analysis, which has been discussed above, however, our current analysis strongly supports our original conclusions. Specifically:

- IAcc scores in the HCT are massively driven by underreports.
- Underreports cannot be interpreted as indicators of interoceptive accuracy as these scores are potentially sensitive to both accuracy and response bias.
- These scores additionally fail to discriminate between participants thought to be very good in interoception (i.e., those who underreport the less) and those thought to "hallucinate" (i.e., those who overreport).
- As a third critical blending issue, HCT performance, under classic task instructions, confounds felt and estimated heartbeats.
- The low correlation between actual and reported heartbeats, consistent with Schandry's conceptualization and dovetailing Ainley and colleagues' validation standards, questions the validity of usual (i.e., resting) task conditions from which IAcc scores are derived.
- The correlation between IAcc scores and actual heartbeats, be it mathematically constrained, makes it challenging to interpret correlates of IAcc scores in terms of cardiac interoception versus cardiac condition.
- Time interval may influence IAcc scores. Although this effect is small, it awaits an explanation and can affect comparisons across studies.

These concerns are consistent with and add to a growing set of findings that question HCT research (Desmedt et al., 2018; Phillips et al., 1999; Ring & Brener, 1996, 2018; Ring et al., 2015; Windmann et al., 1999; Zamariola, Frost et al., 2019). In particular:

- Nearly all HCT studies to date have relied on instructions that allow for a substantial contribution of estimation processes to IAcc scores (Desmedt et al., 2018). This influence of estimation strategies is confirmed in studies that relied on extended debriefing (Zamariola, Frost et al., 2019) or that manipulated instructions (Desmedt et al., 2018). A recent collaborative research effort indicates the role of time estimation (i.e., counting seconds) among these strategies when using non-modified instructions - but, interestingly, *not* when using Desmedt et al. (2018) modified instructions (see Desmedt, Dekeyser, Corneille, & Luminet, 2019), which supports the higher validity of IAcc scores collected under modified instructions.
- IAcc scores show a surprising lack of association with theoretically associated constructs, such as alexithymia (for a recent meta-analysis, see Trevisan et al., 2019) and trait anxiety, even under modified instructions (Desmedt, Dekeyser et al., 2019).
- Efforts to replicate effects predicted on the basis of a sensible theorization, such as the emotional regulatory role of IAcc, have proved unsuccessful (Zamariola, Luminet, Mierop, & Corneille, 2019).

This *by no means* implies that correlational evidence published in the HCT literature should be dismissed. However, this suggests (i) that this evidence may be less robust than one may think, and, probably even more critical, (ii) that, when robust, this evidence should not necessarily be interpreted in terms of interoceptive accuracy. For instance, if IAcc scores capture cardiac condition, then health condition may be responsible for an observed correlation; if IAcc scores capture time estimation, then correlates of time estimation biases, such as depression, may be responsible for the observed association (Desmedt, Corneille et al., 2019); if IAcc scores capture response bias, confidence in one's judgment may be responsible for the correlation; if IAcc scores reflect estimation based on heart rate knowledge, health literacy or mere computation abilities may be responsible for the correlation.

Ainley and colleagues list an impressive number of correlational studies in their Commentary. However, the current analysis urges caution in the interpretation of the effects. As we explained here and had tried to explain in our original article, IAcc scores collected in HCT studies cannot be clearly interpreted. Drawing attention to these issues is not meant to do a disservice to interoception research. On the very contrary, it is meant to advance it. One reason why some effects may be less robust than expected, or why theoretically expected correlations may not be found, could be that HCT performance captures the contribution of a variety of processes, many of which are irrelevant to the tested hypothesis.

Advancing interoception measurement to tease apart the contribution of these various processes (e.g., by relying on modified instructions that reduce the contribution of estimation processes, or by relying on covariate analyses that control for confounded contributors; Desmedt, Corneille et al., 2019) can only *benefit* interoception research. In no way can it hurt it. Consistent with this constructive mindset, we now discuss recommendations for future research.

