Journal of Official Statistics, Vol. 20, No. 2, 2004, pp. 265–287

Relating Respondent-Generated Intervals Questionnaire Design to Survey Accuracy and Response Rate

S. James Press[1] *and Judith M. Tanur*[2]

This article is concerned with relating questionnaire design to estimation accuracy and item response rate in sample surveys, in the context of the Respondent-Generated Intervals (RGI) protocol for asking questions. One RGI procedure for asking survey questions is concerned with recall of facts ("How many times did you visit your doctor in the last year?"). The research addresses the problem that respondents' unequal memory abilities may lead to large nonsampling errors (bias). The novelty of this question protocol is asking respondents both for an answer to the recall question and also for the smallest and largest possible values they think the true answer might be. We find that a Bayesian estimator of the population mean is given by a weighted average of the basic responses, where the weights assigned to respondents' estimates are larger for smaller interval lengths. We summarize four record-check surveys for which the RGI protocol has been applied. We find that interval length is related to the respondent's confidence in his/her answer, and that fine-tuning the way the question is worded is directly related to the response rate, and to the accuracy of population parameter estimates. So by placing strong emphasis upon the questionnaire design we can improve the importance and usefulness of the survey.

Key words: Bayesian; brackets; questionnaires; recall; item nonresponse; surveys.

1. Introduction

We are concerned in this article with relating questionnaire design to survey accuracy and response rate. We take up these issues with respect to a new survey questioning protocol called Respondent-Generated Intervals (RGI) (see Press 1996; 1999; 2002; 2004). In this procedure we ask respondents to provide a basic answer (we call the basic answers the *usage quantities*, since they often refer to the frequency of usage of some behavior) to a question involving recall of a factual question. But we also ask the respondent to provide a lower and an upper bound to where he or she thinks the true value to the question lies. For example, we might ask, "How many times did you visit your doctor last year?" The respondent might then answer, say, six times. But then the respondent might also indicate that the true value is surely no larger than seven times, but also, it is surely larger than four times. So the respondent has now also provided a coverage interval for his/her response.

[1] University of California, Riverside, Department of Statistics, Riverside, CA 92521-0138, U.S.A. Email: jpress@ucrac1.ucr.edu

[2] State University of New York, Stony Brook, New York, P O Box 280, Montauk, NY 11954-0202, U.S.A. Email: jtanur@notes.cc.sunysb.edu

Acknowledgments: This research was partially supported by a grant from the New York Community Trust. We are grateful to Dr. Diane Miller for providing the Minitab macros that simplified computing the numerical results required for the RGI Bayesian estimates presented in this article. For information about these macros contact her at diane.m.miller@ngc.com. We are also grateful to the anonymous referees who provided helpful comments and suggestions for improving the presentation.

It is assumed that the respondent knew the true value at some point (or had sufficient information to construct a correct answer) but because of imperfect recall, he or she is not certain of this value. This article addresses the important issues of how the population parameters should be estimated, the questionnaire design issues of how the bounding questions should be asked, and how the design of the questionnaire affects the estimation accuracy of population parameters and the item response rate.

The population parameter that is often of greatest interest in a survey is the population mean. We derive an estimator for this parameter by using a hierarchical Bayesian model. The derivation was originally developed in Press (2002), and for convenience, it is repeated in the appendix to this article. In Section 2 we discuss the characteristics of the Bayesian point estimator of the population mean. In Section 3 we relate the design of the questions to the properties of the estimator. In Section 4 we describe some experiments in which we (and others) have used the RGI protocol, and summarize some results from them. In Section 5 we discuss what we have learned from these experiments.

2. The Bayesian Point Estimator

2.1. Vague prior for the population mean

For a sample of n independent respondents in a survey, let y_i, a_i, b_i denote the basic usage quantity response, the lower bound response for where the true value to the question lies, and the upper bound response for where the true value to the question lies, respectively, of respondent i, $i = 1, \ldots, n$. Suppose that the y_i's are all normally distributed. Suppose also that we adopt a vague prior distribution for the population mean, θ_0, to represent knowing very little, a priori, about the value of the population mean. It is shown in the Appendix, using a hierarchical Bayesian model, that in such a situation, the posterior distribution of θ_0 is given by:

$$(\theta_0|data) \sim N(\tilde{\theta}, \omega^2) \tag{2.1}$$

where the posterior mean, $\tilde{\theta}$, is expressible as a weighted average of the y_i's, and the weights are dependent upon the intervals defined by the bounds, the smaller the interval the larger the weight. The posterior variance is denoted by ω^2. The posterior mean is expressible as:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i \tag{2.2}$$

where the λ_i's are weights that are given approximately by:

