Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Unsupervised detection of random responding for Likert-type inventories with varying numbers of response categories

Michael John Ilagan and Carl F. Falk*

Department of Psychology, McGill University, Montreal, Quebec, Canada
*Corresponding author. Email: carl.falk@mcgill.ca

Abstract

Likert-type inventories administered online risk "random" responding, such as by bots. To safeguard data quality, we consider unsupervised classification of random vs. non-random responders. Previous work proposed L1P1, an algorithm based on a permutation test with bias-corrected outlier statistics. L1P1 successfully calibrates sensitivity, assuming that for random responders, exchangeability holds for the entire response vector. However, when the items do not have the same "point-scale" (i.e., number of response categories), the same assumption is inapplicable. To extend the L1P1 classifier to inventories with multiple point-scales, we propose generating the null distribution by permuting only within subvectors of the same point-scale (PWP), otherwise following the original L1P1. Such a proposal is in contrast to doing a permutation test per point-scale then combining multiple p-values into a final predicted class. In a simulation study, the main findings were twofold. First, the proposed approach generally outperformed alternatives considered in terms of sensitivity calibration and accuracy. Second, p-values from point-scales with few items failed to calibrate sensitivity to the nominal 95% rate. PWP is implemented in the R package `detranli`, available on Github.

Keywords: Likert, random responding, careless responding, permutation test, classification, unsupervised learning

1. Introduction

While collecting data via administering Likert-type questionnaires is common, there is risk that some participants will be content-nonresponsive (CNR)—that is, they provide responses without regard to the items' content (Meade & Craig, 2012; Nichols et al., 1989). The source of CNR could be randomly responding humans or computer programs ("bots") that masquerade as humans (e.g., Perkel, 2020; Storozuk et al., 2020).

To detect random responding, Ilagan and Falk (2024) proposed the L1P1 algorithm.[1] 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.

L1P1 assumes that CNR response patterns are *exchangeable*—that every permutation is equally probable as the observed pattern. As found in Ilagan and Falk (2024), when this assumption is true,

[1] L1P1 is so named due to having leave-one-out cross validation ("L1") and using permutations as CNR examples ("P1"), in contrast to other methods presented in Ilagan and Falk (2024). These other methods are not in the scope of the present article.

the p-value calibrates sensitivity—for instance, there is a 5% probability that the p-value is below 0.05.

Of interest in the present article, a limitation of L1P1 is that the exchangeability assumption makes sense only if all items of the inventory have the same number of response categories (e.g., all items are on a 5-point-scale). But in some studies (e.g., Smyth et al., 2024), items may have "multiple point-scales", so to speak. To illustrate, the researcher may have administered two inventories: the Depression and Anxiety Stress scales (DASS; Lovibond & Lovibond, 1995), which has 42 items on a 4-point-scale (4PS); and the Ten Item Personality Inventory (TIPI; Gosling et al., 2003), which has 10 items on a 7-point-scale (7PS). When exchangeability does not hold for CNR respondents, sensitivity calibration is no longer guaranteed.

In the present article, we extend L1P1 (Ilagan & Falk, 2024) to inventories with varying number of response categories. In particular, we do the following: we define a null hypothesis characterizing CNR under multiple point-scales; we consider several algorithms, in line with this null hypothesis, that attempt to maintain nominal sensitivity calibration; and finally, we demonstrate these algorithms' properties in a simulation study. The main result is that one algorithm we propose, called *permutation-within-point-scales*, attains sensitivity calibration while dominating its alternatives in terms of specificity. ("Sensitivity" and "specificity" are defined in Section 2.1.)

To be clear, the present article is concerned with how to generate null distributions in response pattern space so that CNR can be dealt with in inventories with multiple point-scales, preserving sensitivity calibration. Details about the test statistic (i.e., the function that converts the response pattern to a p-value) are in Ilagan and Falk (2024) but are not necessary for the purposes of the present article.

2. Interventions

2.1 Notation and data format

Let z_{ij} be the observed response of respondent $i = 1, \dots, n$ on item $j = 1, \dots, m$. Let $z_i = \begin{bmatrix} z_{i1} & z_{i2} & \dots & z_{im} \end{bmatrix}^\top$ be the entire response pattern. For the j-th item, follow the convention that the ordinal response categories are $1, 2, \dots, c_j$. For the i-th respondent, the true class labels are denoted $y_i = 1$ for CNR and $y_i = 0$ for non-CNR. The constraint $c_1 = c_2 = \dots = c_m$ was assumed in Ilagan and Falk (2024), but it is not assumed in the present article.

