Discussion of "Calibrated Bayes, for Statistics in General, and Missing Data in Particular" by R. J. A. Little

Nathaniel Schenker

It is a pleasure and an honor for me to comment on this article by Rod Little, who has contributed greatly to statistics in general and to Bayesian statistics and handling missing data in particular. Little provides a nice discussion of the calibrated Bayes approach, methods for missing-data problems and recent developments (SRMI and PSPP) that increase flexibility in dealing with missing data.

1. DON'T FORGET THE PRAGMATISTS

Little begins his Section 2 by stating that the statistics world is still largely divided into frequentists and Bayesians. Indeed, during the University of Maryland workshop ("Bayesian Methods that Frequentists Should Know") at which Little presented a talk on the topic of his article, many of the speakers declared themselves to be either frequentists or Bayesians. As formal discussant of Little's talk, however, I declared myself to be a "pragmatist," which Little (2006) defined as one who does not have an overarching philosophy and picks and chooses what seems to work for the problem at hand. If I were forced to choose a philosophy, I would probably go with the Bayesian one. But I am happy to use either approach, depending on the context, and many of my statistical colleagues seem willing to use either approach as well. Moreover, although subject-matter specialists with whom I work seem to be primarily familiar with point estimates, standard errors and confidence intervals, they seem to have no problems using Bayesian analogues (e.g., posterior means, standard deviations and credibility intervals) in the same way, when presented with them.

Little (2006) argued that, to enhance the credibility of our profession and avoid confusion and ambiguity,

Nathaniel Schenker is Associate Director for Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, Maryland 20782, USA (e-mail: nschenker@cdc.gov)

it would be preferable not to have the "split personality" that is inherent in the pragmatic approach. He has made a strong case in that article and here for calibrated Bayes as a unified inferential approach that combines strengths of the Bayesian and frequentist approaches. His arguments are compelling, but given the abundance of good and easily accessible frequentist methods that exist and are widely used, I imagine that it would be difficult for our profession to rid itself of this split personality. Moreover, I think the key issue in most applications is the development of realistic models for the data. Thus, I second Little's emphasis on flexible models and methods, such as the SRMI and PSPP methods, and his concluding call for further work on model diagnostics, especially in the area of missing data.

2. THE FREQUENTIST/BAYESIAN SCHISM IS PERHAPS MAGNIFIED IN SURVEY SAMPLING

In the survey sampling world in which I primarily work as a government statistician, the definition of being a frequentist versus being a Bayesian is not necessarily clear, because inferences are often desired about finite-population quantities rather than about model parameters. Such inferences are often made using a design-based paradigm (e.g., Cochran, 1977), that is, based on the distribution of estimators in repeated sampling from the finite population under a given design. Thus, one possible definition of frequentist inference in survey sampling is that it treats the finite-population values, Y, as fixed parameters, and bases inferences about a function of those parameters, say, $Q(Y)$, on a function of the sampled values and its distribution in repeated sampling. The corresponding definition of Bayesian inference (e.g., Rubin, 1987, Chapter 2) is that it places a prior distribution on Y, say, $p(Y|\theta)$, where θ represents hyperparameters with a hyperprior $p(\theta)$, and bases inferences on the posterior predictive distribution of $Q(Y)$ given the sampled values.

The two-by-two table (Table 1) gives a simplified, nonexhaustive depiction of the frequentist/Bayesian dichotomy within survey sampling on the one hand and

TABLE 1
Simplified, nonexhaustive 2-way table depicting the frequentist/Bayesian dichotomy within
survey sampling and in many areas outside of survey sampling

