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

Bot detection: Simulations and application in people-centered health measurement surveys with missing data

Carl F. Falk,[*†] Michael John Ilagan,[†] Mathilde G. E. Verdam,[‡] and Richard Sawatzky[¶]

†Department of Psychology, McGill University, Montreal, Quebec, Canada
‡Department of Methodology & Statistics, Institute of Psychology, Leiden University, Leiden, Netherlands
¶School of Nursing, Trinity Western University, Langley, British Columbia, Canada
Centre for Advancing Health Outcomes, Providence Health Care Research Institute, Vancouver, British Columbia, Canada
Institute of Health and Care Sciences, and Centre for Person-Centered Care (GPCC), Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[*]Corresponding author. Email: carl.falk@mcgill.ca

Abstract

In the context of improving the measurement of pain and emotional well-being among diverse populations, we sought to detect random responders or survey bots that are not responsive to item content. We adapted the L1P1 algorithm by Ilagan and Falk (2024), which uses a permutation test and outlier statistics to compute a p-value and do classification under the null that the response vector is exchangeable. As the response options for the Likert-type items could yield missing data, simulations evaluated two variants of outlier statistic computations that used the expectation-maximization algorithm: one in which means and covariances were pre-computed and re-used for all rows, and another in which a leave-one-out approach was used. Results indicated that the L1P1 algorithm works as expected, but a leave-one-out strategy works best, and respondents with few completed items are flagged at higher rates due to loss of specificity. Based on simulations, we then performed classification for an empirical dataset (N = 11,197) with 76 Likert-type items. Flagging rates were similarly higher for respondents with fewer completed items, but otherwise low. We therefore expect that random responders would likely not have strong influence on subsequent analyses for this measurement project.

Keywords: Patient-reported outcome measures; Person-centered measurement, Survey bots, Careless responding, Machine Learning, Missing Data, Permutation test

The equitable people-centered health measurement (EPHM) project aims to improve the measurement of pain and emotional well-being among diverse populations (e.g., Sawatzky et al., 2024).[1] Towards this end, pain and well-being items (Kopec et al., 2006) are measured in a calibration sample and a mixture item response model is developed for each item bank to incorporate heterogeneity in the measurement parameters. Subsequently, the relation between class probabilities and a host of social determinants of health variables are investigated. The results of these models are then carried forward to develop computer adaptive tests (CAT) based on the mixture model (for supporting research, see Sajobi et al., 2022; Sawatzky et al., 2012, 2024; Sawatzky et al., 2018). It is hoped that this approach will help personalize health measurements by tailoring the selection and scoring of items for individuals.

Since a large and diverse calibration sample was desired, the project team collected online data from various sources. However, any online data collection endeavor must worry about contamination with

[1]Future updates and publications about the project will be posted to: https://www.healthyqol.com

participants who do not take the study seriously, and there is persistent worry regarding computer generated responses by survey bots (Perkel, 2020). Recent articles review strategies to deal with such bots, including deterrence/prevention as well as detection (e.g., Simone, 2019; Storozuk et al., 2020). Some steps were taken to prevent survey bots at the time of data collection (e.g., the survey was not widely distributed on social media), and some information was also made available for detection (e.g., checking postal code versus self-reported location).

Here we focus only on efforts to help ensure the integrity of the EPHM calibration sample by detailing an additional flag for random responders or survey bots based on the work of Ilagan and Falk (2024). Under the assumption that Likert-type items are exchangeable for random responders or survey bots, these authors' L1P1 algorithm calibrates sensitivity for detecting such aberrant responses using an unsupervised classification approach. However, using this algorithm with the EPHM data required some modifications for use with missing data. In the remainder of this manuscript, we first describe some details of the EPHM calibration data as it relates to L1P1. We then describe two strategies for adapting L1P1 to the case of missing data, followed by simulations evaluating these two approaches. Finally, we provide results of flagging rates by L1P1 on the actual EPHM calibration sample.