Our task is to predict the true class, using only the response pattern data $z_1, z_2, \dots, z_n$. The respondent is *flagged* if $\hat{y}_i = 1$ and *spared* if $\hat{y}_i = 0$. 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). An algorithm is said to be sensitivity-calibrated if its sensitivity matches the nominal rate (e.g., true sensitivity and nominal sensitivity are both 95%).

Assuming exchangeability of the CNR-class response patterns, L1P1 (Ilagan & Falk, 2024) produces the p-value p_i for each respondent $i = 1, \dots, n$. To predict classes with sensitivity calibration, have $\hat{y}_i = \mathbb{I}\{p_i \geq \tau\}$ where $1 - \tau$ is the nominal sensitivity (e.g., $\tau = 0.05$ for 95% nominal sensitivity) and $\mathbb{I}$ denotes the indicator function. If the CNR response pattern is not exchangeable, sensitivity calibration is not guaranteed.

2.2 The CNR null hypothesis for multiple point-scales

To motivate the null hypothesis for multiple point-scales, we introduce a toy example. Suppose the inventory is a total of eight items—six items on a 4-point-scale (4PS), and two items on a 7-point-scale (7PS). Furthermore, suppose a CNR respondent who answers items independently, drawing each response from a fair-probability binomial distribution depending on the number of response categories (Hong et al., 2020). Precisely, $z_{ij} - 1 \sim \text{Binomial}(c_j - 1, \frac{1}{2})$. In Table 1, suppose the first row is the resulting observed response pattern, and the remaining rows are some permutations thereof. Clearly, these permutations are not equiprobable. In fact, one of them is impossible, as the response category "5" appears on a 4PS item. Thus, the response pattern is not exchangeable, and L1P1 cannot guarantee calibrating sensitivity.

Table 1. Several permutations of the same response pattern and their log probabilities.

4-point-scale items						7-point-scale items		
Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Log probability
1	2	3	4	1	2	3	5	−12.08
4	3	2	1	1	2	3	5	−12.08
1	1	2	2	3	3	4	5	−10.70
1	2	3	4	5	1	2	3	$-\infty$

To extend L1P1 to mulitple point-scales with sensitivity calibration, we propose a new null hypothesis. In the toy example, note that: the 4PS subvector is exchangeable; the 7PS subvector is exchangeable; and the subvectors are independent. Accordingly, we propose the following broader null hypothesis for CNR.

Multiple point-scale null hypothesis for CNR. Within each unique value of $\{c_1, c_2, \ldots, c_m\}$ the relevant subvector of items is exchangeable. The subvectors are mutually independent.

In Table 1, the two unique values are $\{4, 7\}$. Note that when there is only one point-scale (i.e., $c_1 = c_2 = \ldots = c_m$), this hypothesis reduces to exchangeability of the entire response pattern, so base L1P1 (without modifications for varying number of response categories) is appropriate.

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.