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(b_0 - a_0)^2}{k_2^2}} \right)}{\sum_{i=1}^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(b_0 - a_0)^2}{k_2^2}} \right)}, \quad \sum_1^n \lambda_i = 1, \tag{2.3}$$

where: $a_0 \equiv \min_{1 \le i \le n}(a_i)$; $b_0 \equiv \max_{1 \le i \le n}(b_i)$. The interval $(b_0 - a_0)$ represents the full range of opinions the n respondents have about the possible true values of their answers to the question, from the smallest lower bound to the largest upper bound. In Equation (2.3), k_1 and k_2 denote pre-assigned multiples of standard deviations that correspond to how the bounds should be interpreted in terms of standard deviations from the mean. For example, for normally distributed data it is sometimes assumed that such lower and upper bounds can be associated with 2 standard deviations below, and above, the mean, respectively. With this interpretation, we could take $k_1 = k_2 = 4$ to represent the length of the interval between the largest and smallest values the true value of the answer to the recall question might be for respondent i. If desired, we might take $k_1 = k_2 = k$, and then we would make a choice among reasonable values, such as: $k = 2, 4, 5, 6, 7, 8$, and study how the estimate of the population parameters vary with k. This issue relates to questionnaire design and is discussed further in Section 3.

2.2. *Normal prior for the population mean*

In some situations we are not entirely ignorant of the possible value of the population mean. We may have some preconceived notion of what this mean might be, even though we are still uncertain about its true value. For analysis of such situations where we can take advantage of such prior information, we have also studied the case of a normal prior distribution for the population mean, $\theta_0 : \theta_0 \sim N(\theta^*, \rho^2)$. In this case, it is shown in the Appendix that the posterior distribution for θ_0 becomes:

$$(\theta_0 | data) \sim N(g, \eta^2) \tag{2.4}$$

where for $\xi = \frac{\frac{1}{\omega^2}}{\frac{1}{\omega^2} + \frac{1}{\rho^2}}$, $0 \le \xi \le 1$, and $\tilde{\theta} = \sum_1^n \lambda_i y_i$

$$g = \xi\tilde{\theta} + (1 - \xi)\theta^* \tag{2.5}$$

and

$$\eta^2 = \frac{1}{\frac{1}{\rho^2} + \frac{1}{\omega^2}} \tag{2.6}$$

and the λ_i'*s* are the same weights as for the vague prior; $0 \le \lambda_i \le 1$, $\sum_1^n \lambda_i = 1$. The y_i'*s* are of course the data (usage quantities).

We note the following characteristics of these estimators:

(1) The weighted averages are simple and quick to calculate, without requiring any computer-intensive sampling techniques. The weighted average point estimator may be used nonparametrically, even when the data are not normally distributed, but interval estimation does require normality in small samples.

(2) In the special case in which the interval lengths in Equation (2.3) are all the same, the weighted average in Equation (2.2) reduces to the sample mean, $\bar{y}$, where the weights all equal $(1/n)$. The weighted average estimator will result in a more accurate estimate of the population mean than will the sample mean when some respondents who are accurate give short bounding intervals and the others who are inaccurate give longer bounding intervals.

(3) The longer the interval a respondent gives, the less weight is applied to that respondent's usage quantity in the weighted average. The length of respondent *i*'s interval is a measure of his/her degree of confidence in the usage quantity he/she gives, so that the shorter the interval, the greater degree of confidence that respondent seems to have in the usage quantity he/she reports. (Of course a high degree of confidence does not necessarily imply an answer close to the true value.)

(4) Since the weights sum to one, and must all be nonnegative, they can be thought of as a probability distribution over the values of the usage quantities in the sample. So λ_i represents the probability that $y = y_i$ in the posterior mean.

(5) We see that if we take $k_1 = k_2$ in Equation (2.3), the *k*'s cancel out and the λ_i weights no longer depend upon the *k*'s.

(6) If we define the precision of a distribution as its reciprocal variance, the quantity $\left\{\frac{(b_i-a_i)^2}{k_1^2} + \frac{(b_0-a_0)^2}{k_2^2}\right\}$ may be seen (from the analysis in the Appendix) to be the variance in the posterior distribution corresponding to respondent *i*, and therefore, its reciprocal represents the precision corresponding to respondent *i*. Summing over all respondents' precisions gives:

$$\text{total conditional posterior precision} = \sum_{1}^{n} \left(\frac{1}{\frac{(b_i-a_i)^2}{k_1^2} + \frac{(b_0-a_0)^2}{k_2^2}} \right) \qquad (2.7)$$

So another interpretation of λ_i is that it is the proportion of the total precision in the data attributable to respondent *i*.

3. The Bounding Questions and Their Relationship to Characteristics of the Estimators

In all surveys, the wording of a question strongly drives the estimation accuracy and response rate for that question and perhaps more broadly across the survey instrument. The relationships are usually quite subtle, however, and it is difficult to know how to determine the effects and implications of alternative wordings, though the movement to study cognitive aspects of survey methodology has offered us some principles (e.g., Schwarz and Sudman 1994; Sudman, Bradburn, and Schwarz 1996; Sirken et al. 1999; Tourangeau, Rips, and Rasinski 2000). In an RGI-based survey, there is a clear and overt relationship that can be separately studied to improve the effectiveness of the survey. We need to understand fully how the wording of the bounds question or questions affects respondents' interpretation of how broad the interval they supply should be.