1. Motivation: EPHM calibration data

The main focus of measurement consisted of pain and emotional well-being item banks developed by Kopec et al. (2006). Included were the original item banks (including items that were removed from the final item banks due to poor fit or suspected differential item functioning). Given this setup, a measurement model could be designed to detect random responding (e.g., Jin et al., 2018; Ulitzsch et al., 2022). However, we thought it more expedient to use L1P1 as it does not require a known measurement model for humans and it would be quicker to study with a large sample or in simulations.

Based on Falk et al. (in press), we would expect L1P1 to perform well for EPHM data. As L1P1 was initially developed to handle items with the same number of response options (Ilagan & Falk, 2024), we used just the 76 items (out of 84) with 5 ordinal response options (35 pain, 41 emotional well-being).[2] In simulations with real measurement instruments and 5 category items, L1P1 achieved good classification accuracy (>90%) for inventories with greater than 50 items (Falk et al., in press). Simulations also suggest that uniform information functions (as opposed to peaked information) are good, which appears to be the case for the item banks considered (Kopec et al., 2006).

However, the survey also included "Prefer not to answer" and "Do not know" options for the Likert-type items. For our purposes, these responses were coded as missing data. In addition, some respondents did not complete all survey items or did not finish the questionnaire. As a result, we had data from 11,197 respondents who had at least one complete response on the 76 items, with a total of 10.1% missing data (Figure 1). Of these responses, data were collected from an online survey panel ($N = 5,336$), from partner health organizations ($N = 4,891$), and through a mix of other means ($N = 970$). Since L1P1 does not readily handle missing data, we pursue a modification of it in the following section.

2. L1P1 and missing data

Let $i = 1, \ldots, n$ and $j = 1, \ldots, m$ index respondents and items, respectively, with $z_{ij} \in \{1, 2, \ldots, k_j\}$ the observed response to item j for respondent i, $\mathbf{z}_i = \begin{bmatrix} z_{i1}, & z_{i2}, & \ldots, & z_{im} \end{bmatrix}^\top$ respondent i's full response pattern, and $\mathbf{Z}$ stacking all n response patterns row-wise. Let $y_i \in \{0, 1\}$ be a true class label, with 0 indicating a diligent human and 1 indicating a random responder.

[2] L1P1 has been recently adapted to be able to accommodate inventories with a different number of response options (Ilagan & Falk, 2025b).