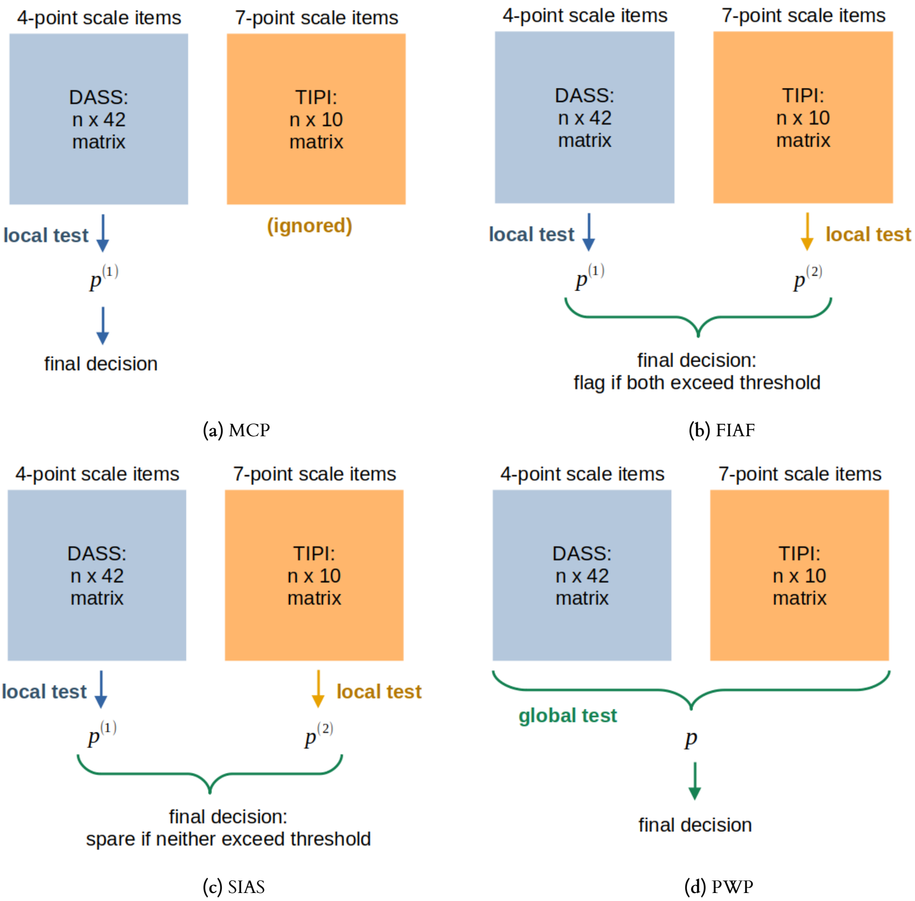
For the multiple point-scale CNR null hypothesis, sensitivity can be calibrated in many ways. In the present article, we consider four algorithms extending L1P1: use only the Most Common Point-scale (MCP); test point-scale-wise, Flag If All Flag (FIAF); test point-scale-wise, Spare If All Spare (SIAS); and test globally, permuting within point-scales (PWP). In fact, we favor the PWP for reasons that will become clear.

In what follows, we walk through each of the four algorithms. For concreteness, we consider the scenario where the respondents answer the DASS as well as the TIPI (henceforth "DASS+TIPI"). Each algorithm is illustrated in Figure 1. Keep in mind that permutation tests based on more items tend to have better specificity (Falk et al., in press; Ilagan & Falk, 2024). Intuitively, with more items, there is more information to tell apart the exchangeable CNR class from the non-exchangeable CNR class. Also keep in mind that sensitivity calibration cannot be guaranteed when there are too few items to permute. For instance, with four items, there are only 4! = 24 possible permutations, whereas Ilagan and Falk (2024) generated 200 random permutations per respondent. Note that when the point-scale is the same for the entire inventory, all algorithms reduce to base L1P1.

2.3 Use only the Most Common Point-scale (MCP)

The simplest among the algorithms, MCP, is as follows.

1. Use only items from the most common point-scale in the inventory, ignoring the rest.
2. On the chosen items, execute base L1P1, getting a single p-value. Thus, participant $i = 1, \ldots, n$ is associated with $p_i^{(1)}$.



(a) MCP

(b) FIAF

(c) SIAS

(d) PWP

Figure 1. Given the Depression and Anxiety Stress scales (DASS) and the Ten Item Personality Inventory (TIPI), four algorithms that attempt to calibrate sensitivity under the multiple point-scale CNR null hypothesis: (a) use only the Most Common Point-scale (MCP); (b) test point-scale-wise, Flag If All Flag (FIAF); (c) test point-scale-wise, Spare If All Spare (SIAS); (d) Test globally, permuting within point-scales (PWP).

In the case of DASS+TIPI, 4PS is more common, so L1P1 would be run on only the 4PS subvector. See Figure 1a for an illustration. Note that there is only one p-value, so the superscript appears superfluous; but the notation is consistent with other algorithms where each point-scale is associated with its own p-value.

MCP is straightforward in that it changes only the input to L1P1 rather than the algorithm itself. It avoids the permutation being based on a small number of items (e.g., only 10 items with 7PS in TIPI). However, the limitation is that information from those items is ignored.