This type of question relates closely with the literature about how to assess prior probability distributions. In Bayesian assessment procedures an entire prior distribution (and/or a utility function) for an individual is assessed by connecting a collection of points on the individual's subjective probability distribution obtained by means of a sequence of elicitation questions (see for example Schlaifer 1959, Ch. 6; and Hogarth 1980, Appendices B and C). Berry (1996, pp. 347–348) assumes the person's belief distribution is normally distributed, and that a person whose prior probability he is trying to assess (the respondent) "would not be very surprised" if there were a 10% chance that the true value exceeds a given stated amount (an upper bound).

In RGI there is a fundamental tension between the way we would ideally like to extract information from respondents and how questions can be asked so that ordinary respondents with no special training can understand them and answer appropriately. If we were to follow the approach typically taken by probability assessors to assess someone's prior distribution for some unknown quantity, such as the true value of some item a respondent is trying to recall, we might ask a sequence of questions ("ethically neutral propositions," to use the terminology of Ramsey 1931), such as:

(1) "Give a number such that it is equally likely that the true value is less than that number, and that the true value is larger than that number." Call the number given, B(0.5). B(0.5) is then the median of the respondent's recall distribution.

(2) "Next suppose I tell you that your true value is really less than B (0.5). Now give another number that is less than B (0.5), and such that it is equally likely that the true value is less than that new number, and the true value is larger than that new number." Call the number given now B (0.25). B (0.25) is the 25th percentile of the respondent's recall distribution.

(3) Now ask the analogous question first advising the respondent that his/her true value is actually larger than B (0.5). The number given now, B (0.75), is the 75th percentile of the respondent's recall distribution.

The three points just found, plus the fact that the respondent's recall cumulative distribution function (cdf) must be bounded by 0 and 1, give us 5 points that define the recall cdf quite well. We can now readily develop the corresponding probability density function. The resulting prior density could ideally now be combined with the likelihood to generate a posterior distribution from which we could estimate the population parameters of interest. The problem, of course, is that most respondents have not been specially trained to be able to address the above three questions with any dependable degree of cognitive ability. Some respondents might be totally confused when faced with such a task. While they might be able to deal with the first question, the remaining two questions would probably be very confusing. In spite of the fact that the answers to these three questions would provide the analyst with the precise information required, we need to formulate alternative ways of developing the required information that would be within the cognitive grasp of ordinary respondents, but would still provide sufficient information to the analyst so that at least a close approximation to the required information becomes available. We have found that a reasonable alternative is available by asking respondents for lower and upper bounds on their true values. While this alternative does provide analogous information to the analyst, it is not itself without some remaining interpretive difficulties, as explained below.

We saw above that under a vague prior on the population mean, the posterior mean estimator of the population mean is a weighted average of the usage quantities. We also saw that under a normal prior on the population mean, the posterior mean estimator of the population mean becomes a weighted average of the prior mean and the Bayesian estimator of the population mean under a vague prior. In fact, the weights in the posterior mean associated with the usage quantities depend explicitly on the intervals defined by the bounds on these quantities. In addition, the weights depend upon two *interpretation* *constants*, k_1,k_2, which relate to how the respondents interpret the bounds questions, or how we assume they do. For simplicity, we will be taking $k_1 = k_2 = k$ throughout this article.

We know that as long as the data are normally distributed, for example, it is unlikely that we would find more than 5% of the observations beyond 2 standard deviations away from the mean. But what does this mean to the typical respondent? Does it mean that the analyst would be safe in assuming that the bounds proffered by the respondent can be placed at plus and minus 2 standard deviations and conclude that the true value the respondent is being questioned about lies in an interval of length 4 standard deviations? In the first test of RGI, described below, we phrased the bounds question, e.g., "Please fill in the blanks," "There is almost no chance that the number of credits I had earned by the beginning of this quarter was less than ______ and almost no chance that it was more than _______." Assuming a normal distribution, we took the interval length given by a respondent to cover the middle 95% of the distribution and thus took $k = 4$. We seek to refine this process of assigning values to k.

Perhaps the way each bounds question is worded signals most respondents to give bounds that exclude just 1% of the chances of finding the true value in the associated interval; perhaps 3% or 5% or 10%. While we have studied the wording of the bounds questions, as detailed below, we have not yet obtained results that permit us to determine how to relate the interpretation of the length of the interval to the question wording. For purposes of this article we use normality and interpret the interval defined by the bounds as a 2 standard deviation interval (k = 4).

We are also concerned about the possibility that the respondent's recall distribution is not normally distributed. For example, suppose the respondent provides a bounding interval (a,b) for which the usage quantity given is close to one of the bounds. The respondent is clearly thinking in terms of some sort of asymmetric recall distribution, certainly not a normal distribution. In such a case we recommend a preliminary transformation of both the usage quantity and the bounds to approximate normality, such as the Box-Cox transformation (see Box and Cox 1964). There is a pull-down menu in MINITAB 13 that will carry out the transformation readily (under "Stat" → "Control Chart"). But all values being transformed must be positive.