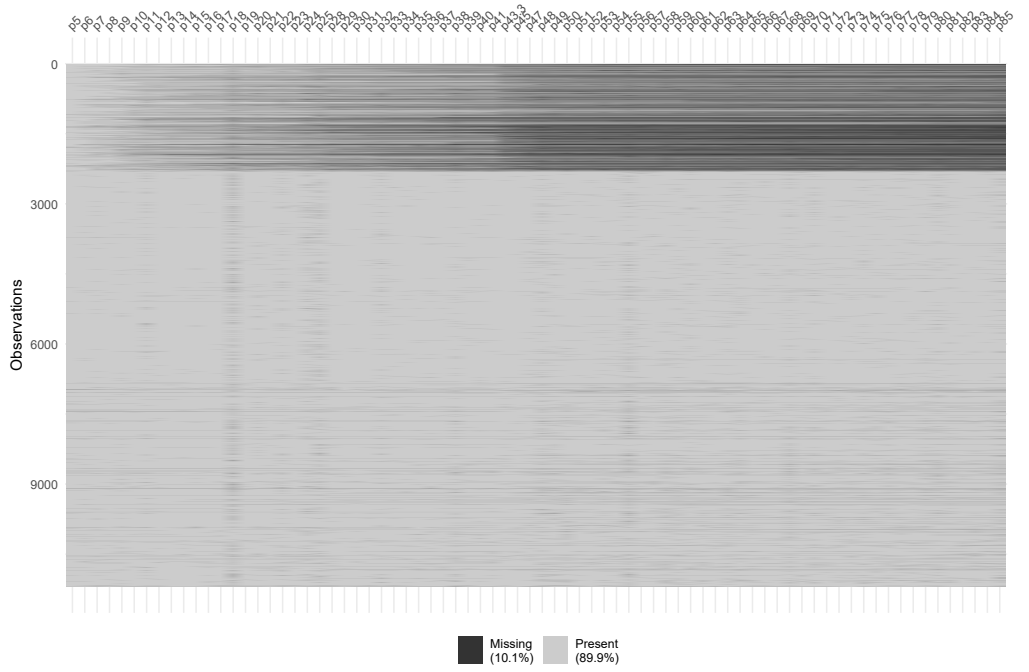

Figure 1. Missing data patterns

For respondent i, classification can be done using the following strategy:

1. Randomly permute $\mathbf{z}_i$, to create B new response patterns: $\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, \ldots, \mathbf{z}_i^{(B)}$.
2. For $\mathbf{z}_i$ and its B permutations, compute outlier statistics, $\mathbf{x}_i$, and $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(B)}$. More than one outlier statistic per response pattern may be computed.
3. Collapse outlier statistics to a single dimension, r_i and $r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(B)}$, and rank-order in terms of suspiciousness. The value p_i indicates the proportion of permutations that are at least as suspicious as $\mathbf{z}_i$.
4. Classify using some threshold τ, $\hat{y}_i = \mathbb{I}\{p_i \geq \tau\}$, where $1-\tau$ corresponds to the desired sensitivity (e.g., $\tau = .05$ is 95% sensitivity).

L1P1 is unique in several respects. At step 1 it assumes random responders have response vectors whose values are exchangeable. Knowledge of the exact data generating mechanism for random responders is not explicitly required, nor is it required for diligent humans.

In addition, step 2 involves computation of outlier statistics. Two examples are Mahalanobis distance (Mahalanobis, 1936; Zijlstra et al., 2011) and person-total cosine similarity,[3] which require an estimate of the item means, $\hat{\boldsymbol{\mu}}$, and covariances, $\hat{\boldsymbol{\Sigma}}$. For instance, Mahalanobis distance for person i is:

$$\sqrt{(\mathbf{z}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{z}_i - \boldsymbol{\mu})} \tag{1}$$

[3] Whereas person-total correlation is the Pearson correlation between a response vector, $\mathbf{z}_i$, and some mean reference vector, e.g., $\hat{\boldsymbol{\mu}}$, person-total cosine similarity is the cosine similarity between these two vectors.

and person-total cosine similarity is:

$$\frac{\mathbf{z}_i^\top \boldsymbol{\mu}}{\|\mathbf{z}_i\|\|\boldsymbol{\mu}\|} \tag{2}$$

where the numerator contains an inner product and the denominator contains Euclidean norms. Thus, $\mathbf{x}_i$ may be a vector of length two that contains Mahalanobis distance and person-total cosine similarity. Importantly, L1P1 uses a leave-one-out strategy where $\mathbf{Z}_{-i}$, the original $\mathbf{Z}$ omitting observation i, is used to compute $\hat{\boldsymbol{\mu}}_{-i}$ and $\hat{\boldsymbol{\Sigma}}_{-i}$, which are then used to compute $\mathbf{x}_i$, and $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(B)}$, from $\mathbf{z}_i$ and $\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, \ldots, \mathbf{z}_i^{(B)}$. For example, $\hat{\boldsymbol{\mu}}_{-i}$ and $\hat{\boldsymbol{\Sigma}}_{-i}$ are substituted for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (1) and (2). At step 3, outlier statistics may be collapsed to a single dimension by computing the distance to some ideal, non-suspicious point.[4] For example, (0, +1) are the least suspicious values for Mahalanobis distance and person-total cosine similarity. The distance between (0, +1) and $\mathbf{z}_i$ is represented by r_i.