2.4 *Test point-scale-wise, Flag If All Flag (FIAF) and Spare If All Spare (SIAS)*

FIAF and SIAS are the same up to the last step. In both, permutation tests are done per point-scale, then the point-scale-wise p-values are combined into a final decision, using their independence under the null, as follows.

1. Split the data by point-scale, indexed by $k = 1, \dots, K$.
2. For each split k, run L1P1, yielding a point-scale-wise p-value. Thus participant $i = 1, \dots, n$ is

associated with p-values $\begin{bmatrix} p_i^{(1)} & \dots & p_i^{(K)} \end{bmatrix}^\top$.

3. To arrive at a final decision for respondent $i = 1, \dots, n$, the FIAF rule is

$$\hat{y}_i = \min\left\{\mathbb{I}\{p_i^{(k)} \geq \tau\} : k = 1, \dots, K\right\}$$

while the SIAS rule is

$$\hat{y}_i = \max\left\{\mathbb{I}\{p_i^{(k)} \geq \tau\} : k = 1, \dots, K\right\}$$

where the threshold τ differs between the two.

In the case of DASS+TIPI, $K = 2$, as there is just 7PS and 4PS. See Figure 1b (FIAF) and Figure 1c (SIAS) for an illustration.

The threshold can be set as a function of K and the nominal sensitivity rate $1 - \alpha$. Under the null hypothesis, all K p-values are independent. Thus, for FIAF, $\tau = 1 - (1 - \alpha)^{1/K}$; and for SIAS, $\tau = \alpha^{1/K}$. For an illustration where $K = 2$ and $1 - \alpha = 0.95$, see Figure 2a (FIAF) and Figure 2b (SIAS).

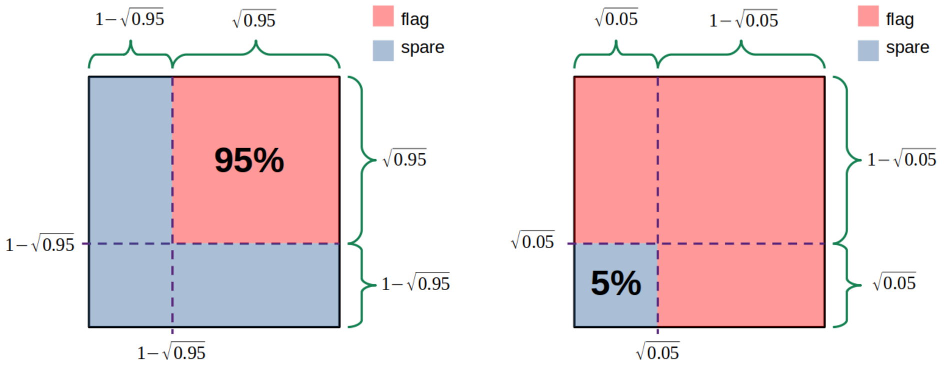
In comparison to MCP, the advantage of FIAF/SIAS is that it uses all items. However, the disadvantage is that some p-values may be based on few items, which MCP is suited to avoid. Note that in FIAF/SIAS, even if all items are used, covariances of items across point-scales are ignored.

2.5 *Test globally, permuting within point-scales (PWP)*

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.

Avoiding drawbacks of the other algorithms, PWP is as follows.

1. The outlier statistic is computed from the entire response pattern, as with base L1P1.
2. A null distribution of the outlier statistic is constructed by computing the same statistic from many CNR examples of the same response pattern. But unlike base L1P1, CNR examples are generated by permuting items only within each unique value of $\{c_1, c_2, \dots, c_m\}$.
3. The p-value is the observed statistic's quantile rank in the null distribution, as with base L1P1.



(a) FIAF

(b) SIAS

Figure 2. Decision boundaries with 95% sensitivity calibration for $K = 2$ point-scales tested separately: (a) Flag If All Flag (FIAF); and (b) Spare If All Spare (SIAS).

See Figure 1d for an illustration on DASS+TIPI. In Table 2, suppose the first response pattern is the one observed, and the remaining rows are several CNR examples generated by PWP. Notice that Items 1–6 are permuted among themselves, as they are all 4PS; Items 7–8 are permuted among themselves, as they are all 7PS; both permutations take place within a single response pattern.