For $k_1 = k_2 = k$, the value of a Bayesian point estimator does not depend upon k. Moreover, regardless of whether $k_1 = k_2$ or not, we see from Equation (A24) in the Appendix that the posterior variance of our estimator (and hence the lengths of credibility intervals) depends on the values assigned to k_1 and k_2, or to k. The higher the value assigned to $k = k_1 = k_2$, the smaller the posterior variance. Thus, much depends on what values we see as justifiable for k_1 and k_2.

4. Some Empirical Studies

In this section we describe four experiments we have carried out to explore the usefulness of the RGI protocol and one experiment carried out by others (Schwartz and Paulin 2000, at the U.S. Bureau of Labor Statistics). The first two experiments were carried out on two university campuses. The third was carried out together with the U.S. Census Bureau, and the last described here was carried out in conjunction with a Health Maintenance Organization (called the HMO experiment).

Both the theory and the estimation procedure for the RGI estimator have evolved during the years we have been studying the procedure. While we have published results using an earlier model than that detailed here, and a different estimation procedure (e.g., Press and Tanur 2000; Press and Marquis 2001), those results are now outdated. In addition, in the lapse of time since the data were collected some data availability has been lost. In particular, because of concerns about confidentiality, individual level verification data for the Census experiment were never released to us. With changing methods of estimation, differing sets of respondents are appropriate for analysis, but because of further concerns about confidentiality, the data set containing the verified values has been destroyed. Hence we can present no formal analysis of the Census data here, though we do present some results from the cognitive testing.

On a somewhat less dramatic note, we find that some of the data files for the campus data have been separated from the demographic information for the respondents that would allow us to assess a realistic proper prior for the population mean. Hence we report findings only for six variables of State University New York at Stony Brook (SUSB) and four at the University of California at Riverside (UCR).

The data collection for the HMO experiment has only recently been completed, so no analysis is possible for that as yet.

There are also a couple of new UCR campus experiments that are currently being fielded; those results will be reported at a later date.

4.1. The two campus surveys

Our first empirical effort was primarily to determine whether respondents were willing to use the new method in a paper-and-pencil survey, but also, it was to explore the accuracy that can be achieved, and to see if the technique reduced item nonresponse (see Press and Tanur 2000).

We carried out a paper-and-pencil survey in spring of 1997 at each of our campuses, UCR and SUSB. We asked students questions about the amounts of fees they paid, registration fee, recreation fee, student activities fee, and health fee, their SAT math and verbal scores, their number of on-campus traffic tickets, their number of library fines, their grade point average, their number of credits earned, their number of grades of C or less, and their expenditures on the food plan during the previous month. All of these quantities could be verified by the appropriate campus office and they were verified for those student respondents who gave permission for us to do so and who supplied an identification number that made such checking possible. For both of the campus surveys the usage question was always asked before the bounds question.

We were able to analyze the following ten items in terms of the accuracy of the RGI procedure for analyses not involving Bayesian estimation we were able to use the full set of 18 items (see e.g., Table 2).

For both campuses: GPA, SAT verbal and math scores, TICKETS.

For SUSB only: CREDITS, FINES.

Before considering the results, it will be useful if we describe in some detail how we derived the prior mean and standard deviation for the variable CREDITS at SUSB. (Far less detailed accounts will be given of how we derived the other prior means.)

Respondents had told us the year in school they were in. Of course, some people who said "first year" are in their first semester of their freshman year, some in their second, and similarly for other years. From conversations with undergraduate directors, we assumed that a student takes about 17 credits a semester for each semester of the first two years and 15 credits a semester for the third year. (We ignored 4th year, as it did not figure into our calculations because we had no 4th year respondents.) Thus we estimated that a first semester 2nd year student would have finished 34 credits, a second semester 2nd year student would have finished 51 credits, a first semester 3rd year student would have finished 66 credits and a second semester 3rd year student would have finished 81 credits. Taking into account that most students start school in the fall and that the survey was carried out in a spring semester, we estimated that 80% of the students were second semester in their respective years and 20% first semester. Thus we took the mean of the credits completed by those students reporting 2nd year status to be $.8*51 + .2*34 = 47.6$ and the mean of the students reporting 3rd year status as $.8*81 + .2*66 = 78$. Since the mean "year in school" reported by respondents who answered the credits question was 2.68 we took .68 * the difference between 78 and 47.6 and added that to 47.6 to get our prior mean for credits as 68.27 or 68.3.

We chose to set the prior standard deviation at 20% of the prior mean or 13.65. Three times 13.65 or 41 credits above and below our prior mean (27.3–109.3) seemed to encompass almost all likely numbers of credits students might have earned. It was because we did not have the year in school for respondents from UCR that we were unable to derive a prior mean and hence unable to use the current model for estimation.

We took a prior mean for $\text{GPA} = 2.5$, since a "C" average (2.0) is required to remain in school, and most students achieve a somewhat larger GPA. We took the standard deviation of the prior distribution to be 10% of the prior mean.