To handle missing data, we pursue modifications to steps 1 and 2. First, step 1 permutes $\mathbf{z}_{i,c}$ to obtain new response patterns, where $\mathbf{z}_{i,c}$ is respondent i's response vector containing only their complete responses. Second and at step 2, under ignorable missing data mechanisms, consistent estimates of means and covariances of the items can often (though not always) be obtained using direct maximum likelihood under the assumption of multivariate normality (Yuan, 2009). Under the original L1P1 algorithm, leave-one-out would be used for each permutation test. We may omit observation i when obtaining $\hat{\boldsymbol{\mu}}_{-i}$ and $\hat{\boldsymbol{\Sigma}}_{-i}$ to maximize the log-likelihood:

$$l_{-i}(\boldsymbol{\theta}) = \sum_{i' \neq i}^{n} \left(-\frac{1}{2} \log |\boldsymbol{\Sigma}_{i',c}(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{z}_{i',c} - \boldsymbol{\mu}_{i',c}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{i',c}(\boldsymbol{\theta})^{-1} (\mathbf{z}_{i',c} - \boldsymbol{\mu}_{i',c}(\boldsymbol{\theta})) - \frac{1}{2} m_{i'} \log(2\pi) \right) \tag{3}$$

where $\boldsymbol{\theta}$ is a vector of parameters (all means and covariances), $\boldsymbol{\mu}_{i,c}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{i,c}(\boldsymbol{\theta})$ are partitions of $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ that correspond to complete observations for respondent i, and m_i is their number of complete responses.

Alternatively, since n is large and we need to perform many permutation tests, we may obtain estimates $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ just once that maximize the log-likelihood for all respondents:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log |\boldsymbol{\Sigma}_{i,c}(\boldsymbol{\theta})| - \frac{1}{2} (\mathbf{z}_{i,c} - \boldsymbol{\mu}_{i,c}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{i,c}(\boldsymbol{\theta})^{-1} (\mathbf{z}_{i,c} - \boldsymbol{\mu}_{i,c}(\boldsymbol{\theta})) - \frac{1}{2} m_i \log(2\pi) \right) \tag{4}$$

In either case, although the full $\hat{\boldsymbol{\mu}}$ (or $\hat{\boldsymbol{\mu}}_{-i}$) and $\hat{\boldsymbol{\Sigma}}$ (or $\hat{\boldsymbol{\Sigma}}_{-i}$) are obtained for all items, only the complete subset of elements for respondent i are used to compute outlier statistics, $\mathbf{x}_i$, and $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(B)}$, from $\mathbf{z}_{i,c}$ and its B permutations.

For both strategies, we obtain estimates using software with the expectation-maximization (EM) algorithm (Falk, 2024; Städler & Bühlmann, 2012). When $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are re-used for each permutation test, we refer to this as "pre-computed", and when leave-one-out is used for each permutation test we refer to this as "LOO". The pre-computed strategy slightly violates the premises of the L1P1 algorithm, which could result in a loss of sensitivity. However, it is unclear whether this would occur at such a large sample size. Pre-computing also takes much less time as the item means and covariances only need to be computed once, whereas under LOO Equation 3 needs to be maximized for each row in the dataset.

[4] Any distance metric may suffice, though we used an equation that itself resembled Mahalanobis distance, encompassing the covariance among outlier statistics $\mathbf{x}_i$ and $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(B)}$. For more details, we refer to Ilagan and Falk (2024).

3. Simulations

We conducted a small set of simulations to evaluate algorithm performance for classification of random responders specifically for conditions similar to the EPHM data. We also wanted to evaluate the two strategies for computing item means and covariances under missing data.