6.2. Recommendations for improving HCT and interoception research

The current exchange points to clear directions for improving HCT research on interoceptive accuracy. Below, we briefly discuss six recommendations.

1 Be cautious when analyzing ratio variables.

We thought the Commentary by Zimprich and colleagues was well-articulated, useful, and cautious in addressing "the intricacies of analyzing ratio variables" (p. 1). These authors explain that "interpreting findings of analyses based on ratio variables (…) may result in unintended inference and incorrect conclusions (e.g., Beaupre & Dunham, 1995)" (p. 1). This is a legitimate concern (but see Firebaugh & Gibbs, 1985; Kasarda & Nolan, 1979; MacMillan & Daft, 1980), although one that does not necessarily impact our conclusions. In particular, we explained that the correlation between HCT performance and heart rate creates interpretational issues even though this correlation may be mathematically expected.

The intricacies and unwarranted inferences associated with ratio variables analyses may actually be more concerning for HCT research itself. In the article by Kronmal (1993) referred to by Zimprich and colleagues, the author argues that ratio variables may lead to erroneous

interpretations also when related to a third variable (not just one of its components). Kronmal (1993) explains that: "The common practice of using ratios for either the dependent or the independent variable in regression analyses can lead to misleading inferences and rarely results in any gain" (Kronmal, 1993, p. 391). This points to the risk of faulty interpretations in HCT research. As two additional, although probably less critical concerns, Zimprich and colleagues point to bounded values (i.e., from 0 to 1) and heteroscedasticity in IAcc scores. Overall, then, mathematical issues coming with ratio analyses support our conclusion that one should be cautious in interpreting effects associated with IAcc scores.

2 Promote Open Science research.

We recommend that researchers in interoception make their data public, thereby allowing for cross-checks and, whenever necessary, correction. The recalculation by Ainley and colleagues of our quintile analysis is a perfect illustration of the value of transparent research practice. We also recommend that predictions and analyses are preregistered in HCT studies, and that exploratory findings are discussed as such until replication. Given important degrees of freedom that exist in, for instance, setting exclusion criteria, deciding what factors should enter a covariance analyses, or deciding on the general analytic strategy, non-preregistered research is at risk of feeding false positives. Finally, we recommend that authors do not hesitate to contact each other in case of uncertainties. Several methodological concerns raised by Ainley and colleagues could have been swiftly resolved through direct contact with the authors of the original article.

3 Set standard procedure criteria.

Many comments raised by Ainley and colleagues have to do with procedures. For instance, how should outliers be identified, what measurement procedure should be considered valid, what heart rate should be considered as representing rest state? These procedural considerations are essential, and we should have done a more precise job at reporting them. Unfortunately, clear guidelines are currently lacking. Ainley and colleagues offer in their Commentary a comprehensive set of recommendations for valid measurement, pointing to the validity of the sensors provided by the supplier and to controls for regular athletic activity and caffeine consumption hours before testing. Although we repeat that our findings were replicated across labs (our lab, GB's lab) and across recording devices (Polar Watch, Pulse Oximeter), procedures should not be neglected. We encourage our research community to reach a consensus on what set of criteria make for valid measurement of IAcc. Should some procedures be considered invalid, then (i) these procedures should not be used anymore, and (ii) evidence collected using these procedures should be interpreted cautiously.

4 Rely on modified instructions

Because one should care about the validity of measurement procedures, then modified HCT instructions should be used, and past studies that did not rely on modified instructions (nearly all of them) should be interpreted with great caution until effects are confirmed using these more valid instructions. As we discussed, usual HCT instructions make room for a substantial contamination of IAcc scores by estimation strategies (Desmedt et al., 2018). We are not talking here about factors that contribute to small fluctuations in measurement. Instead, we are talking about a simple change in instructions that reduces IAcc scores by 50 %. It is important to note, however, that even modified instructions still do not guarantee the task's validity (in particular, response bias and accuracy are still mixed under modified instructions). What the modified instructions guarantee, however, is that reliance on estimation strategies is substantially reduced.