Mode of inference	Within survey sampling	In many areas outside of survey sampling
Frequentist	• Estimate $Q(Y)$ by $\hat{Q}(Y_{inc}, I)$, where Y = population values, Y_{inc} = sampled values, and I = indicators of inclusion in the sample.	• Formulate $p(y\|\theta)$, where y = data and θ = parameters.
	• Base inferences for $Q(Y)$ on $p(\hat{Q}(Y_{inc}, I)\|Y)$ induced by the distribution of the indicators I in repeated sampling, $p(I\|Y)$.	• Base inferences for θ on $p(\hat{\theta}(y)\|\theta)$, where $\hat{\theta}(y)$ estimates θ.
Bayesian	• Formulate $p(Y\|\theta)$ and $p(\theta)$, in addition to $p(I\|Y)$.	• Formulate $p(y\|\theta)$ and $p(\theta)$.
	• Base inferences for $Q(Y)$ on $p(Q(Y)\|Y_{inc}, I)$.	• Base inferences for θ on $p(\theta\|y)$.

in many areas outside of survey sampling on the other. As Table 1 shows, both within and outside of survey sampling, there are differences between the frequentist and Bayesian approaches concerning which quantities are treated as random, as well as whether prior distributions are specified. However, within survey sampling, there is an additional distinction, which is perhaps the most important in practice. The reference distribution for inferences under the frequentist, or design-based approach, is not induced by a model for the finite-population values, Y, whereas the Bayesian posterior predictive distribution does involve such a model.

Much has been written on design-based versus model-based inference in sample surveys, but I would particularly like to cite Hansen, Madow and Tepping (1983) and Little (2004). Hansen, Madow and Tepping (1983) concluded basically that for descriptive inference from reasonably large, well-designed sample surveys, design-based inference is to be preferred, because it avoids errors due to model misspecification that are possible with model-based inference, and because it loses little efficiency relative to model-based inference. They acknowledged, however, that model-based methods for sample surveys can be useful and important in the contexts of sample design, inference for small samples, inference in the presence of nonsampling errors, and situations in which inferences under a model are of intrinsic interest. One issue regarding the conclusions of Hansen, Madow and Tepping (1983) is that it is not always clear how large a sample is large enough. Moreover, lately there has been increasing interest in "pushing the data as far as possible," for example, by using a national survey to obtain estimates for a small subpopulation.

Little (2004) concluded that the Bayesian paradigm is flexible enough to provide practical and useful inferences in the context of survey sampling. He pointed out that the models used in Bayesian inference for surveys need to properly reflect features of the sample design, such as weighting, stratification and clustering, or else inferences are likely to be distorted. Similar points were made in the discussions of Hansen, Madow and Tepping (1983), in particular, those by Rubin, who clarified the role of the probabilities of selection in Bayesian modeling for sample surveys, and Little, who advocated the use of model-based estimators that are design-consistent. Hansen, Madow and Tepping (1983) agreed with those points in their rejoinder. However, as I will discuss in Section 4.3 in the context of applications to be presented in the next section, reflecting sample design features can be complicated in some problems for which model-based inference can be particularly useful. Thus, I believe that further development of methods for reflecting design features will be an important area of research.

3. A MAJOR REASON WHY THIS PRAGMATIST LIKES BAYESIAN METHODS

From a pragmatic point of view, one of the major attractions of Bayesian methods is their ability to handle problems with complex data structures such as missing data in a relatively straightforward manner. As Little points out, this has been true especially since the development of Markov chain Monte Carlo methods and multiple imputation. To complement Little's discussion and to illustrate some of his points, I will now describe a few applied projects for which Bayesian techniques were very helpful.

3.1 Survival Analysis with Intermittently Observed Covariates

Faucett, Schenker and Elashoff (1998) analyzed the relationship between post-operative smoking and sur-

vival using data from a clinical trial on survival of patients after surgery for lung cancer. At follow-up visits, the patients had been asked about their current smoking status. Faucett, Schenker and Elashoff (1998) discretized time using narrow time intervals, and they specified a Markov chain model for current smoking status together with a time-dependent proportional hazards model with a piecewise constant baseline hazard for survival given smoking behavior and covariates. Gibbs sampling was used to approximate the joint posterior distribution of the model parameters under diffuse prior distributions.

