BMJ Open

Identifying patterns of item missing survey data using latent groups: an observational study

Adrian G Barnett,[1] Paul McElwee,[1,2] Andrea Nathan,[1,2] Nicola W Burton,[3] Gavin Turrell[1,2]

To cite: Barnett AG, McElwee P, Nathan A, *et al*. Identifying patterns of item missing survey data using latent groups: an observational study. *BMJ Open* 2017;**7**:e017284. doi:10.1136/bmjopen-2017-017284

► Prepublication history and additional material for this paper are available online. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2017-017284).

Received 12 April 2017
Revised 23 July 2017
Accepted 25 July 2017

[1]School of Public Health and Social Work and Institute of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, Queensland, Australia
[2]Institute for Health and Ageing, Australian Catholic University, Melbourne, Victoria, Australia
[3]School of HumanMovement Studies, The University of Queensland, Brisbane, Queensland, Australia

Correspondence to
Dr Adrian G Barnett;
a.barnett@qut.edu.au

ABSTRACT

Objectives To examine whether respondents to a survey of health and physical activity and potential determinants could be grouped according to the questions they missed, known as 'item missing'.

Design Observational study of longitudinal data.

Setting Residents of Brisbane, Australia.

Participants 6901 people aged 40–65 years in 2007.

Materials and methods We used a latent class model with a mixture of multinomial distributions and chose the number of classes using the Bayesian information criterion. We used logistic regression to examine if participants' characteristics were associated with their modal latent class. We used logistic regression to examine whether the amount of item missing in a survey predicted wave missing in the following survey.

Results Four per cent of participants missed almost one-fifth of the questions, and this group missed more questions in the middle of the survey. Eighty-three per cent of participants completed almost every question, but had a relatively high missing probability for a question on sleep time, a question which had an inconsistent presentation compared with the rest of the survey. Participants who completed almost every question were generally younger and more educated. Participants who completed more questions were less likely to miss the next longitudinal wave.

Conclusions Examining patterns in item missing data has improved our understanding of how missing data were generated and has informed future survey design to help reduce missing data.

Strengths and limitations of this study

- A better understanding of item missing data could help improve survey design and help determine if the data are missing at random.
- Identifying patterns in missing data is just part of the process and researchers need to use additional methods to impute or weight data to account for the potential biases of missing data.
- We show results for two different groups sizes (3 and 14) and choosing the optimal group size may be based on a combination of goodness of fit statistics and the researchers' preferences.

INTRODUCTION

Missing data is an almost ubiquitous problem for surveys, and participants may fail to complete whole surveys or partially complete surveys. Ignoring missing data by using a complete case analysis can create potentially serious biases when the ratio of missing to complete data is large, or when participants with missing data are very different from participants with complete data.[1] For example, participants with complete data may be healthier, which would be a particular problem if we were interested in measuring general health.

Longitudinal studies often have both wave missing data, where a participant failed to complete an entire wave, and item missing data, where a participant partially completed a wave.[2] In this paper, our focus is on item missing data in returned mail surveys. Wave missing usually creates more serious biases than item missing as more information is lost, however, item missing data can compound the biases of wave missing data.

An understanding of the potential causes of item missing data may inform procedures to impute missing data, and may help to improve future data collections by changing the way questions are worded, displayed or ordered.[3 4] A better understanding of the patterns of missing data may also help determine if the data are missing at random.[5] To identify the missing data mechanism, we need to use as much information as possible to determine why the questions were not completed.[6]

In this analysis, we identify patterns of item missing data in survey responses using a case study of a survey of physical activity and potential predictors of activity such as neighbourhood perceptions, attitudes, health and sociodemographics. We used latent class analysis to identify groups of participants who tended to miss similar questions.

For example, to find a group of participants who rarely completed potentially sensitive questions such as income.

There is a large literature on wave missing data and item missing data, but most papers have been concerned with imputing missing data in order to overcome the potential biases caused by missing data.[7] Examining the pattern of missing data using tables and plots of both wave missing and item missing (yes/no) has been recommended by Rubin and Little[8] as this can inform the method of imputation and highlight problems such as variable combinations that were never jointly observed. This can be achieved using the 'mi' package in R which plots the overall pattern of missing data with the option of ordering participants by their proportion missing.[9] Pattern-mixture models have been used to model common patterns of missing as a between-subject factor for longitudinal analysis.[10] In this paper, we model patterns of item missing data in order to identify problem questions or groups of questions, and to examine whether there are groups of participants who miss similar questions.

MATERIALS AND METHODS

Data

We used data from the longitudinal HABITAT (How Areas in Brisbane Influence healTh And acTivity) study.[11] This is a population-based study of people aged 40–65 years in 2007 living in Brisbane, Australia. The aims of the study are to understand the factors enabling and limiting physical activity in mid-age. Most questions in the self-completed mail surveys were either answered by ticking a box on a Likert scale or writing a number in a box with just a few free text (open-ended) responses. Some questions could be legitimately missing, for example, hours of work only needed to be completed if the respondent was currently working. We excluded these conditional questions from this analysis as they could be legitimately missing.