5 Constrain the IAcc construct. Get rid of the "accuracy" terminology.

In all likelihood, only a small portion of IAcc scores has the potential to inform us about stable interindividual differences in interoceptive accuracy. This portion might grow when using modified instructions (Desmedt et al., 2018), or when considering covariates (Murphy, Brewer, Hobson, Catmur, & Bird, 2018). Of critical importance, however, this portion is mostly a function of the conceptualization of the IAcc construct itself.

We understand from the current exchange with Ainley et al. (2020) and Zimprich et al. (2020) that our core disagreement relates to divergent conceptualizations of the interoceptive accuracy construct. If interoceptive accuracy is to be defined as a mere rate of under-reporting, whatever set of processes contribute to participants' underestimation (including contributors that are physical rather than psychological, non-interoceptive such as for estimation strategies, or interoceptive but driven by response bias rather than cardiac sensitivity), then this portion is admittedly broad. However, it is fair to ask how theoretically sensible and useful this conceptualization is. In our view, it clearly is not as sensible and useful as it would be if conceptualization and measurement were further advanced.

Admittedly, one may hold different views on what interoceptive accuracy means. However, these views should then be precisely spelled out, and researchers should specify their preferred construct definition. If IAcc scores are meant to reflect mere undercounting in a theoretically wild way, then perhaps "underestimation" or "performance" instead of "accuracy" is the right label. "Interoceptive awareness" may also be used. However, IAcc scores would still mix instances of true awareness (i.e., hits) and pseudo-awareness (i.e., false alarms). This leads to our sixth and final recommendation, one we already stressed at the end of our original paper:

6 Our community should try to gain control over the interoceptive signal.

As we explained, in the absence of control over the internal signal, IAcc scores cannot be univocally interpreted. This is because performance on the task reflects a mix of accuracy and response bias, and one cannot tease apart hits and false alarms. A most challenging endeavor for future research on interoception is to try and design tasks allowing for a valid, convenient, and ethically acceptable, manipulation of the signal (e.g., Khalsa & Lapidus, 2016).

6.3. Conclusions

We are grateful that the two Commentaries gave us the opportunity to clarify these important points. We encourage scholars interested in advancing interoception research to consider and discuss within our research community the questions and recommendations discussed here and in the two Commentaries. We believe interoception research can only benefit from these exchanges.

Authors' note

We are grateful to Geoffrey Bird for comments provided on an earlier version of this manuscript, and to Adrien Mierop for running an independent test and preparing the r codes for the comparative quintile analysis. Olivier Desmedt is a FRESH grantee of the Fonds de la Recherche Scientifique – FNRS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ainley, V., Herbert, B. M., Tsakiris, M., Pollatos, O., & Schulz, A. (2020). Comment on "Zamariola et al. (2018), Interoceptive accuracy scores are problematic: Evidence from simple bivariate correlations"- The empirical data base, the conceptual reasoning and the analysis behind this statement are misconceived and do not support the authors' conclusions. Manuscript under review at *Biological Psychology*.

Beaupre, S. J., & Dunham, A. E. (1995). A comparison of ratio-based and covariance analyses of a nutritinal data set. *Functional Ecology,* 876–880.

Bestler, M., Schandry, R., Weitkunat, R., & Alt, E. (1990). Cardiodynamic determinants of heart perception. *Zeitschrift fur experimentelle und angewandte Psychologie, 37*(3), 361–377.

Bornemann, B., & Singer, T. (2017). Taking time to feel our body: Steady increases in heartbeat perception accuracy and decreases in alexithymia over 9 months of contemplative mental training. *Psychophysiology, 54*(3), 469–482. https://doi.org/10.1111/psyp.12790.

Dale, A., & Anderson, D. (1978). Information variables in voluntary control and classical conditioning of heart rate: Field dependence and heart-rate perception. *Perceptual and Motor Skills, 47*(1), 79–85.

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

Desmedt, O., Corneille, O., Luminet, O., Murphy, J., Bird, G., & Maurage (2019). Contribution of time estimation and knowledge to heartbeat counting task performance under original and adapted instructions. Manuscript under review at *Biological Psychology*.