The use of Gibbs sampling facilitated analyses under two different survival models, one with current smoking as the time-dependent covariate and another with cumulative smoking as the time-dependent covariate. It was found that the coefficient for cumulative smoking (in the latter model) had much more posterior probability mass to one side of zero than did the coefficient for current smoking (in the former model). Thus, the evidence was stronger for a detrimental effect of cumulative smoking than for a detrimental effect of current smoking. The application of Faucett, Schenker and Elashoff (1998) is an example of joint modeling of longitudinal and survival data, which has been a popular area of research in the past decade.

3.2 Incorporating Auxiliary Variables into Survival Analysis via Multiple Imputation

In a different type of application that jointly modeled longitudinal and survival data, Faucett, Schenker and Taylor (2002) developed an approach, based on multiple imputation, to using auxiliary variables to recover information from censored observations in survival analysis. Applications of this type are mentioned by Little in his Section 5, point (a) and Section 7. Faucett, Schenker and Taylor (2002) analyzed data from an AIDS clinical trial comparing zidovudine and placebo, in which the outcome of interest was the time to development of AIDS, and in which CD4 count was a time-dependent auxiliary variable. Because AIDS can take a long time to develop, most of the observations were censored. Faucett, Schenker and Taylor (2002) specified a hierarchical change-point model for CD4 counts and a time-dependent proportional hazards model for the time to AIDS given CD4 and covariates. Markov chain Monte Carlo methods were then used to multiply impute event times for the censored cases.

The use of multiple imputation facilitated drawing inferences about quantities whose posterior distributions could not be approximated directly using the output of the Markov chain Monte Carlo simulations. For example, Kaplan–Meier estimates of survival under treatment and placebo were compared, and the coefficient of treatment in a Cox regression analysis was examined as well. Comparisons with analyses of the censored data without imputation, and accompanying simulation results, suggested that incorporating the auxiliary variables via multiple imputation can lead to improved efficiency as well as partial corrections for dependent censoring. This application illustrated use of a nonBayesian complete-data analysis with multiple imputation; see Little's Section 5, point (c).

3.3 Multiple Imputation for Missing Data in Surveys

As Little discusses in Section 5, points (a) and (b), multiple imputation has particular benefits in the context of public-use data. SRMI was used recently in two major applications of multiple imputation to public-use data from the National Center for Health Statistics. One involved missing income data in the National Health Interview Survey (NHIS) (Schenker et al., 2006), and the other involved missing body-scan data from dual-energy X-ray absorptiometry (DXA) in the National Health and Nutrition Examination Survey (NHANES) (Schenker et al., 2011). DXA scans are used to measure body composition such as soft tissue composition and bone mineral content. Public-use data with multiple imputations from both applications have been released online (http://www.cdc.gov/nchs/nhis/2009imputedincome.htm; http://www.cdc.gov/nchs/nhanes/dxx/dxa.htm).

Both applications involved nontrivial amounts of missing data—roughly 30% for the NHIS income data and 20% for the NHANES DXA data—with missingness related to characteristics of the persons surveyed, so that analysis of only the complete cases would likely result in biases as well as inefficiencies. The use of SRMI facilitated inclusion of large numbers of predictors of different types (e.g., categorical, continuous, count) in each application, with some of the predictors having missing data themselves, although usually at much lower levels than the main variables of interest. As discussed in Meng (1994), Rubin (1996), Little and Raghunathan (1997) and Reiter, Raghunathan and Kinney (2006), the reasons for including several predictors in imputation models, besides of course to help predict the missing values, are to help "explain" the missingness, that is, to make the assumption of missingness at random more tenable, and to promote compatibility between the imputation models and the analyses that

would ultimately be carried out by secondary users of the data. The issue of possible incompatibility is discussed further in Section 4.4.

Each application had other interesting features, some of which were handled especially well by SRMI. For example, in the NHIS project, for the majority of the missing family income values, respondents had provided coarse income categories, so that bounds were available for the missing values. Also, there were sometimes structural dependencies between variables that needed to be imputed. For example, a person could not have earnings unless he/she was employed, and occasionally, employment status was missing along with earnings.