The study began in 2007 and had follow-up waves in 2009 and 2011. The response rates per wave for all participants sent a questionnaire were 68% in 2007 (n=11 035 returned questionnaires), 72% in 2009 (n=7867) and 67% in 2011 (n=6901).

Data were collected using detailed mail surveys of 250 or more questions. The survey questions were grouped in sections such as physical activity participation, neighbourhood characteristics (eg, traffic), general health and lifestyle, and demographics (eg, employment). The full surveys are available online at Institute for Health and Ageing's website (https://iha.acu.edu.au/research/research-projects/habitat-project). A number of strategies were used to encourage a high response rate, including advanced notice, personalised cover letters, surveys labelled with suburb area, and reminder letters.[11] Participants were encouraged to answer every question. The instructions at the front of the survey stated, "Some of the questions may sound the same. However, it will help us greatly if you answer all questions".

The HABITAT study including the procedure for participants providing informed consent was approved by the human research ethics committee of the Queensland University of Technology.

Statistical methods

Our overall aim was to find similar patterns of item missing data using latent groups. This could also be called unsupervised classification or clustering.[12]

We used a mixture of multivariate multinomial distributions to create latent groups with similar patterns of item missing using Rmixmod.[13] The multinomial distributions modelled the average probability of missing across the 286 questions from the 2011 survey. Each participant was allocated the latent group that best matched their pattern of item missing. We allowed unequal group sizes, so, for example, two latent groups might have 90% of participants in one group and 10% in the other. The probability density function is a weighted sum of multinomial distributions:

$$g(\mathbf{x}_i) = \sum_{k=1}^{K} p_k f(\mathbf{x}_i; \pi_k), \quad i = 1, \ldots, n,$$

where $\mathbf{x}$ is the observed binary data of missing (yes/no), n is the total number of observations, p are the latent group probabilities that sum to 1 and π_k are the missing probabilities in each group. A key question is the optimal number of latent groups needed to capture the range of missing patterns in the data (K in the above formula). We chose the optimal number of latent groups using the Bayesian information criterion (BIC) which makes a trade-off between a good fit to the data and model complexity.[14] The models were fitted using the Rmixmod package in the R software V.3.3.1 (code and data are available at GitHub (https://github.com/agbarnett/item.missing)).[15] We tested groups sizes between 2 and 20 and chose the optimal number using the smallest BIC. We discarded models which included any group with fewer than 1% of participants, as such small groups were considered too small to be meaningful.

To examine the differences between latent groups, we plotted the mean estimated missing probability per question against question order as we believed that fewer questions would be missed at the start of the survey. Using the equation notation the plotted mean is $\hat{\pi}_{j,k}$ for question j and group k. We used a kernel density smoother to illustrate the average probability of missing by question order.[16]

We calculated the overall mean probability of item missing in each latent group, using the below equation:

$$\bar{\pi}_k = \sum_{j=1}^{Q} \hat{\pi}_{j,k} / Q,$$

where Q is the total number of questions. We also give the range in each group. To ease comparisons in the tables and plots, we numbered the latent groups using this mean probability from lowest to highest. We also refer to

the latent groups based on their mean overall probability, such as 'low missing' and 'high missing'.

We expected the characteristics of the participants to vary by latent group, as a previous analysis of similar physical activity data found that individual characteristics, such as gender, predicted the number of missing items.[17] We examined whether a participant's age, gender and education predicted whether they would be in the latent group with the least amount of missing. We used predictors at the same wave (year 2011), but if these were missing then we used the value from the previous wave with an adjustment for age (eg, age from 2009 plus 2 years). We fitted this multiple logistic regression model using a Bayesian paradigm and used WinBUGS V.1.4.3[18] and presented the results as prevalence ratios.[19] We used a burn-in of 5000 Monte Carlo Markov chain (MCMC) iterations followed by a sample of 15 000 thinned by 3. The Bayesian multiple logistic regression model is given below:

$$L(i) \sim \text{Bernoulli}\,[p(i)], \quad i = 1, \ldots, n,$$
$$\text{logit}\,[p(i)] = \alpha \mathrm{X_i},$$
$$\text{(prior distribution)}\ \alpha_k \sim \text{Normal}\,(0, 10^4), \quad k = 1, \ldots, 11,$$

where $L(i)$ is the binary dependent variable of being in latent group 1 and $\mathbf{X}$ is an $n \times 11$ matrix of an intercept and the predictors of gender, age and education which has nine categories with the lowest educational level as the reference category. To aid convergence of the MCMC samples, we standardised age by subtracting 56 years (the mean) and dividing by 10 years. A standard frequentist logistic regression model could equally have been used.

We used the highest probability to assign each participant to their latent class, $L(i)$. However, there could be some uncertainty in this assignment as a participant may have a relatively high probability for multiple groups. For example, if there were five latent groups a participant with a probability of 0.2 of being in each group would reflect great uncertainty. To examine this issue, we used summary statistics for the highest probability for all participants.

