DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 287–289, http://dx.doi.org/10.1515/JOS-2016-0014

Discussion

James O. Chipperfield[1]

The traditional survey paradigm has been to collect *all* variables from *all* respondents. This paradigm, which results in the well-known Single Phase Design (SPD), is being challenged by recent trends. These trends include: a decrease in response rates; a demand for more information to be collected, as analysts become more sophisticated; availability of inexpensive, but perhaps less-reliable, secondary sources of data (e.g., administrative data, the internet or Big Data) which in turn may be used as a substitute for survey data; an increase in the burden that may be imposed on respondents (e.g., require medical procedures, such as taking blood samples); and widespread use of computer-assisted interviewing that allows flexible sequencing of respondents through a questionnaire (e.g., household screening in order to target units of special interest or rare subpopulations); and increase in costs of administering surveys.

In response to these trends, survey organisations are looking to alternative, more flexible and efficient survey paradigms such as Split Questionnaire Designs (SQDs). SQDs relax the constraint of "collecting *all* variables from *all* respondents", which in turn allows more flexible ways of redressing these current trends. The article of Ioannidis et al. is therefore a timely and welcome contribution. I enjoyed reading it and I hope many more like it will follow.

While current trends do encourage the use of SQDs, the idea of SQDs is not new. For about the last 15 years, reviews by statistical agencies of their survey data collection strategies have recommended use of SQDs in some form, citing many of the benefits noted in the introduction of Ioannidis et al. So why are SQDs still not standard practice? Perhaps it is because SQDs present new and difficult methodological problems that can significantly increase the complexity of the 'survey cycle' and require substantial investment in new systems. The intricacy of the SQD design problem discussed by Ioannidis et al. is a case in point! So perhaps much more methodological work, some of which is discussed below, is required before SQDs are standard practice.

I now turn to comment on Ioannidis et al. The article is about optimal sample design. The objectives of the sample design are described in the traditional way of balancing accuracy and cost. However, what is far from traditional is that the authors consider the optimal sample design for an SQD rather than an SPD. The optimal SQD is defined in terms of $\mathbf{n} = (n_1, \ldots n_j \ldots, n_k)$, where n_j is the number of respondents assigned instrument j, an instrument is made up of a selection of questionnaire modules, and k is the

[1] Associate Professor (Adjunct), National Institute for Applied Statistics Research, University of Wollongong NSW 2522, Australia. Email: james.chipperfield@abs.gov.au

total number of instruments used in the design. I would like to make a few comments about the set-up of the sample design problem.

First, the set-up assumes that the Horvitz-Thompson (HT) estimator is used for estimation. This estimator does not exploit correlations between variables collected in different modules. Exploiting these correlations to improve the accuracy of estimates, whether using a model-based likelihood approach (Rubin and Little 2002) or by using a finite sampling model-assisted approach (Merkouris 2004), could perhaps be factored into the design problem. For example, consider the situation whereby Modules A and B contain variables that have a known and high correlation. Collecting Module A but not Module B from a respondent would contribute, due to the correlations, would also contribute, a non-zero amount to the effective sample size of Module B.

Second, traditional survey designs have almost exclusively been designed for estimating means or totals. Analysts interested in model-fitting are often called *secondary analysts*, because they are not the primary consideration during the survey design process. This is perhaps because, given the wide variety of possible analyses, designing for analysts' requirements is difficult. Nevertheless, traditional survey designs have historically met the needs of analysts for two reasons: (1) all modules are collected from all respondents meaning there is no loss of information about interactions between variables collected by different modules; and (2) the sample size for accurate estimates of subpopulation means is sufficient for accurate estimates of model parameters, where subpopulation is often treated as a marginal effect. However, in the case of SQD these reasons may not apply. For example, if an SQD only collects two out of five modules from any respondent, then information about two-way interactions would be available but no information about three-way (or higher) interactions would be available. So the SQD design problem may need to explicitly take into account the needs of analysts. While Ioannidis et al acknowledge the needs of analysts via 'enforcing crossings', I wonder whether measures of accuracy for a broad class of analysis could be incorporated into the design, as they are for population means.

Third, instruments are assigned to respondents independently of their characteristics. This means data not collected by the SQD are Missing Completely At Random (MCAR). We could instead assign instruments to respondents with a probability that depends upon the respondent's characteristics. This means the data not collected by the SQD are Missing at Random (MAR). Chipperfield et al. (2013) considered assigning instruments to respondents with a probability that depended upon the respondent's diabetes status collected during the interview (diabetes effects about 5% of people in the Australian state of NSW). In a logistic model with diabetes as the outcome variable, a person *with* diabetes contributes about the same amount of *information* (in a likelihood sense) as 400 people *without* diabetes. So given diabetes status, collecting the model's covariates from people *with* diabetes is much more efficient than collecting them from people *without* diabetes.

It is also worth mentioning search algorithms for the optimal SQD. When Chipperfield and Steel (2009, 2011) and Chipperfield et al. (2013) search for the optimal SQD they do not impose a constraint on the set of instruments (i.e., combination of modules). In other words, they allow *all* $k = 2^m - 1$ possible instruments to be used in the optimal design, where m is the number of modules. However, this is computationally infeasible even for moderate m. Ioannidis et al avoids this computational problem by considering only a

limited set of instruments, denoted by the matrix **A**, at each iteration (i.e., $k \ll 2^m - 1$). Across iterations, the algorithm searches for the optimal set of instruments. So Ioannidis et al. optimises over both **A** and **n**, and allows k to be set at the design stage rather than determined by the value m. This is a very useful development for moderate and large m.

In conclusion, it is hard to ignore that administrative data and Big Data will shape the way official agencies collect data in the future. I can see a role for an MAR-SQD whereby a respondent is assigned each module with a particular probability, where this probability depends on the information that is known about them from an administrative source. For example, if a person's health record shows that they are an unusually high user of medicines given their demographic characteristics, they may be more likely to be given a 'health' module. Their response values to the health module may affect the probability that they are given an 'education' module, and so on. These probabilities could be set to improve the efficiency of the SQD.

References

Chipperfield, J.O. and D.G. Steel. 2009. "Design and Estimation for Split Questionnaire Designs." *Journal of Official Statistics* 25: 227–244.

Chipperfield, J.O. and D.G. Steel. 2011. "Efficiency of Split Questionnaire Surveys." *Journal of Statistical Planning and Inference* 141: 1925–1932. Doi: http://dx.doi.org/10.1016/j.jspi.2010.12.003.

Chipperfield, J.O., M. Barr, and D.G. Steel. 2013. "Split Questionnaire Designs: Are They an efficient Design Choice?". In Proceedings of the 59th ISI World Statistics Congress, 25–30 August 2013. 311–316. Hong Kong. Available at: http://2013.isiproceedings.org/Files/IPS033-P1-S.pdf (accessed February 2016).

Merkouris, T. 2004. "Combining Independent Regression Estimators from Multiple Surveys." *Journal of the American Statistical Association* 99: 1131–1139. Doi: http://dx.doi.org/10.1198/016214504000000601.

Rubin, D.B. and R.J.A. Little. 2002. *Statistical Analysis of Missing Data* (2nd Edition). John Wiley and Sons.