Table 2. A response pattern and three CNR examples under Permuting Within Point-scale (PWP)

4-point-scale items						7-point-scale items	
Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	2	3	4	1	2	3	5
1	1	2	2	3	4	3	5
1	2	3	4	1	2	5	3
1	1	2	2	3	4	5	3

Like FIAF and SIAS, PWP uses all items. However, PWP has advantages. PWP does not risk a p-value being based on too few items, as long as there are enough items in total; PWP incorporates covariances between items that have different point-scales; and PWP avoids an arbitrary scheme of point-scale-wise thresholds, as in Figure 2. Thus, while all the algorithms considered attain sensitivity calibration with enough items, PWP is anticipated to have better specificity.

3. Simulation study

We conducted a simulation study to evaluate the four algorithms. We were particularly interested in verifying that all four algorithms calibrate sensitivity and that PWP dominates them in specificity. In each replicate, we generated a sample having some CNR respondents, applied the algorithms, then calculated three outcome measures: sensitivity, specificity, and accuracy. In addition to the four sensitivity-calibrated algorithms, we included a naive application of L1P1 (Ilagan & Falk, 2024) as a baseline, to demonstrate the ill effects of neglecting multiple point-scales.

Inventories and non-CNR data. Inventories used were DASS+TIPI. To generate non-CNR response patterns, we sampled from the DASS+TIPI dataset from the Open Psychometrics Project (n.d.), which had $N = 39775$ rows. In DASS (42 items, 4PS), there were no missing values. In TIPI (10 items, 7PS), about 2% of the rows had exactly one item missing, while about 1.5% of the rows had more than one item missing.

Simulation design. Denote as n_1 the number of true CNR response patterns in the sample. We varied four factors:

- The total sample size, $n \in \{100, 300, 900\}$;
- The contamination rate, $\frac{n_1}{n} \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$;
- The CNR distribution per item, either a uniform distribution or a fair-coin binomial distribution; and
- Which items were used, either all items (i.e., DASS 42 items 4PS + TIPI 10 items 7PS) or only the even-numbered items (i.e., DASS 21 items 4PS + TIPI 5 items 7PS).

Varying the items used was to demonstrate the effect of having fewer items to work with. In each cell, there were 500 replicates.

Simulation constants. Simulation constants were set as follows, in line with Ilagan and Falk (2024). Toward computing p-values, Mahalanobis distance and person-total correlation (Curran, 2016; Zijlstra et al., 2011) were used as intermediate outlier statistics, which were then combined to the final outlier statistic proposed in Ilagan and Falk (2024). For each permutation test, 200 permutations were generated. Nominal sensitivity was $1 - \alpha = 0.95$ in all scenarios. For response patterns with missing values, the algorithm was applied only to the nonmissing items. If the permutation test could not be computed for any reason (e.g., only one nonmissing item), the respondent was flagged by default.

Software. The entire simulation study was conducted in R version 4.4.0 (R Core Team, 2021). For permutation testing, the package `detranli`, available on Github, (Ilagan & Falk, 2025) was used.[2] This package implements the original L1P1 (Ilagan & Falk, 2024) as well as PWP. Custom functions were written in R to implement MCP, FIAF, and SIAS. For parallel processing, packages `future` (Bengtsson, 2021) and `furrr` (Vaughan & Dancho, 2021) were used.

4. Results and discussion

4.1 Base L1P1

We start with results for base L1P1. Results are summarized by the mean across the 500 replicates in Figure 3. Results were similar across CNR distributions, so results only for uniform CNR distribution are presented.

- As expected, sensitivity was far below the nominal rate of 95% (Figure 3a).
- There was more specificity for all-items scenarios (Figure 3b).
- There was a wide range for accuracy (Figure 3c): when contamination rate was low, accuracy was very high; but accuracy was very low when contamination rate was high. Longer inventories and larger samples mostly had less accuracy, which is undesirable.

With very low sensitivity and very high specificity, we can deduce that the original L1P1 was barely flagging anyone—which for low-contamination scenarios, was accurate by luck.

4.2 Algorithms for multiple point-scales

We then look at the results for the four proposed algorithms—MCP, FIAF, SIAS, and PWP. We first go over results for all items, then see the effect of having fewer items. Results were similar across CNR distributions, so results only for uniform CNR distribution are presented.