We used logistic regression to examine whether the amount of item missing in a survey predicted wave missing for the next survey. Our hypothesis was that greater item missing data would increase the probability that a participant failed to complete the next wave. We used the proportion of item missing in the 2007 survey (wave 1) to predict the return of the 2009 survey (wave 2) and the proportion of item missing in the 2009 survey (wave 2) to predict the return of the 2011 survey (wave 3). We modelled these two waves using a generalised linear mixed model with a random intercept per participant to account for the non-independence of data from the same participant.[20] Participants who did not complete the 2009 survey were only included once.

The association between the proportion of item missing data and missing the next wave could be non-linear, for example, with a stronger effect at higher proportions. To allow for a range of non-linear curves, we used the fractional polynomials approach to give interpretable curves using the equation:

$$W(i, w) \sim \text{Bernoulli}\,[\mathrm{p}(i, w)], \quad i = 1, \ldots, N,\ w = 2, 3,$$
$$\text{logit}\,[\mathrm{p}(i, w)] = \beta_0 + \beta_1 m_{i,w-1}^{P} + \gamma_i,$$
$$\gamma_i \sim \mathrm{N}(0, \sigma_\gamma^2),$$

where $\mathrm{W}(i, w)$ is the binary dependent variable of wave missing for participant i at wave w; $m_{i,w-1}$ is the proportion of item missing from the previous survey; γ is a random intercept to control for repeated results from the same participant and the power P is one of: –2, –1, –0.5, 0, 0.5, 1, 2, 3. For P=0 we use $\log(m_{i,w-1})$ in place of $m_{i,w-1}^{P}$. We chose the best model (best P) as that with the smallest deviance.[21] To evaluate the best model, we computed the overall goodness of fit statistic as compared with a model that did not include the proportion of item missing from the previous survey: $\text{logit}\,[\mathrm{p}(i)] = \beta_0 + \gamma_i$.

The model was fitted using the glmer function in the lme4 library in R.[20]

Table 1 Summary statistics on the percentage of missing data for questions and participants

Unit of analysis	N	Min	First Q	Median	Mean	Third Q	Max
Question	286	0.24	1.04	1.40	1.84	1.88	13.97
Participant	6901	0.00	0.00	0.35	1.78	1.40	99.65

RESULTS

Summary statistics on the amounts of missing data by participant and question are shown in table 1. The question with the least amount of missing data was "I live on or near a main road or busy throughway for motor vehicles" with just 0.24% missing, followed by gender with just 0.32% missing. The question with the most amount of missing data was "Do you plan to use the Bike Hire Scheme?", with 13.97% missing. A small number of participants (0.1%) missed 82% or more of the questions.

The plot of the BIC for choosing the maximum number of latent groups is shown in figure 1. The best model (smallest BIC) is for a maximum latent group size of 14. However, 7 of the 14 groups contained less than 1% of the sample, with the smallest group having just 28 participants. We considered these groups to be too small to have a meaningful interpretation. We therefore selected a maximum group size of three as this was the smallest BIC where all group sizes were above 1%. We examine smaller latent groups in a sensitivity analysis below.

Missing pattern by question order

The estimated probabilities of missing for the three latent groups are shown in figure 2. Group 1 had a low average probability of missing of just 0.01 (range <0.0001 to 0.13) and this was relatively constant throughout the survey except for an unusually high probability of missing of over 0.1 for two questions on bicycle use and a question

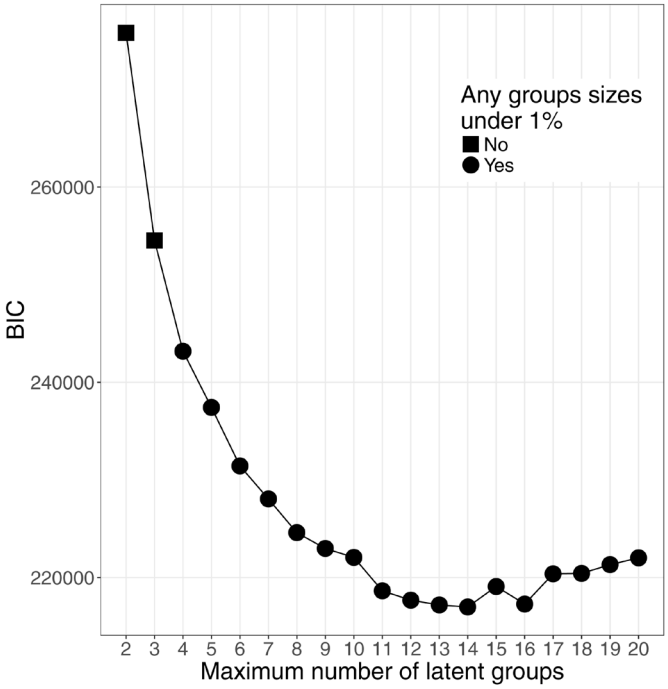

Figure 1 Bayesian information criterion (BIC) for choosing the maximum number of latent groups. Plotting symbols show whether any latent group had fewer than 1% of participants.

on sleeping hours on weekend days. Group 2 had an average missing probability of 0.05 (range 0.002 to 0.45), with a high missing probability of 0.30 for the set of 10 questions on days walking to places in the last month. Group 3 had the highest average missing probability of 0.19 (range 0.03 to 0.35). In this group the probability of missing increased from the start to around the middle of the survey, with a reduction towards the end of the survey. The numbers (percentages) in each group were 5726 (83%), 877 (13%) and 298 (4%).