For TICKETS and FINES we assumed that many respondents would have no instances (either because they are law abiding or because they do not own a car or never borrow a book from the library). We understood that these "true zero" students would probably use a different strategy to derive their estimates than would those who had actual instances to retrieve and count (see Conrad et al. 1998; Gentner and Collins 1981). Thus the actual recall distribution would be a mixture. Nevertheless, we decided to use a normal prior, and used a mean of 1 in order to account for those with multiple instances. We used a standard deviation of 0 for all these variables.

For the SAT scores we used a prior mean of 575 for the math scores (because both campuses attract students who hope to major in technical subjects) and 475 for verbal scores (because both campuses attract a large number of students who are not native English speakers). In all cases we assumed a standard deviation of 50.

Table 1 shows the accuracy results. It compares the RGI estimate with the sample average as estimators of the true mean and a 95% credibility interval around the posterior mean with a 95% confidence interval around the sample average. The estimate that is closer to truth is shown in boldface; the interval estimates that cover truth are also in boldface.

Table 1. Comparing sample and RGI posterior means for estimating population means in Campus experiments using normal priors (boldface point estimates denote winners; boldface interval estimates denote intervals that cover truth)

	truth	$\bar{x}$	postmean	conf_int	cred_int
SUSB					
credits	67.53	63.13	**63.69**	**(56.12, 70.14)**	**(55.54, 71.84)**
gpa	2.91	2.99	**2.97**	**(2.89, 3.09)**	**(2.85, 3.09)**
satm	570.80	593.72	**591.97**	(572.40, 615.00)	**(553.15, 630.79)**
satv	503.20	526.00	**519.01**	(503.80, 548.20)	**(478.52, 559.50)**
tickets	0.53	**0.92**	0.95	(.56, 1.28)	**(.32, 1.58)**
fines	1.52	2.25	**1.00**	**(0, 5.41)**	**(.03, 1.96)**
UCR					
gpa	3.05	3.10	**3.04**	**(3.00, 3.20)**	**(2.88, 3.21)**
satm	574.1	572.60	**574.05**	**(549.50, 595.60)**	**(537.54, 610.56)**
satv	485.40	503.00	**500.38**	**(481.20, 524.80)**	**(463.74, 537.02)**
tickets	0.21	**0.51**	0.63	(.27,.75)	**(.09, 1.16)**

We see in Table 1 that in eight of the ten cases we were able to examine, the RGI posterior mean was closer to truth than was the sample average. Moreover, the 95% credibility intervals covered truth for all 10 out of the 10 items, whereas the 95% confidence intervals covered truth for only 6 of the 10 items. The two items for which the sample mean out-performed the point RGI estimator were for the same item at the two campuses: "TICKETS." Our explanation is that the distribution of the data for TICKETS was so nonnormal as to severely violate the fundamental assumptions of the RGI modeling (which assumes normal distributions throughout). A histogram of the TICKETS data for the SUSB data is the very nonnormal distribution displayed in Figure 1. The TICKETS data for UCR are analogously nonnormally distributed. In nonnormal data situations, transformations to normality may help, although in this situation, the data are largely zeroes, so the Box-Cox transformation described above does not help.

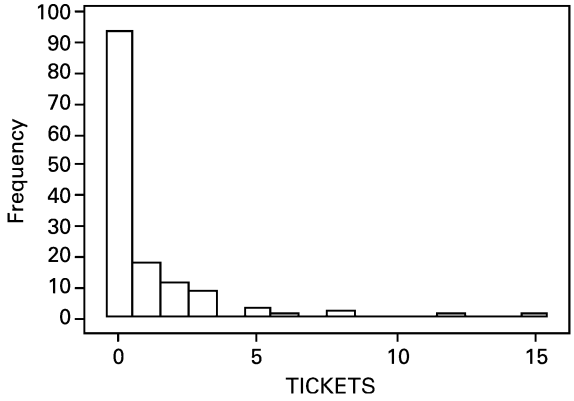

Fig. 1. Histogram of tickets data at SUSB

Table 2. Rank order correlations between length of interval and accuracy with outliers deleted

	UCR			SUSB		
VARIABLE	*n*	*r*	sig	*n*	*r*	sig
Credits	116	.25	.008	124	.21	.021
C’s	108	.69	< .001	125	.33	< .001
GPA	121	.37	< .001	137	.16	.064
SATM	102	.34	.001	76	.07	.524
SATV	90	.36	< .001	79	.26	.023
Tickets	130	.52	< .001	139	.39	< .001
RegFee	618	−.17	< .001			
RecFee	651	.12	.002			
SAFee				407	−.20	< .001
Health				407	−.13	.010
Food				63	.02	.901
Fines				124	.42	< .001

The campus experiments also permitted us to check our assumption that interval length is related to accuracy. For this analysis we were able to use all the data from the campus experiments, and we reported the results in Press and Tanur (2003). The key result of that analysis appears in Table 2. Note that in Table 2 the correlations are between interval length and absolute error (the absolute value of the difference between the respondent’s answer and verified “truth”), so positive correlations indicate that our expectations are fulfilled. Also note that in Table 2 outliers in the distribution of interval lengths have been deleted because they tend to be influential observations unduly inflating the correlations. Here we have defined an outlier as any observation more than 1.5 times the inter-quartile range above the third quartile or below the first quartile. We see that the correlations were not terribly high, nevertheless the RGI procedure was often able to improve on the performance of the sample average. If we can find ways to improve the questioning so as to improve the correlation, we should also be able to improve the accuracy of the RGI estimator.