3.1 Data generation and analysis

To mimic EPHM calibration data, the total sample size for each generated dataset was fixed at 11,197. We manipulated random responder contamination rate (.05, .5, and .95). We generated hypothetical human data by utilizing estimated model parameters for EPHM data based on a 2-dimensional graded response model with a logit link function (Reckase, 2009; Samejima, 1969), and with very simple structure utilizing all 84 items for both pain and emotional well-being item banks.[5] Only item parameters from the 76 5-category items were then used. These model parameter estimates were treated as true values in generating hypothetical responses when $\gamma_i = 0$. Although not corresponding to a mixture model and possibly contaminated with random responders, parameter estimates from this analysis looked reasonable in light of work by Kopec et al. (2006). Thus, we argue this is a reasonable strategy for generating hypothetical humans. For $\gamma_i = 1$, each item response was drawn from a uniform distribution over the 5 possible response categories. Once a complete dataset with the desired proportion of humans and bots was generated, these rows were randomly sorted, and missing data was induced using the exact same missing data patterns as under the EPHM calibration sample.

We generated 100 datasets per each contamination rate and analyzed them using the pre-compute strategy. Due to computational time, we generated 20 datasets per each contamination rate for use with the LOO strategy. For both strategies, $B = 1,000$ permutations were used and person-total cosine similarity and Mahalanobis distance were used as outlier statistics. Since we later wished to be cautious about accidentally flagging humans on the real EPHM data, we used $\tau = .1$ for 90% sensitivity as we expected it would exhibit more specificity than $\tau = .05$ as used by Ilagan and Falk (2024). Custom R code and results for these simulations are available on the Open Science Framework: https://osf.io/t7br2/.

3.2 Results

We report three calibration metrics: sensitivity ($\sum_{i=1}^{n} \mathbb{I}\left\{\hat{\gamma}_i = \gamma_i = 1\right\} / \sum_{i=1}^{n} \mathbb{I}\left\{\gamma_i = 1\right\}$), specificity ($\sum_{i=1}^{n} \mathbb{I}\left\{\hat{\gamma}_i = \gamma_i = 0\right\} / \sum_{i=1}^{n} \mathbb{I}\left\{\gamma_i = 0\right\}$), and classification accuracy ($\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{\hat{\gamma}_i = \gamma_i\right\}$). Each were averaged across all datasets for all cells of the design, but also binned by the number of complete responses available for any given row.

LOO was able to maintain the target sensitivity as it had around 90% sensitivity across all contamination rates and across rows with different amounts of missing data (bottom of Figure 2). However, pre-computing experienced a loss of sensitivity, especially under lower contamination and when the row had more complete responses (top of Figure 2). Presumably, a more complete response pattern may allow the row to influence mean and covariance estimates, making it look less suspicious.

While specificity looked similar across pre-computing and LOO, it only achieved acceptable rates (> 90%) with 20-30 complete responses or more, and when contamination was .5 or .05 (Figure 3).

Classification accuracy tended to be slightly better under LOO than it was under pre-computing (Figure 4). Though in some cases, accuracy was the same or differed by only a percentage point.

Overall then, LOO should be slightly preferred as it may flag a few more random responders than will pre-computing, but either strategy should yield rather similar results.

[5] Reverse-worded items were reverse-coded prior to these analyses.