Group characteristics

Age, gender and education were all strongly associated with being in the 'low missing' latent group 1 (table 2). Every 10-year increase in age reduced the probability of being in the low missing group with a mean prevalence ratio of 0.945 (95% CI 0.921 to 0.965), so older people had generally more missing data. Women also had more missing data as their prevalence ratio of being in the low missing group was 0.940 (95% CI 0.910 to 0.969). Higher levels of education were associated with fewer missing items. Compared with the reference group with the least amount of education, every other education group had a mean prevalence ratio above 1, although for two groups the 95% CIs included 1.

Ninety-nine per cent of participants had a maximum probability of latent group membership above 0.99. Hence uncertainty in latent group membership is unlikely to be an issue.

Item missing predicting subsequent wave missing

There was a rising probability of wave missing with greater item missing (figure 3), and this association was strongly statistically significant (overall model fit χ^2=66.2 (df=1), p<0.0001). The best non-linear curve using fractional polynomials was a square-root transformation (p=0.5). The non-linear curve sharpens as the proportion tends to 0, which likely reflects the increased diligence of those participants who miss only a few questions.

Smaller latent group sizes

We initially used a minimum group size of 1%, but this may hide small latent groups with interesting patterns of missingness. We therefore plot the mean probabilities of missing using the minimum BIC of 14 latent groups (figure 4). The largest group was 4745 participants who almost fully completed the questionnaire with a mean missing probability of 0.003 (range 0.0001 to 0.11). The next largest group had a similar pattern of missingness to group 1, but with a higher mean missing probability 0.006 (range 0.0008 to 0.14).

Some of the groups had high probabilities of missing to specific groups of questions, such as: group 3 for the questions on number of days in the last month walking to each of 10 specified places; group 4 for the questions on 12 long-term health conditions and group 11 for the 16 questions on attitudes about transport and 15 questions on attitudinal barriers and motivations to physical activity. Two other interesting patterns (both with less than 50 participants) were group 12 who generally completed fewer questions with increasing question order and group 13 who—after the first 50 questions—generally completed more questions with increasing question order.

DISCUSSION

Mixture modelling identified three latent groups that made logical sense as they represented excellent, good and poor item completers. The average probability of missing in the three groups was 0.01, 0.05 and 0.19. The good news for the HABITAT study was that 83% of participants were excellent completers, with just 4% as poor completers.

For the low missing group the question with the most missing was: "On a usual 24 hour day, how many hours do you spend sleeping on a weekend day?" (11.3% missing in low missing group), and this was a clear outlier compared with most of their other answers (figure 2). The question as it appeared in the survey is shown in figure 5, which shows how it was paired with a question on sleeping hours during the week. It is possible that people skipped the weekend question because their answer was the same as the weekday answer. But as this group had such low rates of item missing for the rest of the survey there may have been some misunderstanding. Starting the question with 'On a usual 24 hour day' may have primed people to think about all days. Alternatively it may be that the weekend numbers

