

Response to editor's comments on IMPS 2024 proceedings (paper number 7)

2025 June 4

We would like to thank the editor very much for their thorough review!

In the current revision, changes were made for clarity, in line with their comments.

Below are our responses in detail.

I found the article interesting, and it was well-written. I think the authors lean a little bit too heavily on their 2024 paper. I think this confuses the reader and buries the contribution of this work. I think quickly summarizing what L1P1 (does it stand for anything?) does can be done in the introduction and would be better than building a motivation for L1P1, reducing much of the first paragraph.

In the current revision, motivation for L1P1 (about training data and measurement models) has been removed from the 1st paragraph, in line with the editor's advice. Doing so also gives us more space.

In the current revision, the name "L1P1" is now briefly explained as a footnote in the 2nd paragraph.

In the current revision, in Section 1, some changes were made to make clearer the present manuscript's contribution.

- In the 2nd paragraph, L1P1 is explained more carefully, but still concisely: "L1P1 is a permutation test (Nyblom, 2015), done for each respondent, in three steps. First, an outlier statistic (e.g., Mahalanobis distance) is computed from the response pattern. Second, a null distribution of the outlier statistic is constructed by computing the same statistic from many random permutations of the same response pattern. Finally, the p-value is the observed statistic's quantile rank in the null distribution."
- In the 3rd paragraph, it is made explicit that sensitivity calibration comes from exchangeability of response patterns from those who are content non-responsive.
- In the 4th paragraph, it is made explicit that when there are multiple point-scales, exchangeability cannot be taken for granted, so neither can sensitivity calibration.
- In the last two paragraphs, it becomes clear that our goal is to keep sensitivity calibration even when the point-scale is not the same across items.

Similarly, there are these tidbits of information about the L1P1 throughout the text that is extraneous, leaving the reader confused, and perhaps require that the reader read the 2024 article. For example, 2.1 introduces notation and then adds information about L1P1's p-value, but we never use this notation again in the article. I found it a bit confusing and would recommend that the authors describe in words what sensitivity calibration is (and that the L1P1 does it). Similarly in 3.1, there is some information given and some information that says see the 2024 paper.

We address several issues based on this comment.

First, we believe that the notation introduced in Section 2.1 is either used in subsequent sections or enhances understanding of the proposed method.

- z_{ij} (the item response) appears also in Section 2.2. It is at this level the Binomial distribution is applied.
- z_i (the response pattern) is not formally invoked but is represented as rows in Table 1 and Table 2. It is at this level permutations occur.
- y_i (the true class label) and \hat{y}_i (the predicted class label) appear again in Section 2.4. They are the basis for calculating accuracy, specificity, and sensitivity.
- p_i is the L1P1 p-value itself, which is modified by the methods MCP, FIAF, SIAS, and PWP. Modified versions appear in Section 2.3 and Section 2.4.
- τ (the threshold) appears again in Section 2.4. P-values are converted to predicted class labels by applying the threshold.

Second, we agree that some information about L1P1 is extraneous to the present manuscript.

- In particular, part of L1P1 is the use of outlier statistics, but the present manuscript does nothing to change how they are computed. Thus, it suffices to tell the reader that there are outlier statistics, though the details can be left to the 2024 paper. This is stated at the end of section 1, both in the previous and current revision. However, in the current revision, Mahalanobis distance is mentioned as a concrete example of what statistic might be used, though the actual statistic used is substantially more complicated.
- We agree this sentence may have been confusing in Section 3: "Simulation constants were set in line with Ilagan and Falk (2024)." In fact, it is part of a paragraph labelled "simulation constants", and the rest of the paragraph details these constants, which were in line with the 2024 paper. In the current revision, that sentence has been rephrased to make it clearer that these constants are described in the same paragraph: "Simulation constants were set as follows, in line with Ilagan and Falk (2024)".

Finally, in Section 2.1, the current revision now explicitly defines sensitivity calibration to facilitate understanding: “For any algorithm, its sensitivity is the flag rate among CNR respondents; its specificity is the spare rate among non-CNR respondents; and its accuracy is the rate of correct predictions (Niessen et al., 2016).” It also explicitly states that if the exchangeability assumption (that every permutation is equally likely as what was observed) is in doubt, then so is sensitivity calibration: “If the CNR response pattern is not exchangeable, sensitivity calibration is not guaranteed.”

Sections 2.2 and 2.3

I think Section 2.2 was the most confusing part of the manuscript for me. I didn't understand how the binomial distribution would produce a response pattern of 4 to item 8. I understand that this is impossible, but I did not make the connection between the L1P1 and calibrating sensitivity. I think it would be helpful for the reader to understand what is "fed" to the L1P1.

The current revision makes the following changes for clarity.