All items. Results are summarized by the mean across the 500 replicates in Figure 4.

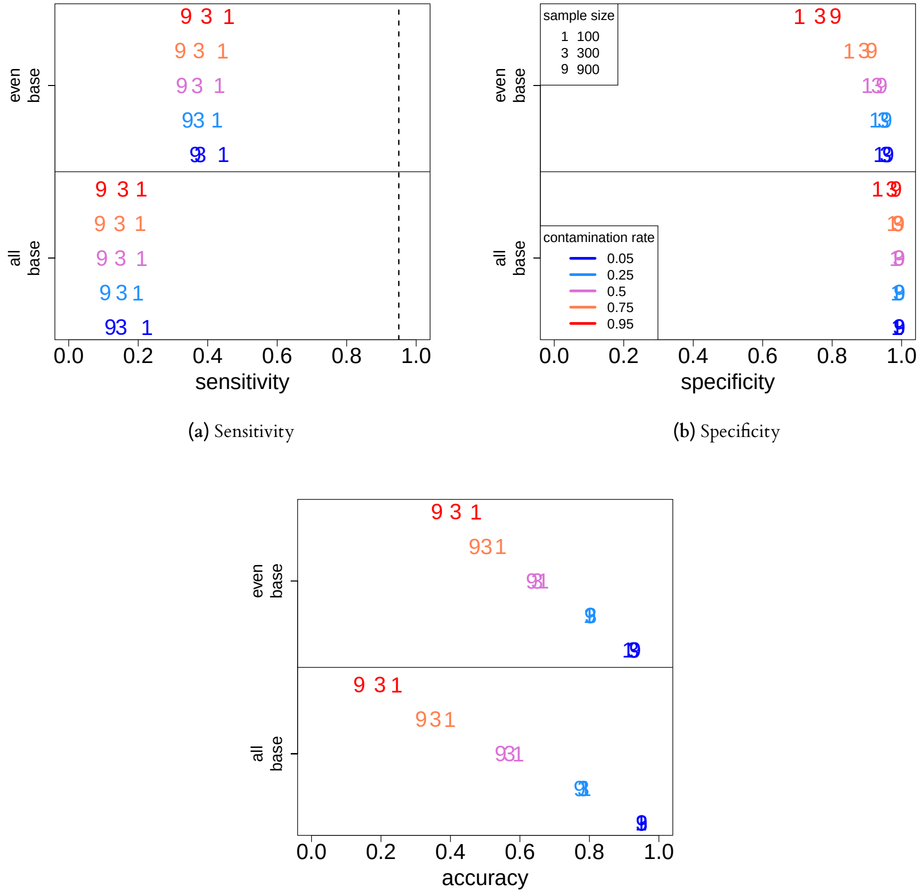
- As expected, MCP, FIAF, SIAS, and PWP all calibrated sensitivity (Figure 4a). The farthest sensitivity was off by no more than one percentage point.
- For specificity (Figure 4b) PWP was better than other sensitivity-calibrated methods, as expected. Interestingly, MCP had better specificity than did SIAS, which indicates calculating a p-value particular to the 10-item TIPI worsened performance.
- In line with Ilagan and Falk (2024), scenarios with higher contamination had higher accuracy (Figure 4c). As expected, PWP was more accurate than other sensitivity-calibrated algorithms.

Even-numbered items only. Results are summarized by the mean across the 500 replicates in Figure 5.

- With fewer items, it becomes clear that FIAF and SIAS have more sensitivity (Figure 5a) than the nominal rate.
- The same trends as in all-items scenarios can be seen when comparing algorithms on specificity (Figure 5b), but specificity is worse across the board with fewer items.
- The same trends as in all-items scenarios can be seen when comparing algorithms on accuracy (Figure 5c), but accuracies are worse across the board with fewer items.