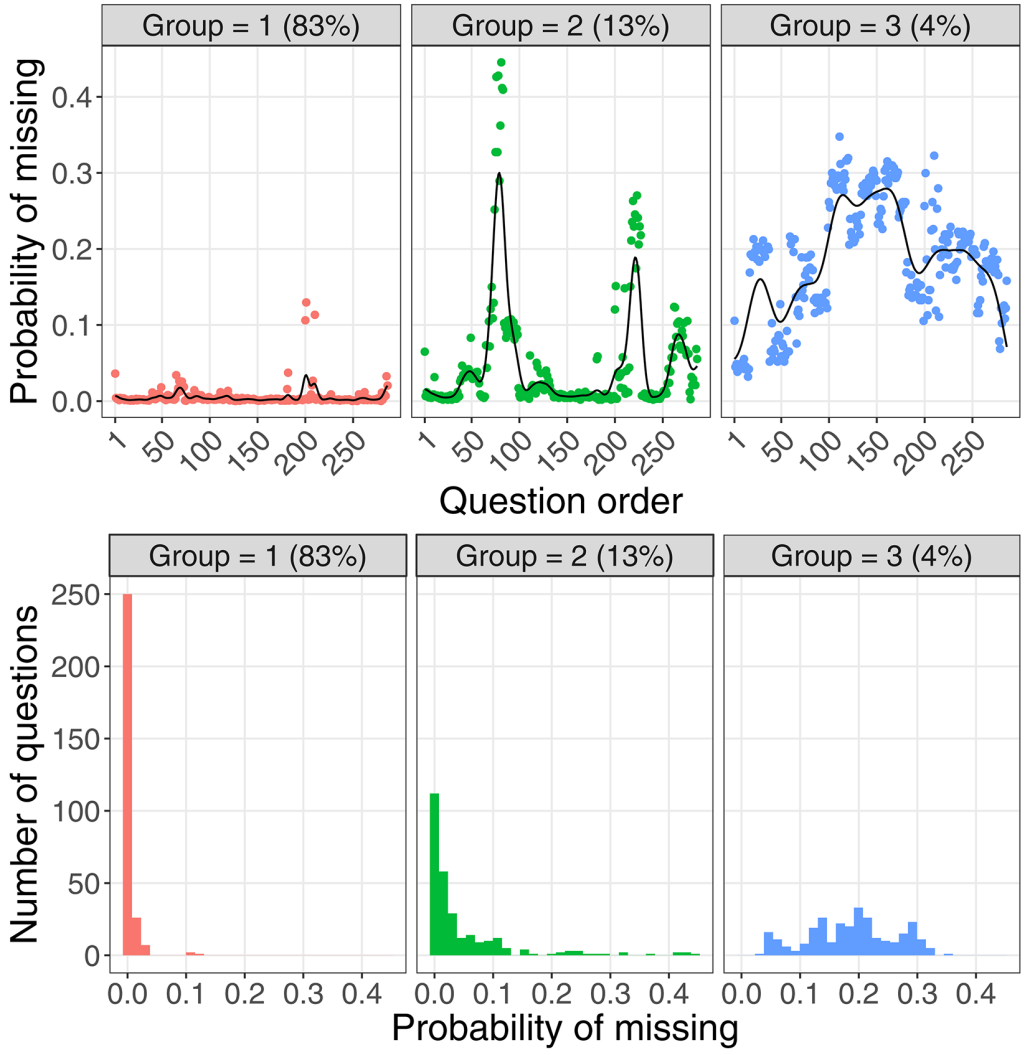

Figure 2 Estimated mean probability of missing by question order (top row) and histogram of mean probabilities (bottom row) for three latent groups. The black lines in the top row are the smoothed mean missing probability using a kernel smooth with a manually chosen window size of 7 for group 1, 11 for group 2 and 23 for group 3. Panel headings show the latent group number and sample per cent.

Table 2 Prevalence ratios and 95% CIs for the probability of being in latent group 1 with the lowest probability of missing

	Per cent change	95% CI	p Value
Age (+10 years)	0.945	0.921 to 0.965	<0.001
Gender=Female	0.940	0.910 to 0.969	<0.001
Education			
Year 10	1.066	1.014 to 1.125	0.010
Year 11	1.064	0.991 to 1.138	0.088
Year 12	1.057	0.998 to 1.121	0.060
Certificate (trade/business)	1.077	1.023 to 1.139	0.002
Diploma/Associate degree	1.060	1.006 to 1.125	0.029
Bachelor degree	1.102	1.048 to 1.164	<0.001
Graduate diploma or Certificate	1.069	1.009 to 1.133	0.021
Postgraduate degree	1.099	1.041 to 1.167	<0.001

The reference group for education was 9 years or less (the lowest education level). Estimates from a Bayesian multiple logistic regression model.

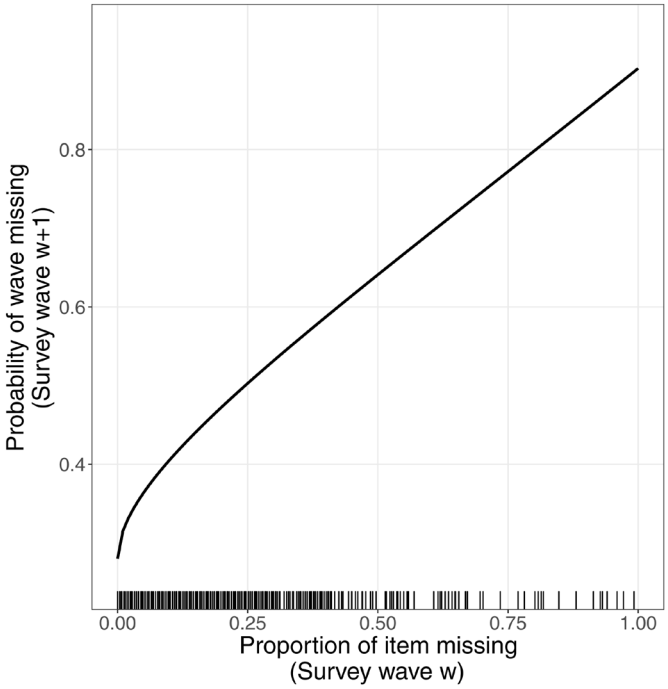

Figure 3 Estimated mean probability of wave missing dependent on the proportion of item missing from the previous survey. The rug plot along the x-axis shows the observed proportions of item missing.

looked like they were the minutes of the weekday question, even with the large 'AND', and so were left blank because the participants were thinking of whole hours. Another explanation is that most other questions in the survey used one response per row. For example, just above this question was a question on weight that had the two options of kilograms or stone and pounds on the same row. Changing the layout of this question to have separate rows for weekend and weekday sleeping could reduce missingness.