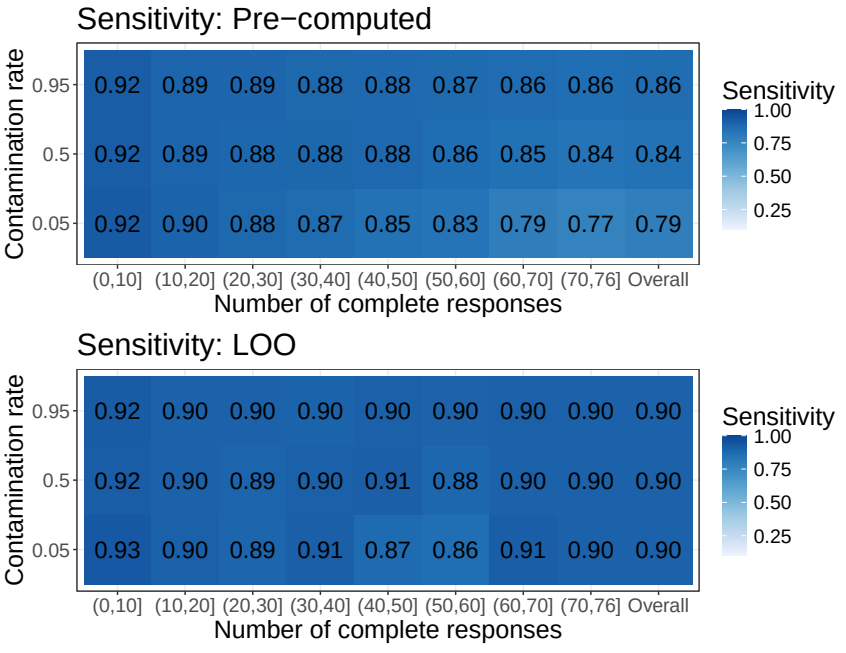

Figure 2. Sensitivity

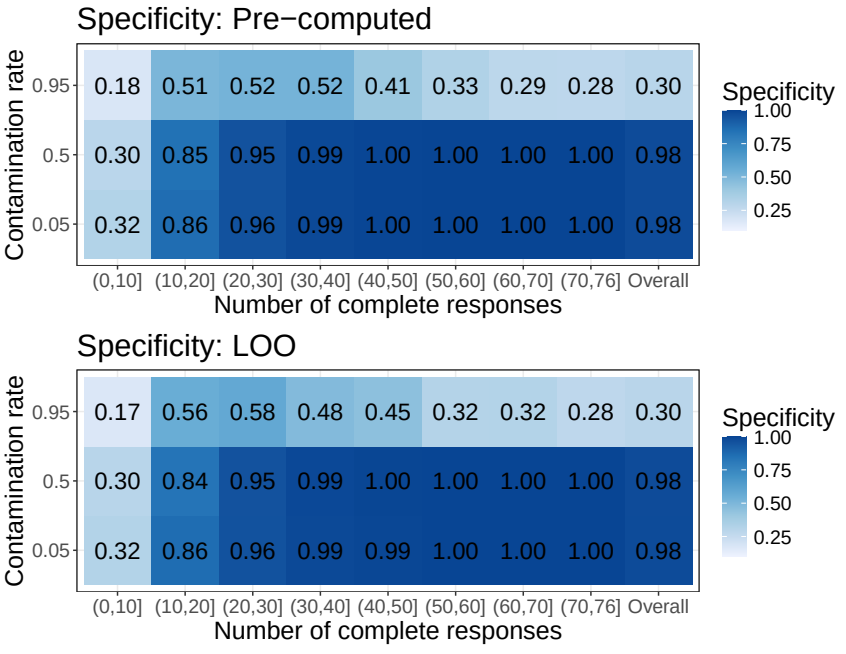

Figure 3. Specificity

4. Empirical results

On the actual EPHM data, we applied L1P1 by pre-computing means and covariances and with LOO with $B = 1,000$ and $\tau = .1$ for 90% sensitivity. We report results of LOO since it performed slightly