Overall, besides our expectations of the four algorithms' properties being supported, truisms from Ilagan and Falk (2024) were also reconfirmed. A miscalibrated algorithm can have better accuracy in some scenarios just by luck—but researchers must beware, as larger datasets (more respondents

[2] https://github.com/michaeljohnilagan/detranli



(a) Sensitivity

(b) Specificity

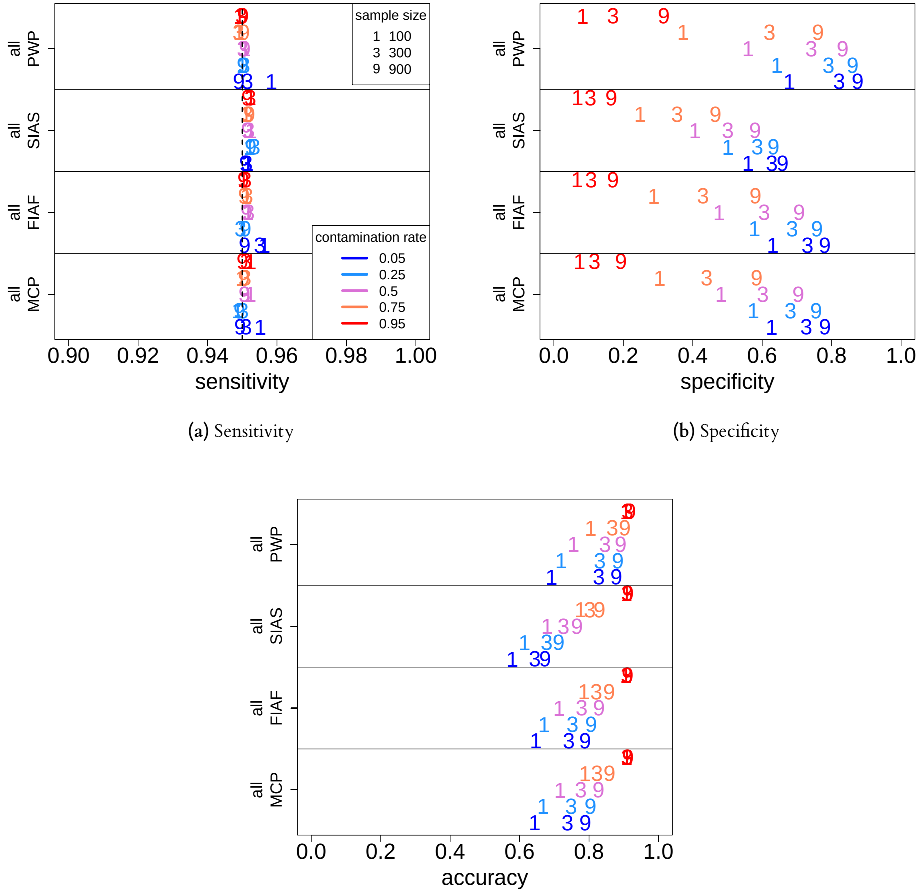
(c) Accuracy

In panel (a), the dashed vertical line marks 95% sensitivity.
base = base L1P1; all = all-items scenarios; even = even-items scenarios.

Figure 3. All-items and even-items scenarios: For base L1P1, mean across replicates for four metrics: (a) sensitivity; (b) specificity; and (c) accuracy.

or more items) do not always improve performance. There is a trade-off between sensitivity and specificity. Once sensitivity is calibrated, having a larger sample improves accuracy. And finally, under some scenarios, even the best accuracy may be low.

Funding Statement We acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC), (funding reference number RGPIN-2018-05357 and DGECR-2018-00083), and the Fonds de recherche du Québec–Nature et technologies (2022-PR-298903 and 2023–2024-B2X-330469). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence RGPIN-2018-05357 et DGECR-2018-00083] et les Fonds de recherche du Québec–Nature et technologies (2022-PR-298903 et 2023–2024-B2X-330469).



(a) Sensitivity

(b) Specificity

(c) Accuracy

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.