In the NHANES project, the missing data were highly multivariate. There were 32 DXA variables, some of which were highly interrelated; and sometimes the DXA data were only partially missing. Perhaps the most interesting feature of the project, however, was that missingness of DXA data often occurred for people with high levels of truncal adiposity, because the adiposity interfered with the ability to obtain valid measurements. Thus, the levels of missingness tended to be high at the largest values of other variables measured in the NHANES, such as BMI and waist circumference. This necessitated some extrapolation beyond the range of the observed DXA values.

3.4 Combining Information from Two Surveys to Enhance Small-Area Estimation

A multi-organization project led by the National Cancer Institute used Bayesian methods to compute small-area estimates of the prevalence of cancer risk factors and cancer screening by combining information from two surveys for the years 1997–2003 (Raghunathan et al., 2007; Davis et al., 2010). The surveys were the Behavioral Risk Factor Surveillance System (BRFSS), a large, state-based survey conducted by telephone, and the NHIS, a smaller, face-to-face survey. The BRFSS included most of the counties in the United States in its sample and thus provided some direct information about them. However, it obtained data only from households equipped with telephones, and its nonresponse rates tended to be relatively high, as is often the case with telephone surveys. The NHIS surveyed both telephone and nontelephone households, asked a question to identify the telephone status of the household, and generally had lower nonresponse rates than the BRFSS. However, its sample only included about 25 percent of the counties.

A Bayesian, trivariate extension of the Fay–Herriot (1979) model was formulated. Markov chain Monte Carlo methods were used, together with county-level telephone coverage rates from the 2000 census, to approximate the posterior distributions of the small-area rates. Estimates from the project have been released publicly (http://sae.cancer.gov/).

4. SOME AREAS FOR FURTHER RESEARCH

4.1 Flexible Models and Methods

In Section 1 I mentioned the need for more flexible models and methods. SRMI and PSPP are two examples of techniques that have increased flexibility (see, e.g., Section 3.3 for examples in which SRMI was used), and the development of more such techniques would be welcome. For example, perhaps a flexible univariate prediction model such as PSPP could be used for each univariate regression in SRMI to develop a robust procedure for multivariate imputation.

4.2 Diagnostics for Models

In Section 1 I also seconded Little's call for work in the area of model checking, especially for missing-data problems. With missing data, checking prediction models for the missing values is especially difficult, for the obvious reason that the missing values are unavailable for use in model checking. Diagnostics for imputations of the general types mentioned in Abayomi, Gelman and Levy (2008) were used in the NHANES multiple-imputation project discussed in Section 3.3.

Little (Section 2) mentions methods such as posterior predictive checks as being frequentist in spirit. It would be helpful to investigate more fully the link between use of such techniques and achieving well-calibrated analyses, such as Bayesian credibility intervals with good frequentist coverage properties. Also useful for survey practitioners would be more research on evaluating models from a design-based point of view, especially in the context of complex sample designs.

4.3 Incorporating Complex Sample Design Features into Models

As I mentioned in Section 2, incorporating complex sample design features into models for survey data can be complicated. In the context of multiple imputation, inclusion of survey weights and indicator variables for strata and primary sampling units (PSUs) has been advocated (Rubin, 1996; Reiter, Raghunathan and Kinney, 2006). Such techniques were used in the NHIS

and NHANES multiple-imputation projects described in Section 3.3 above, although in the NHANES project, there was some concern about parsimony, so a smaller number of variables related to PSU selection were substituted for the full set of indicator variables. Further work on methods for increasing parsimony, such as via use of random effects, would be helpful.

In addition, incorporating complex sample design features can be difficult in problems that involve combining information across surveys, because the design features of the two surveys might not be comparable. This was one reason for using an area-level (Fay–Herriot) rather than person-level model in the small-area estimation project discussed in Section 3.4; see Schenker and Raghunathan (2008). Schenker, Raghunathan and Bondarenko (2010) also discussed such issues in the context of using multiple-imputation to combine information from two surveys.