The question on usual sleeping hours on the weekend was the most frequently missing across all questions (12.7% missing over all participants), so this question could have been identified as a problem using simple percentages. However, the fact that it was so often missing in the group that completed almost all the survey strongly suggests that the number of missing responses could be decreased.

For the middle missing group (group 2), there were three questions with relatively high missing probabilities (figure 2), these were: the questions on number of days in the last month walking to each of 10 specified places (30.4% missing), the questions on 12 long-term health conditions, for example, diabetes (19.3% missing), and the question on usual weekend sleeping hours (17.4% missing).

The walking questions may have been poorly completed because they asked about incidental behaviour such as walking to cafés and restaurants, which is an unstructured behaviour that may be hard to recall. Participants may have found the count of walking during the past month too difficult to attempt. Responses required a written answer (compared with a tick box), which may have been perceived as too burdensome. It is also possible that some people never walked to particular places (such as the library), but failed to write 'zero' as requested in the instructions. The lowest number of missing answers was for the supermarket, which arguably most people would need to visit, and the highest missing was for work, which may not be applicable for retired respondents (see online supplementary appendix). This suggests the need for clearer instructions about writing 'zero' rather than skipping the question, and possibly also including a 'not applicable' or 'do not know' response option.

The 12 questions on long-term health conditions had a simple dichotomous response option (yes/no), but may have had more missing data because they asked about potentially sensitive issues such as depression. It is also possible that people found it hard to answer because conditions needed to have lasted for 6 months or more, and this may be difficult to recall especially for conditions that occurred some time ago or intermittently.

For the high missing group, the two least completed questions were the set of 15 questions on barriers to and motivations for physical activity (85.6% missing) and the set of six questions on confidence to do physical activity in the context of competing demands, for example, fatigue (75.5% missing). These were in the middle section of the survey that was most missed by this group (figure 2), and hence it may be the questions' placement rather than content that was important. However, these questions also require more insight and personal reflection compared with other more factual questions (such as neighbourhood features), and so some respondents may have been reluctant to take the extra time needed.

The average pattern in the high missing group of a rise and fall in the probability of missing through the survey (figure 2) could be because many participants completed some questions at the start before losing interest and skipping to the end. For future survey waves, we could consider sending this group a shorter survey[22] or reordering the survey to put the most important questions at the start or end. However the current survey starts with some general 'warm-up' questions (asking people about their neighbourhood area) and putting more challenging questions first may increase item and wave missing numbers.

The strong association between higher item missing predicting future wave missing (figure 3) is not surprising and could well be because people who partially completed the survey were not engaged with the study and are also likely to miss a wave. This association could be used when trying to impute wave missing data, as we could choose to give greater weight to the responses from partial completers when using techniques such as inverse probability weighting.[23] However, the obvious catch is that by definition these participants may also be missing the variables of interest. This problem could be somewhat overcome by putting the most important questions early in the survey.