Desmedt, O., Dekeyser, S., Corneille, O., & Luminet, O. (2019). *Investigating the link between interoceptive accuracy, Alexithymia and trait anxiety under adapted heartbeat counting task instructions.* Manuscript in preparation.

Durlik, C., & Tsakiris, M. (2015). Decreased interoceptive accuracy following social exclusion. *International Journal of Psychophysiology, 96*(1), 57–63. https://doi.org/10.1016/j.ijpsycho.2015.02.020.

Firebaugh, G., & Gibbs, J. P. (1985). User's guide to ratio variables. *American Sociological Review,* 713–722.

Georgiou, E., Matthias, E., Kobel, S., Kettner, S., Dreyhaupt, J., Steinacker, J. M., et al. (2015). Interaction of physical activity and interoception in children. *Frontiers in Psychology, 6*, 502. https://doi.org/10.3389/fpsyg.2015.00502.

Kasarda, J. D., & Nolan, P. D. (1979). Ratio measurement and theoretical inference in social research. *Social Forces, 58*(1), 212–227.

Kellen, D., & Klauer, K. C. (2018). *Elementary signal detection and threshold theory. Stevens' handbook of experimental psychology and cognitive neuroscience*, *5*, 1–39.

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

Krautwurst, S., Gerlach, A. L., & Witthöft, M. (2016). Interoception in pathological health anxiety. *Journal of Abnormal Psychology, 125*(8), 1179. https://doi.org/10.1037/abn0000210.

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 156*(3), 379–392. https://doi.org/10.2307/2983064.

MacMillan, A., & Daft, R. L. (1980). Relationships among ratio variables with common components: Fact or artifact? *Social Forces, 58*(4), 1109–1128.

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

Norton, J. W., & Corbett, J. J. (2000). Visual perceptual abnormalities: Hallucinations and illusions. *Seminars in Neurology, 20*(1), 0111–0122 Thieme Medical Publishers, NY, USA.

Petersen, S., & Ritz, T. (2010). Dependency of illness evaluation on the social comparison context: Findings with implicit measures of affective evaluation of asthma. *British Journal of Health Psychology, 15*(2), 401–416. https://doi.org/10.1348/135910709X466676.

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

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

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

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

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

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology, 18*(4), 483–488.

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

Trevisan, D. A., Altschuler, M. R., Bagdasarov, A., Carlos, C., Duan, S., Hamo, E., et al. (2019). A meta-analysis on the relationship between interoceptive awareness and alexithymia: Distinguishing interoceptive accuracy and sensibility. *Journal of Abnormal Psychology, 128*(8), 765–776. https://doi.org/10.1037/abn0000454.

Windmann, S., Schonecke, O. W., Fröhlig, G., & Maldener, G. (1999). Dissociating beliefs about heart rates and actual heart rates in patients with cardiac pacemakers. *Psychophysiology, 36*(3), 339–342. https://doi.org/10.1017/S0048577299980381.

Zamariola, G., Cardini, F., Mian, E., Serino, A., & Tsakiris, M. (2017). Can you feel the body that you see? On the relationship between interoceptive accuracy and body image. *Body Image, 20*, 130–136. https://doi.org/10.1016/j.bodyim.2017.01.005.

Zamariola, G., Frost, N., Van Oost, A., Corneille, O., & Luminet, O. (2019). Relationship between interoception and emotion regulation: New evidence from mixed methods. *Journal of Affective Disorders, 246*, 480–485.

Zamariola, G., Luminet, O., Mierop, A., & Corneille, O. (2019). Does it help to feel your body? Evidence is inconclusive that interoceptive accuracy and sensibility help cope with negative experiences. *Cognition and Emotion,* 1–12.

Zimprich, D., Nusser, L., & Pollatos, O. (2020). Are interoceptive accuracy scores from the heartbeat counting task problematic? A comment on Zamariola et al. (2018). Manuscript under review at *Biological Psychology*.