4.4 Impacts of Secondary Analysts Using Variables not Included in the Imputation Model

Little notes (Section 5) that an attractive feature of multiple imputation is that the imputation model can include variables not included in the final analysis. I agree with this, and, furthermore, I have found multiple imputation to be a very general and flexible method for allowing secondary analysts of public-use data to assess the uncertainty due to imputation.

A concern of mine, which applies to single imputation as well as multiple imputation, is biases that can occur in point estimates of interest when a secondary analyst uses the imputed data together with variables that were not included in the imputation model. As mentioned in Section 3.3, the NHIS and NHANES projects used large numbers of predictors in order to avoid such incompatibilities, and the predictors were listed for secondary analysts in the technical documentation for the projects. However, it is likely in general that some secondary analysts of public-use data will attempt analyses that "go beyond" the imputation model. The biases in point estimates for such analyses will depend in a sense on how well the variables included in the imputation model account for the relations being studied in the secondary analysis. Further research on the possible extent of such biases, and guidelines and diagnostics for secondary analysts, would be useful areas for research.

4.5 Real-Life Examples of the Utility of the Calibrated Bayes Approach

As I mentioned in Section 1, I imagine that it would be difficult to move our field completely away from having a "split personality" and toward following Little's (2006) "Bayes/Frequentist Roadmap." Excellent papers such as Little's current one will provide nudges in that direction. Also helpful will be more real-life examples of how the calibrated Bayes approach can help to achieve substantial gains in solving problems that could not be achieved otherwise.

REFERENCES

ABAYOMI, K., GELMAN, A. and LEVY, M. (2008). Diagnostics for multivariate imputations. *J. Roy. Statist. Soc. Ser. C* **57** 273–291. MR2440009

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

DAVIS, W. W., PARSONS, V. L., XIE, D., SCHENKER, N., TOWN, M., RAGHUNATHAN, T. E. and FEUER, E. J. (2010). State-based estimates of mammography screening rates based on information from two health surveys. *Public Health Rep.* **125** 567–578.

FAUCETT, C. L., SCHENKER, N. and ELASHOFF, R. M. (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates. *J. Amer. Statist. Assoc.* **93** 427–437.

FAUCETT, C. L., SCHENKER, N. and TAYLOR, J. M. G. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58** 37–47. MR1891041

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

HANSEN, M. H., MADOW, W. G. and TEPPING, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **78** 776–807.

LITTLE, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Amer. Statist. Assoc.* **99** 546–556. MR2109316

LITTLE, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* **60** 213–223. MR2246754

LITTLE, R. J. A. and RAGHUNATHAN, T. E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods* 617–622. Amer. Statist. Assoc., Alexandria, VA.

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion and rejoinder). *Statist. Sci.* **9** 538–573.

RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Amer. Statist. Assoc.* **102** 474–486. MR2370848

REITER, J. P., RAGHUNATHAN, T. E. and KINNEY, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Surv. Methodol.* **32** 143–149.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1996). Multiple imputation after 18+ years (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **91** 473–489, 507–515, 515–517.

SCHENKER, N. BORRUD, L. G. BURT, V. L. CURTIN, L. R. FLEGAL, K. M. HUGHES, J. JOHNSON, C. L. LOOKER, A. C. and MIREL, L. (2011). Multiple imputation of missing dual-energy X-ray absorptiometry data in the National Health and Nutrition Examination Survey. *Stat. Med.* **30** 260–276.

SCHENKER, N. and RAGHUNATHAN, T. E. (2008). Discussion of "Assessing the value of Bayesian methods for inference about finite population quantities" by Sedransk. *J. Official Statist.* **24** 507–512.

SCHENKER, N., RAGHUNATHAN, T. E. and BONDARENKO, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Stat. Med.* **29** 533–545.

SCHENKER, N., RAGHUNATHAN, T. E., CHIU, P.-L., MAKUC, D. M., ZHANG, G. and COHEN, A. J. (2006). Multiple imputation of missing income data in the National Health interview survey. *J. Amer. Statist. Assoc.* **101** 924–933. MR2324093