Note that in Table 2 the highest correlations are for the variables that constitute frequencies – number of grades below C, traffic tickets, and library fines. Exploring these data shows us that those respondents who had zero or one occurrence were almost always quite accurate and gave short intervals, while those who had a large number of occurrences were less accurate and had longer intervals. This is what would be expected if respondents with few (and especially no) occurrences were using an actual recall strategy and those with many occurrences were actually estimating (see Conrad et al. 1998; Gentner and Collins 1981). Part of this potential bias can be taken into account, as detailed above, by a strategic choice of a prior mean, but the data in Table 1 suggest that in most cases of frequency variables both the sample mean and the RGI estimator overestimate truth.

The campus experiments also let us examine the effects of the RGI protocol on item nonresponse. These results are reported in Press and Tanur (2001). Those respondents who gave an interval but did not give a usage quantity constitute an appreciable percentage of those who did not give a usage quantity and thus were potential nonrespondents to each item. Indeed, those percentages are never less than 4% and twice are over 40%. We can interpret these results as estimated conditional probabilities of giving an interval among those who did not give a usage quantity. We can use the midpoint of the RGI as a point estimator and the interval from the average of the lower bounds to the average of the upper bounds (the Average Respondent-Generated Interval, ARGI) as an interval estimator, for those respondents who offered interval but no usage quantity responses. We can then inquire into the accuracy of these estimates for the fee data (where sample sizes are large and verification data unnecessary because of the uniformity of the fees across respondents). We find that the average midpoints overestimate usage for 3 of the 4 cases, but the ARGI covers the true value in all cases.

Thus in the Campus Experiments in a substantial proportion of cases, respondents who do not supply an estimate of usage quantities do supply intervals which are reasonably accurate, thus reducing the amount of item nonresponse appreciably.

There were both advantages and disadvantages to using university students as experimental subjects. The advantages were that such subjects were conveniently available to the experimenters, and record checks of the accuracy of their answers were readily available from campus administrators. The disadvantages were that there was internal evidence that a few students were not completely cooperative in terms of giving serious or truthful answer, and that only about half the students were willing to let us check their academic records. So we were eager to try out the RGI technique with more mature respondents, as well as to vary some other conditions.

4.2. The Census experiment

This experiment was designed to test for any differences in the order of asking the bounds and usage quantity questions, to test whether the technique can be used in a telephone interview, and to test the usefulness of the RGI proceedure for sensitive questions, such as a respondent's income (see Marquis and Press 1999; and Press and Marquis 2001). As detailed above, we are unable to report on the accuracy of the performance of the RGI estimator for this experiment because the verification data have been lost. But we can shed some light on the effects of question wording.

A frame of households was developed from the U.S. Census Bureau's commercial and administrative records containing households that filed joint tax returns having wage and salary income for the previous five consecutive years. The frame covered the four states in which the American Community Survey (ACS) held its first pilot tests. A sample of about 2,000 households was drawn from this frame, and each household was assigned to an experimental interviewing treatment. From this sample the U.S. Census Bureau's Hagerstown Telephone Facility obtained a quota of 500 completed CATI interviews, eliminating households that had become ineligible through retirement, death, divorce or other circumstances that precluded observing the joint wage and salary income on the tax return. Respondents answered questions about their income from salary and wages and from interest and dividends for each of the past two years, and for the change in both these types of income over the previous five years. Since the frame information also included data from administrative records about household income, we eventually linked the survey responses to the administrative records to evaluate the validity of the telephone survey responses, but those data have been lost.

This experiment used extensive cognitive pretesting for the form of the interval question. There had been some hope that sometimes the upper bound question could be asked before the lower bound question, but it was found that such an ordering made pretest respondents uncomfortable, so the experiment was designed to always ask for the lower bound before the upper bound. Also as a result of the cognitive testing the final instrument asked for the usage quantity as a "best estimate" in order to reinforce the notion that respondents might well be uncertain about their answers. One other outcome of the cognitive testing was to add a question about how confident the respondent was about his/her best estimate, as a way of introducing the intent of the bounds questions that immediately followed. This was done in a split-panel experiment in which 75 percent of the cases were asked the two bounds questions first, followed by the usage question. The other 25 percent of the cases were asked the usage question first, then the confidence rating, followed by the two bounds questions. This 25% sample enabled us to test our assumption that interval length is related to confidence. The results appear in Press and Tanur (2002) and are summarized in Table 3. These correlations do suggest that, except for the change variables which were very difficult for respondents to calculate, relative interval length (interval length/usage quantity) is a good proxy for confidence.