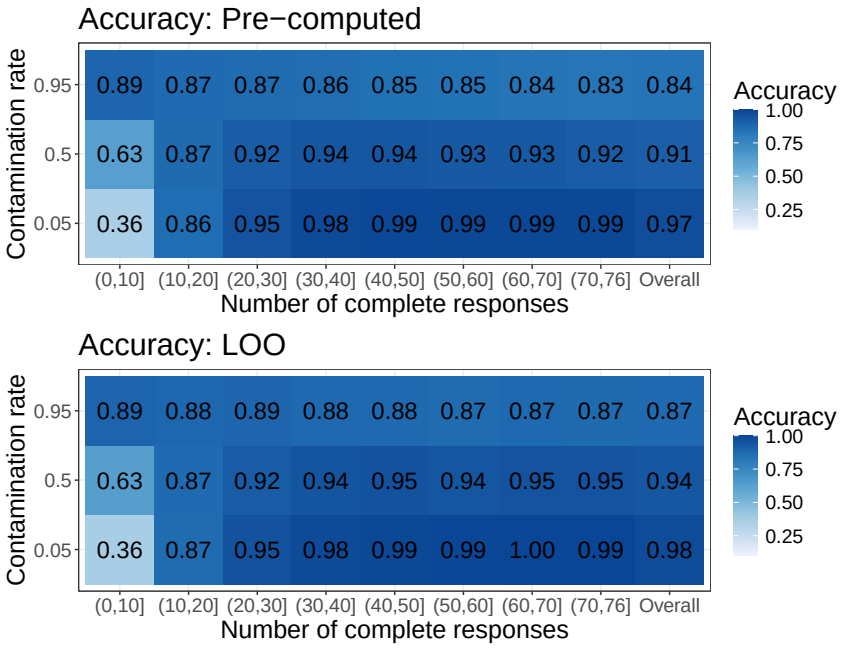

Figure 4. Accuracy

better in simulations and resulted in flagging only an additional 16 respondents. The number of flagged respondents (i.e., $\hat{y}_i = 1$) was also binned by the number of complete responses (Table 1). For some respondents a flag could not be generated and was marked as "NA" for missing; this typically occurred when the respondent utilized only one or two response categories across all completed items and also when the respondent completed very few items (Table 1). Though flagging rates also tended to be higher when the respondent used few response categories (Table 2). In examining the data source, we noticed that the highest flagging rates were among the online survey panel participants, as opposed to those recruited by partner health organizations or other sources (Table 3).

Table 1. Flag rates for EPHM data by number of completed items

Complete Items	N	N Flag	Prop. Flag	N NA	Prop. NA
(0,10]	415	172	0.41	174	0.42
(10,20]	272	73	0.27	36	0.13
(20,30]	245	42	0.17	17	0.07
(30,40]	391	65	0.17	14	0.04
(40,50]	108	10	0.09	4	0.04
(50,60]	130	17	0.13	2	0.02
(60,70]	251	23	0.09	1	0.00
(70,76]	9385	254	0.03	65	0.01
Overall	11197	656	0.06	313	0.03

N = count or sample size; Prop. = Proportion; NA = Missing, no flag could be generated.

Table 2. Flag rates for EPHM data by number of categories used

	1	2	3	4	5
$\gamma_i = 0$	0	148	478	3498	6104
$\gamma_i = 1$	0	260	178	104	114

Table 3. Flag rates for EPHM data by data source

	N	N Flag	Prop.Flag	N NA	Prop. NA
Online survey panel	5336	490	0.09	230	0.04
Partner health organizations	4891	122	0.02	64	0.01
Other	970	44	0.05	19	0.02

N = count or sample size; Prop. = Proportion; NA = Missing, no flag could be generated.

5. Conclusion

L1P1 can be combined with modern ways to handle missing data such as the EM algorithm to compute means and covariances for use with outlier statistics. Doing this using LOO for each row appeared to perform best in simulations. Although the performance of L1P1 with real data from heterogeneous populations has been evaluated (Falk et al., in press; Ilagan & Falk, 2024), the present simulations could be improved by introducing some heterogeneity for simulated humans. Nonetheless, we sought quick answers as to whether such an algorithm was feasible for the real EPHM data. Handling of missing data is now available in a forked version of the `detranli` package (Ilagan & Falk, 2025a) and may eventually be incorporated into the main repository.

Combining results from simulations and the actual EPHM analyses, application of L1P1 with LOO suggests that data collection was not overrun with random responders. Although 6% and 3% of respondents were either flagged or a flag could not be generated, the majority of these respondents completed few items. Based on simulations, for those with few complete responses (e.g., less than 20–30), the results of L1P1 may not be trustworthy as many diligent humans may be accidentally flagged. Furthermore, respondents with few completed items would presumably have little influence on any subsequent analyses (i.e., development of the mixture CAT) as they would contribute few responses to the estimated model(s). Such a conjecture could be tested by performing a sensitivity analysis with and without flagged respondents.