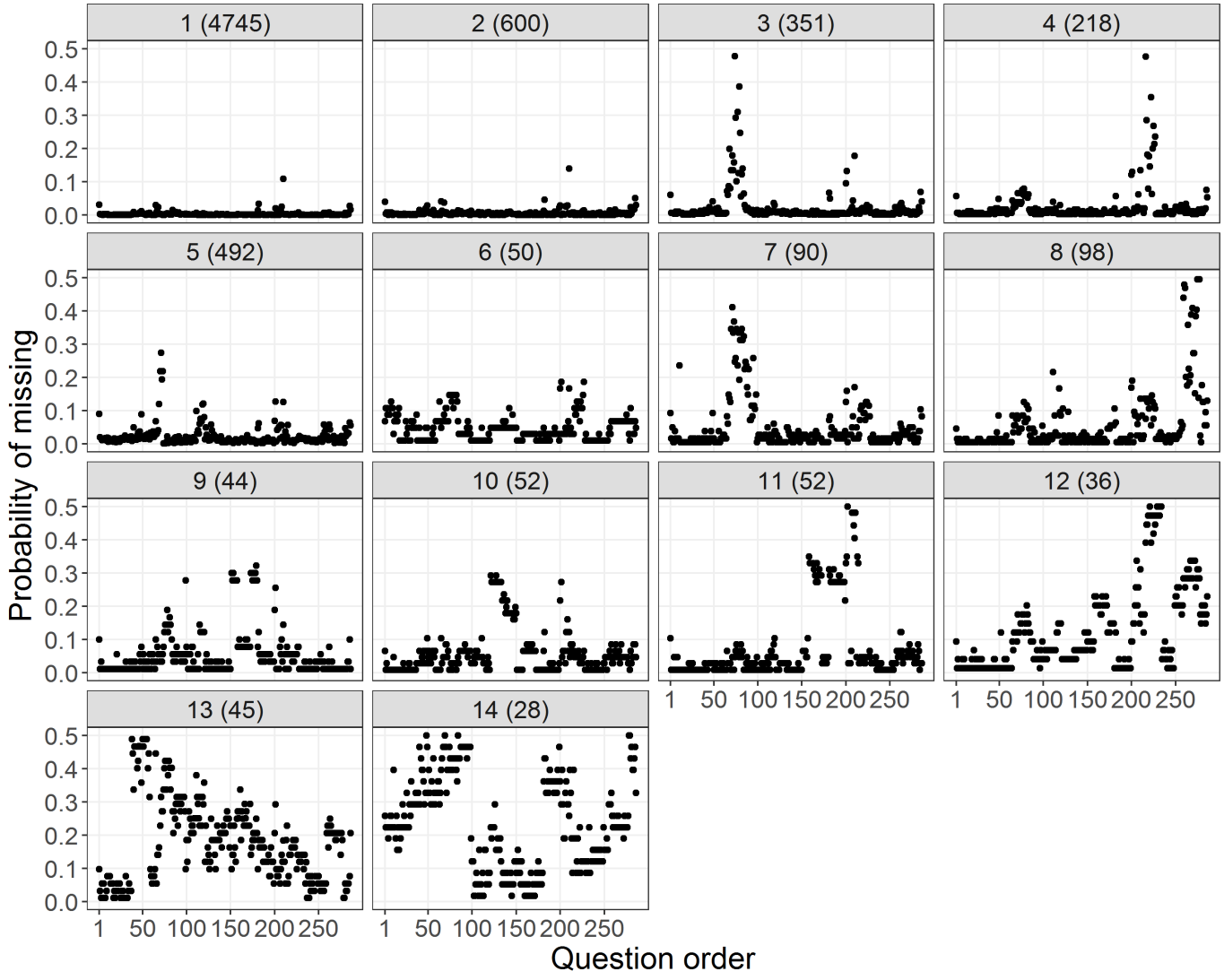

Figure 4 Estimated mean probability of missing by question order for 14 latent groups. Panel headings show the latent group number and sample size (total sample size=6901). Groups are ordered by size.

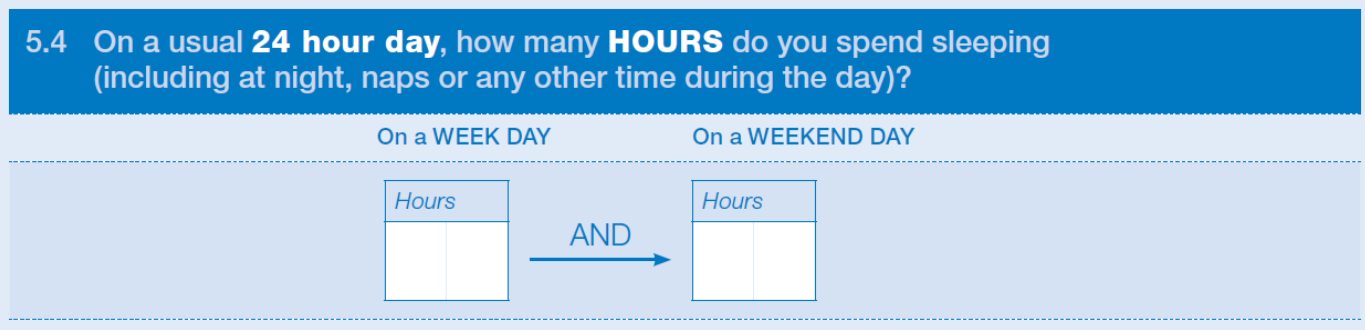

Figure 5 Two questions on sleeping (weekday and weekend day) as they appeared in the 2011 How Areas in Brisbane Influence healTh And acTivity survey.

Related methods and extensions

Latent classes have previously been used to identify patterns of wave missing in longitudinal data[24] and to identify participants who did not complete a question because it was not relevant.[25] Classification trees have been used to identify patterns in participant characteristics using the count of item missing per participant.[26]

An initial dimension reduction step may have been useful, such as principal components analysis, to combine questions with a similar pattern of missing.[27] However, this

may have combined questions with an interesting difference in the pattern of missing between two latent groups.

We originally used a Bayesian multinomial mixture model that gave very similar results to those shown here as it also selected a maximum of three latent groups (using the deviance information criterion) with similar patterns of missing to those in figure 2. However, the Bayesian model did not perform well in simulation studies with poor convergence using multiple MCMC chains which too often found divergent modes in the likelihood. This poor convergence occurred because of the complexities of label switching and because the model has hundreds of parameters.[28] One advantage of a Bayesian model was that a longitudinal structure could be added and we were able to track patterns of item missing over time in the same participants. The Rmixmod package used here can only model cross-sectional data.

We could have given more structure to the latent classes, for example, by making the participants' latent class membership dependent on variables such as gender using multinomial regression. More structure would also be provided by making the probability of missing dependent on question order using a spline for each latent group. Such structure could be included in a Bayesian model that accounts for repeated data from the same participant if the convergence issues could be addressed.