Competing Interests None

References

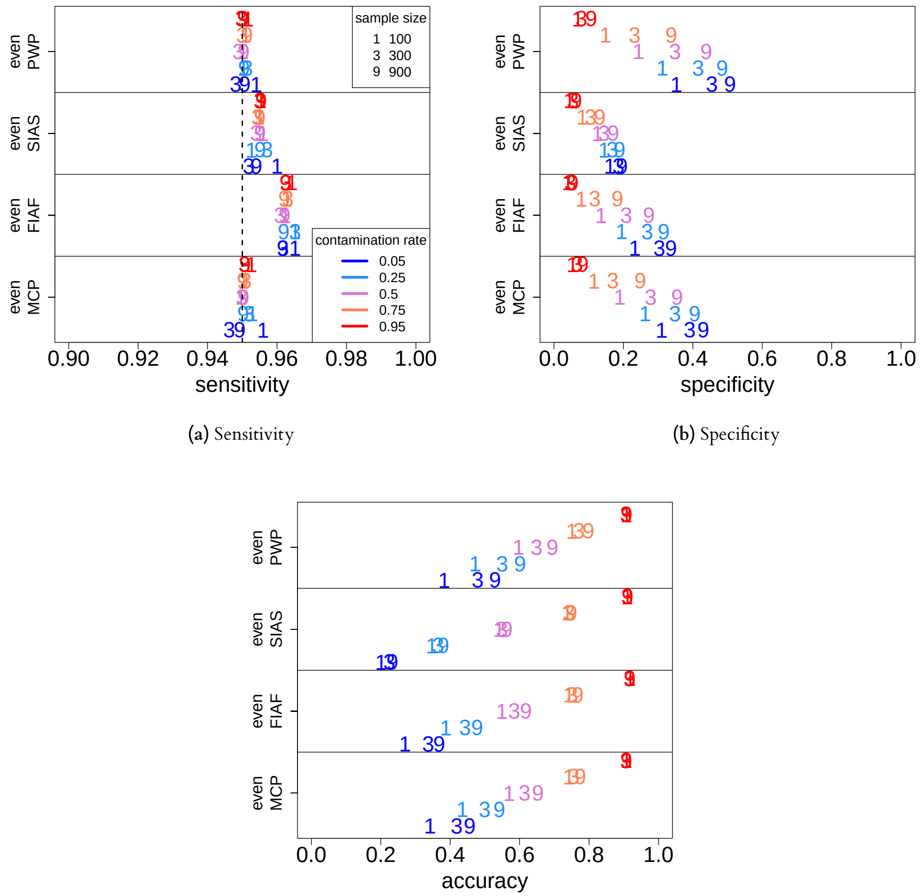
Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in R using futures [R package version 1.21.0]. https://journal.r-project.org/archive/2021/RJ-2021-048/index.html

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Falk, C. F., Huang, A., & Ilagan, M. J. (in press). Unsupervised [randomly responding] survey bot detection: In search of high classification accuracy. *Psychological Methods*.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality*, *37*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Hong, M. R., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312–345. https://doi.org/10.1177/0013164419865316



(a) Sensitivity

(b) Specificity

(c) Accuracy

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

Ilagan, M. J., & Falk, C. F. (2024). Model-agnostic unsupervised detection of bots in a likert type questionnaire. *Behavior Research Methods*, *56*, 5068–5085. https://doi.org/10.3758/s13428-023-02246-7

Ilagan, M. J., & Falk, C. F. (2025). Detranli [https://github.com/michaeljohnilagan/detranli].

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the depression anxiety stress scales*. Sydney, Psychology Foundation.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2). https://doi.org/10.1002/1097-4679(198903)45:2\%3C239::AID-JCLP2270450210\%3E3.0.CO;2-1

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Nyblom, J. (2015). Permutation tests in linear regression. In K. Nordhausen & S. Taskinen (Eds.), *Modern Nonparametric, Robust and Multivariate Methods*. Springer. https://doi.org/10.1007/978-3-319-22404-6_5

Open Psychometrics Project. (n.d.). Raw data from online personality tests. https://openpsychometrics.org

Perkel, J. M. (2020). Mischief-making bots attacked my scientific survey. *Nature*, *579*, 461. https://doi.org/10.1038/d41586-020-00768-0

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Smyth, A. P. J., Juneau, C., Hong, S., Ilagan, M. J., & Knäuper, B. (2024). Facing obstacles with equanimity: Trait equanimity attenuates the positive relations between values obstruction and symptoms of depression, anxiety, and stress. *Mindfulness*, *15*, 945–957. https://doi.org/10.1007/s12671-024-02338-1

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472–481. https://doi.org/10.20982/tqmp.16.5.p472

Vaughan, D., & Dancho, M. (2021). *furrr: Apply mapping functions in parallel using futures* [R package version 0.2.2]. https://CRAN.R-project.org/package=furrr

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, *36*(2). https://doi.org/10.3102/1076998610366263