This cognitive testing showed that some telephone respondents had difficulty understanding and holding in memory a single question that asked for both lower and upper bounds. The solution was to split the question into two and ask, e.g. "What is the highest dollar amount you think this could have been?" and "What is the lowest dollar amount you think this could have been?" Interviewers reported considerable difficulty for some respondents in understanding this question, but the large majority of respondents were able to carry out the task successfully, supplying a lower bound that was lower than the usage quantity and an upper bound that was higher than the usage quantity.

In the Census Experiment, although many respondents did not supply usage quantities, in only a few such cases did they supply bounds information. Hence in this case RGI did little to reduce item nonresponse.

Table 3. Correlations between relative length of interval and confidence in the 25% Census sample

Item	*r*	*n*
Salary/Wages last year	.287**	102
Salary/Wages previous year	.247*	106
5-yr Change Salary/Wages	−0.014	95
Income/Dividends last year	.450**	81
Income/Dividends previous year	.319**	83
5-yr Change Income/Dividends	.155	71

**$p < .01$; *$p < .05$.

4.3. The HMO experiment

A fourth experiment has been fielded (see Miller and Press 2002) in order to test whether respondents are willing to answer the bounds question without being offered the usage question at all, and to explore which option they will choose if they are permitted to choose between the bounds question and the usage question. We are, of course, also interested in the accuracy of the responses in both these new situations.

Mail questionnaires were sent to 3,000 female members of an HMO (Health Maintenance Organization) asking questions about the length of their membership in the HMO; dates on which they had their most recent pap smear, mammogram, and influenza vaccination; date their most recent child was born in the HMO, and the birth-weight of that child; date of most recent blood test to measure cholesterol and the level of that cholesterol measurement. There were five groups of respondents: a control group that was asked the usage quantity only, another control group which was asked the questions in the form currently used by the HMO (respondents classify themselves into one of several interval options predetermined by the questionnaire designer), one group that received only the bounds questions, and two groups that were offered a choice of answering either the bounds question or the usage question (with the bounds question being offered first to one group and the usage question being offered first to the other group).

In the HMO experiment we modify RGI for one experimental group by deleting the requirement for respondents to give both a usage quantity and bounds information and asking them only for bounds information. (Two other groups can choose between the bounds and usage questions.) How will this affect the results? Respondent burden would certainly be reduced. A reasonable estimator of the population mean, might be a weighted average of the average of the lower bounds, $\bar{a}$, and the average of the upper bounds, $\bar{b}$. Such a weighted average, with weights ξ^*, could be expressed as

$$\tilde{\theta} = \xi^* \bar{a} + (1 - \xi^*)\bar{b} \tag{4.1}$$

Of course with a proper (normal) prior, we would need a further weighted average of the prior mean and the posterior mean with respect to a vague prior (given in Equation (4.1)). But how should the ξ^*-weights be selected? If they are chosen to be equal, the result is the "midpoint estimator" as used in our study of item nonresponse in the Campus experiments (see Press and Tanur 2001). (The same result is obtained by choosing the midpoints of all ranges given, and averaging these midpoints.) Another choice would be to select ξ^* to depend on the saliences of the questions to the respondents, and on the respondents' demographic characteristics. Yet another choice would involve the variances (and precisions) of the bounds information. Define the variances of the bounds:

$$\hat{\sigma}_a^2 = \frac{1}{n}\sum_1^n (a_i - \bar{a})^2, \quad \hat{\sigma}_b^2 = \frac{1}{n}\sum_1^n (b_i - \bar{b})^2$$

Reasonable ξ^*-weights could be taken to be:

$$\xi^* = \frac{\dfrac{1}{\hat{\sigma}_a^2}}{\dfrac{1}{\hat{\sigma}_a^2} + \dfrac{1}{\hat{\sigma}_b^2}}, \quad (1 - \xi^*) = \frac{\dfrac{1}{\hat{\sigma}_b^2}}{\dfrac{1}{\hat{\sigma}_a^2} + \dfrac{1}{\hat{\sigma}_b^2}}$$

Now $\tilde{\theta}$ weights $\bar{a}$ and $\bar{b}$ by precision proportions, as with the ordinary RGI estimator developed in the Appendix. Which estimation procedure would be best for such a *modified RGI protocol*? Analyses of the results of the HMO experiment may offer some information here.

4.4. *The Schwartz-Paulin Bureau of Labor Statistics experiment*

In face-to-face mock Consumer Expenditure Quarterly Survey interviews that compared RGI with unfolding brackets and conventional survey-designer-generated ranges, Schwartz and Paulin (2000) report some interesting findings. They used the following wording: "While we're talking about income, what I'd like you to do is tell the range within which you would feel almost certain that your actual income would fall. This is like completing the sentence, '*Oh yes, during the past 12 months, I must have earned between* ______ *and* ______. During the past 12 months did you receive any money in wages or salary? What do you think the range would be?" These authors found (p. 969): ". . .participants liked the RGI technique primarily because it afforded them some degree of control over their disclosures. Surprisingly, when respondents were given freedom to choose their own ranges, they did not opt for huge, relatively meaningless ranges that obscured their real financial situation. Instead respondent-generated intervals tended to be smaller than those generated by researchers. In this study, RGI was the only technique that resulted in respondents providing an exact value rather than a range."