The methods shown here do not help identify data that is missing not at random where the probability of missingness depends on the missing value. Our methods also do not help adjust for the potential bias caused by missing data and researchers need to use other methods for imputing missing data or weighting observed data.[4 8]

CONCLUSION

Examining item missing data in detail has improved our understanding of how missing data were generated in these surveys. The latent groups were distinctly different and had characteristics of non-responders that make sense to the study team. Thinking about the types of responders will help us improve the design or instructions of future surveys and the wording and layout of questions. Our results also highlighted a potentially poorly presented question that could be changed in future surveys to hopefully produce fewer missing responses.

Twitter aidybarnett

Acknowledgements Computational resources and services used in this work were provided by the High Performance Computer and Research Support Group, Queensland University of Technology, Brisbane, Australia. We thank Dr Xing Lee for useful advice on using multinomial mixture models.

Contributors AGB had the original idea, ran the statistical analysis, wrote the first draft of the paper and is the study guarantor. GT and NWB critiqued the methods used, interpreted the results and commented on drafts of the paper. PM and AN helped collect, manage and clean the data, commented on the results and read drafts of the paper.

Funding The HABITAT project was awarded funding by the Australian National Health and Medical Research Council (NHMRC) (ID290521; ID497236; ID1047453). At the time the manuscript was written, GT was supported by an NHMRC Senior Research Fellowship (ID1003710). AGB is supported by an NHMRC Senior Research Fellowship (ID1117784).

Competing interests None declared.

Ethics approval The Queensland University of Technology Human Research Ethics Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All the data and R code are freely available on GitHub (https://github.com/agbarnett/item.missing).

REFERENCES

1. Kalton G. Compensating for missing survey data. Research report series. Survey Research Center, Institute for Social Research: the University of Michigan, 1983.
2. Hedeker D, Gibbons R. *Longitudinal data analysis. Wiley series in probability and statistics*: Wiley, 2006.
3. de Leeuw ED. Reducing missing data in surveys: an overview of methods. *Quality and Quantity* 2001;35:147–60.
4. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920–31.
5. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
6. Curran D, Bacchi M, Schmitz SF, *et al*. Identifying the types of missingness in quality of life data from clinical trials. *Stat Med* 1998;17:739–56.
7. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Stat Med* 2007;26:3057–77.
8. Rubin DB, Little RJA. *Wiley series in probability and statistics: statistical analysis with missing data*. 2nd Edition: Wiley, 2014.
9. Ys S, Gelman A, Hill J, *et al*. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Softw* 2011;45:1–31.
10. Hedeker D, Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol Methods* 1997;2:64–78.
11. Burton NW, Haynes M, Wilson LA, *et al*. HABITAT: A longitudinal multilevel study of physical activity change in mid-aged adults. *BMC Public Health* 2009;9:76.
12. Bergé L, Bouveyron C, Girard S. HDclassif : an *R* package for model-based clustering and discriminant analysis of high-dimensional data. *J Stat Softw* 2012;46:1–29.
13. Lebret R, Iovleff S, Langrognet F, *et al*. Rmixmod : the *R* package of the model-based unsupervised, supervised, and semi-supervised classification mixmodlibrary. *J Stat Softw* 2015;67:1–29.
14. Burnham K, Anderson D. *Model selection and inference: A practical information-theoretic approach*. New York: Springer, 2013.
15. Langrognet F, Lebret R, Poli C, *et al*. *Rmixmod: supervised, unsupervised, semi-supervised classification with MIXture MODelling (Interface of MIXMOD Software)*, 2016. R package version 2.1.1.
16. Wand M, Jones M. *Kernel smoothing. Chapman & Hall/CRC monographs on statistics & applied probability*: Taylor & Francis, 1994.
17. Burton NW, Turrell G, Oldenburg B. Item nonresponse in a population-based mail survey of physical activity. *JPAH* 2004;1:344–62.
18. Lunn D, Jackson C, Best N, *et al*; *The BUGS book: A practical introduction to bayesian analysis. chapman & Hall/CRC texts in statistical science*: Taylor & Francis, 2012.
19. Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. *Occup Environ Med* 2008;65:501–6.
20. Bates D, Mächler M, Bolker B, *et al*. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67:1–48.
21. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964–74.
22. Sahlqvist S, Song Y, Bull F, *et al*. Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. *BMC Med Res Methodol* 2011;11:62.

23. Hogan JW, Roy J, Korkontzelou C. Handling drop-out in longitudinal studies. *Stat Med* 2004;23:1455–97.
24. Lin H, McCulloch CE, Rosenheck RA. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* 2004;60:295–305.
25. Spycher BD, Minder CE, Kuehni CE. Multivariate modelling of responses to conditional items: New possibilities for latent class analysis. *Stat Med* 2009;28:1927–39.
26. Tierney NJ, Harden FA, Harden MJ, *et al*. Using decision trees to understand structure in missing data. *BMJ Open* 2015;5:e007450.
27. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: A review. *Comput Stat Data Anal* 2014;71:52–78.
28. Jasra A, Holmes CC, Stephens DA. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science* 2005;20:50–67.