- In section 2.1, the connection between L1P1 and sensitivity calibration is made explicit.
- In section 2.1, it is explicitly stated that L1P1 is predicting using only the response pattern data z_1, z_2, \dots, z_n so that is what is "fed" into L1P1.
- In section 2.2, it is made clearer that using the binomial distribution is something that the CNR respondent is doing. Furthermore, the reader is asked to imagine the first row of Table 1 as the observed response pattern, while the other rows are permutations of it. Then from Table 1, it can be seen that the way the CNR responded cannot be captured by the exchangeability assumption L1P1 is using. Thus, sensitivity calibration is not implied.

It might help the editor to walk through the example in Table 1. We suppose a CNR respondent. Because Item 1 is 4PS, the respondent draws the item response from Binomial(3, 0.5). The same binomial distribution is used for Item 2 to Item 6, as they are all 4PS. But for Item 7 and Item 8, the respondent draws from Binomial(6, 0.5) instead, as they are 7PS. A response pattern resulting from this process is the first row of Table 1. The probabilities shown in Table 1 are calculated assuming this process. Because permutations of this process do not have the same probability, such a response process is not exchangeable.

I found the jumping between the toy example and the DASS+TIPI examples confusing. Tables 1 and 2 show the 4 PS and 2PS scales, but Figure 1 shows 4PS and 7PS scales. I would recommend doing one or the other, but not both when explaining this to the reader.

To reduce confusion, we changed the 8-item example in the revised manuscript. The point-scales are now 4PS and 7PS, which is analogous to the DASS+TIPI.

We understand that jumping between a small (8 items) and a large example (DASS+TIPI, 52 items) is not ideal. But there are constraints that force us to do so.

- On one hand, we do not advise permutation tests be done on short inventories. In fact, Section 2.2 explicitly warns against doing so in the original version, which is retained. Thus, in Figure 1, we show the algorithms for the 52-item DASS+TIPI.
- On the other hand, to illustrate how permutation works, it is necessary to show examples where the entire response pattern can be seen on the page. This is impractical for the 52-item DASS+TIPI. Thus, Table 1 and Table 2 show permutations for an 8-item example.

Altogether, we have to compromise by showing permutations for an 8-item inventory while using DASS+TIPI everywhere else. Word limits prevent us from explicating these considerations in the text.

MCP, FIAF, SIAS were easy to understand, but I still do not understand what PWP is doing. Each set of items are permuted among themselves, but I don't understand what it means for the separate permutations to be "concatenated back into a single response pattern" that can be fed into L1P1. How does permuting within each PS result in a single response? I would recommend some more clarity here.

In the current revision, PWP is now better foreshadowed.

- In section 2.1, in the first paragraph, we now explicitly mention that the 2024 paper has the constraint $c_1=c_2=\dots=c_n$ while the present article does not.
- In section 2.2, the formal definition of the multiple point-scale null hypothesis now talks about the point-scales, "unique values" of $\{c_1, c_2, \dots, c_m\}$ which PWP loops over. It is also stated that $\{4, 7\}$ are the unique values for Table 1.
- In section 2.2, we add a new paragraph that talks about CNR example generation. "Calibrating sensitivity comes down to producing CNR response pattern examples that are in line with the null hypothesis. Permuting the observed response pattern produces a CNR example in line with exchangeability as the null hypothesis, which is exactly what base L1P1 does. But for the multiple point-scale null hypothesis, doing the same is not in line, as seen in Table 1."
- In section 2.5, just before describing PWP, we add a new paragraph that emphasizes its difference with other algorithms. "So far, the algorithms turn out to be applications of base L1P1. MCP simply changes the input to base L1P1; while FIAF and SIAS do multiple applications of L1P1, then combine the multiple p-values into a single final output. In all these algorithms, CNR examples are generated by simply permuting the entire input response pattern. In contrast,

our recommended algorithm, PWP, changes how CNR examples are generated from the input response pattern.”

In the current revision, the description of PWP is parallel to how base L1P1 is described in section 1. The change is only in Step 2, keeping Step 1 and Step 3 the same.

The paragraph after these three steps walks through Table 2. Take the first row as the observed response pattern. The other rows are produced by permuting only within 4PS (Items 1 to 6) and only within 7PS (Items 7 to 8). The language of “concatenating back into a single response pattern” has been dropped.

Section 4.1

It is very easy to miss what PIE stands for - I had to do a search, just a fyi that it might be easier to rename this.