Acknowledgement

We are grateful to the EPHM team, including researchers, healthcare providers, patient partners, and partner organizations who were part of the original project proposal and subsequent implementation: https://webapps.cihr-irsc.gc.ca/decisions/p/project_details.html?applId=450658&lang=en

Funding Statement The EPHM project is supported by the Canadian Institutes of Health Research (CIHR), (Project Grant #468626), and the Canada Research Chairs program (# CRC-2022-00155). 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). Le projet EPHM est soutenu par les Instituts de recherche en santé du Canada (IRSC), [subvention de projet #468626], et le programme des chaires de recherche du Canada (# CRC-2022-00155). 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).

Competing Interests None.

References

Falk, C. F. (2024). *Emgaussian: Expectation-maximization algorithm for multivariate normal (Gaussian) with missing data* [R package version 0.2.1]. https://CRAN.R-project.org/package=EMgaussian

Falk, C. F., Huang, A., & Ilagan, M. J. (in press). Unsupervised [randomly responding] survey bot detection: In search of high classification accuracy. *Psychological Methods*. https://doi.org/10.1037/met0000746

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. (2025a). Detranli: DEtection of RANdom LIkert-type responses [https://github.com/falkcarl/detranli].

Ilagan, M. J., & Falk, C. F. (2025b). Unsupervised detection of random responding for likert-type inventories with varying numbers of response categories. In *Proceedings of the 89th annual International Meeting of the Psychometric Society*.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, *21*(1), 197–225. https://doi.org/10.1177/1094428117725792

Kopec, J. A., Sayre, E. C., Davis, A. M., Badley, E. M., Ambrahamowicz, M., Sherlock, L., Williams, J. I., Anis, A. H., & Esdaile, J. M. (2006). Assessment of health-realted quality of life in arthritis: Conceptualization and development of five item banks using item response theory. *Health and Quality of Life Outcomes*, *3*, 33. https://doi.org/10.1186/1477-7525-4-33

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, *12*, 49–55.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Sajobi, T. T., Lix, L. M., Russell, L., Schulz, D., Liu, J., Zumbo, B. D., & Sawatzky, R. (2022). Accuracy of mixture item response theory models for identifying sample heterogeneity in patient-reported outcomes: A simulation study. *Quality of Life Research*, *31*, 3423–3432. https://doi.org/10.1007/s11136-022-03169-0

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, *17*.

Sawatzky, R., Ratner, P. A., Kopec, J. A., & D., Z. B. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research*, *21*, 637–650. https://doi.org/10.1007/s11136-011-9976-6

Sawatzky, R., Verdam, M. G. E., Mehdipour, A., Ratner, P. A., Antonio, M., Courtney, K., Cuthbertson, L., Falk, C. F., Gadermann, A., Jackson, J., Kwon, J.-Y., Lix, L. M., Liu, J., Sajobi, T. T., Schick-Makaroff, K., Wong, H., & Zumbo, B. D. (2024). Are social determinants of health associated with measurement bias? *Quality of Life Research*, *33 (Suppl 1)*, S53–54. https://doi.org/10.1007/s11136-024-03786-x

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*, *27*, 1745–1755. https://doi.org/10.1007/s11136-017-1680-8

Simone, M. (2019). How to Battle the Bots Wrecking Your Online Study. *Behavioral Scientist*. https://behavioralscientist.org/how-to-battle-the-bots-wrecking-your-online-study/

Städler, N., & Bühlmann, P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, *22*, 219–235. https://doi.org/10.1007/s11222-010-9219-7

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

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3). https://doi.org/10.1111/bmsp.12272

Yuan, K.-H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis*, *100*, 1900–1918. https://doi.org/10.1016/j.jmva.2009.05.001

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