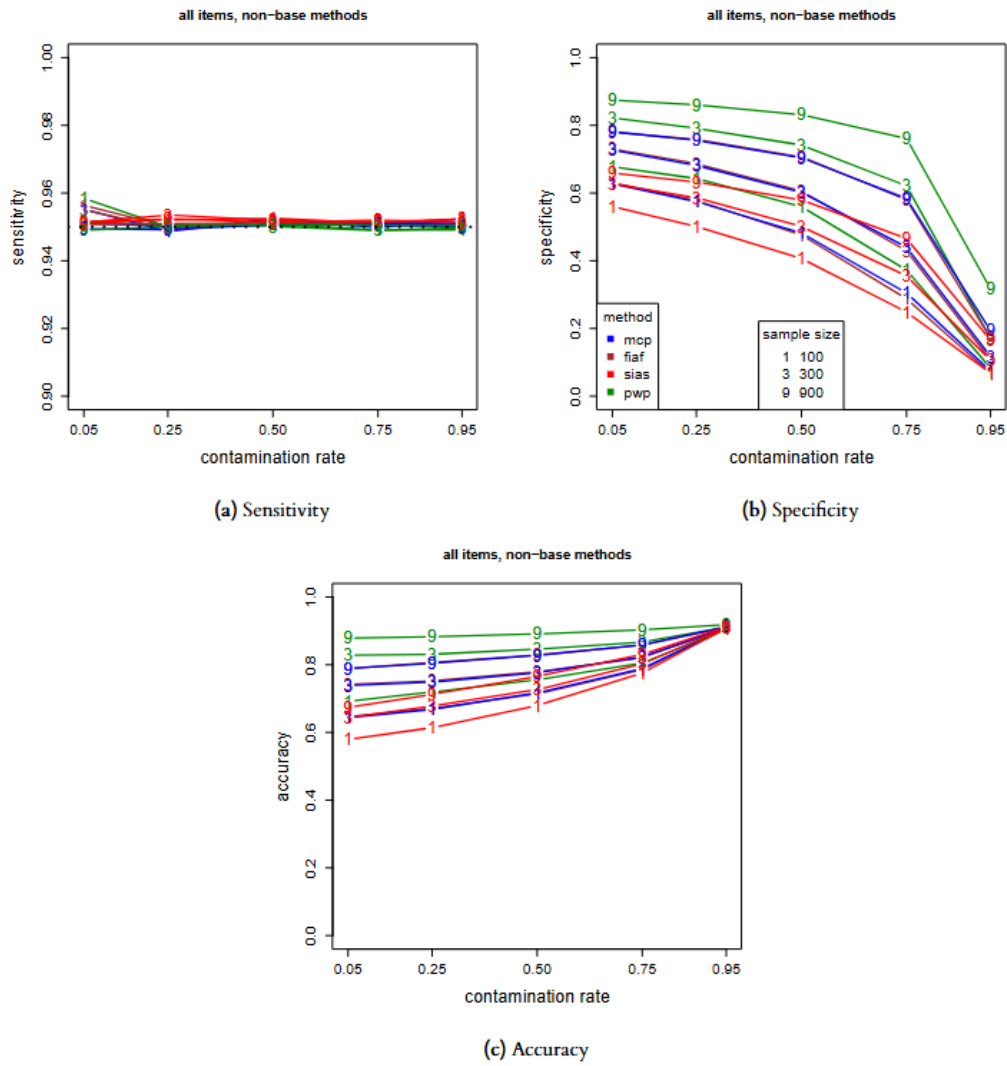
In the current revision, PIE has been renamed to "base". In Section 2.2, it is made explicit that “base L1P1” refers to L1P1 “without modifications for varying number of response categories”.

Figure 3 too challenging to read. It is extremely difficult to visually make sense of changing colors and changing numbers at the same time, especially given the axes. Figures almost always have the outcome of interest on the Y-axis, I would recommend rotating the plot so that the outcome is vertical and plotting something like the contamination rate on the x-axis. Line graphs of different darkness or different colors (showing the sample size) would be much easier to understand. Same comments about Figures 4 and 5.

We acknowledge that the plots are dense, as they account for four simulation factors (contamination rate, sample size, method, items) together. However, they are carefully: colors are contamination rates (e.g. red is 95% contamination), numbers are sample sizes (e.g. “3” is n=300), and boxes are item-method pairs. Colors and numbers have proper legends.

However, to reduce confusion, we adjusted the heights of the points so that scenarios of the same contamination rate (i.e. same color) are of the same height. The result can be read like a Cleveland dot-plot, which typically has the outcome on the x-axis.

We tried to create line plots where the horizontal axis was contamination rate and the vertical axis was the outcome measure (sensitivity, specificity, or accuracy). However, such line plots create some issues. Notice that for Figure 4a (sensitivity for non-base methods), proper calibration meant that most of the lines would have overlapped at the correct rate of 95%. At high contamination rates, there is a lot of overlap as well for specificity (Figure 5b) and accuracy (Figure 5c).



In panel (a), the dashed vertical line marks 95% sensitivity.
 base = base L1P1; fiaf = flag if all flag; sias = spare if all spare; pwp = permute within point-scale.

Figure 4. All-items scenarios: For the four algorithms for multiple point-scales, mean across replicates for four metrics: (a) sensitivity; (b) specificity; and (c) accuracy.

We find the Cleveland-style plots to maintain readability even when the various conditions have similar outcome measures.