Schwartz and Paulin (2000) found that the use of an interval technique reduced item nonresponse from 18.1% to 9.5%, though their sample size is too small to report these percentages separately for each of the three interval techniques they compared. They do note, however, that this improvement in item nonresponse came exclusively from those whose response to the usage quantity question were "don't know" rather than a refusal.

5. Discussion

We have seen that the RGI protocol (in the Campus experiments, at least, where we can test this assertion) quite often improves on the sample average as an estimator of the population mean. While we see that the correlations between interval length and accuracy (Table 2) seem to be reasonable, we are exploring the use of a form of asking the questions that results in a stronger relationship between accuracy and interval length. Such a higher correlation would ensure the success of the RGI protocol in reducing bias. An experiment is currently in the field that aims to direct those respondents most confident of their recall to give short intervals and those less confident to give longer ones. If confidence is related to accuracy then we should be able to improve the RGI procedure by this strategy. While we assume such a relationship exists, at least to some extent, evidence from the literature is mixed – for reviews see Bothwell, Deffenbacher, and Brigham (1987); Wells (1993); Wells and Murray (1984). We hope to be able to differentiate the kinds of questions for which this strategy will be effective from those for which it will not.

It may still be possible to improve the calibration of the wording of the bounds question in such a way that we can communicate its intended coverage more clearly to respondents. We have done some informal cognitive testing, asking respondents, for example, to choose an interval which they would be as sure of as they would be sure of drawing a white ball from an urn containing 99 white balls and 1 black ball. Respondents tended to stare at us in puzzlement. We should not be surprised at this outcome – not only does common sense suggest that such question wording would be puzzling to respondents, but much of the literature on probability estimation in surveys demonstrates the difficulty respondents experience in trying to estimate probabilities (and the resulting inaccuracy). See, for example, Tversky and Kahneman (1974) and Tversky and Koehler (1994). While this question asks for an application of probability, the difficulty should be similar.

Our next steps in investigating question wording will be to do more systematic empirical work to try to determine what respondents see as inclusion probabilities for the intervals they offer.

We have also seen differences across experiments in the effectiveness of RGI in reducing item nonresponse. Why? There may be an effect of the sensitivity of the questions interacting with mode of interview. There were sensitive questions about income in the Census experiment and in the Schwartz and Paulin (2000) experiment, less sensitive questions in the Campus experiments. In the paper-and-pencil Campus experiments it was easy to fill in part of a question, whether sensitive or not; it is less easy to answer part of a question, especially a sensitive one, posed by an interviewer over the telephone. In the Schwartz and Paulin (2000) experiment, respondents were interviewed face-to-face at a lab; such a setting might well encourage extra effort for questions in which the immediate recall is difficult. The type of respondent, type of interviewer, and survey sponsor may matter. Compared to the laboratory situation using paid respondents described by Schwartz and Paulin (2000), the Campus experiments involved undergraduate student respondents, students distributing questionnaires, and an "academic" survey. The Census experiment interviewed respondents from established households, who were presented with questions by professional interviewers representing the U.S. Census Bureau. Overall, there was greater respondent cooperation in this government survey by telephone than we found in our earlier campus-based experiments.

Thus, while several empirical questions remain open, we believe that the RGI protocol has demonstrated its usefulness in improving the accuracy of estimation of a population mean. In some cases it also is useful in reducing item nonresponse. The forthcoming results of the HMO experiment and of the currently fielded experiment encouraging shorter intervals from confident respondents and longer ones from less confident respondents should serve to answer some of the outstanding questions and to make the technique even more useful.

Appendix

In this Appendix we develop a hierarchical Bayesian model for estimating the posterior distribution of the population mean for data obtained in a sample survey by using the RGI protocol.

Suppose as his/her answer to a factual recall question respondent i gives a point response y_i, and bounds (a_i, b_i) for the true value of the answer, such that $a_i \le y_i \le b_i$, $i = 1, \ldots, n$. Assume:

$$\left(y_i | \theta_i, \sigma_i^2\right) \sim N\left(\theta_i, \sigma_i^2\right) \tag{A1}$$

The normal distribution will often be appropriate in situations for which the usage quantity, y_i, corresponds to a change in some quantity of interest. Assume the means, θ_i, of the usage quantities are themselves exchangeable, and normally distributed about some unknown population mean of fundamental interest, θ_0, so that:

$$(\theta_i | \theta_0, \tau^2) \sim N(\theta_0, \tau^2) \tag{A2}$$

Thus, respondent i has a recall distribution whose true value is θ_i (each respondent is attempting to recall, say, a different number of visits to the doctor last year). We would like to estimate θ_0. Assume $\left(\sigma_1^2, \ldots, \sigma_n^2, \tau^2\right)$ are known; they will be assigned later. Denote the column vector of usage quantities by $\utilde{y} = (y_i)$, and the column vector of means by $\utilde{\theta} = (\theta_i)$. Let $\utilde{\sigma}^2 = \left(\sigma_i^2\right)$ denote the column vector of data variances. The joint density of the y_i's is given in summary form by:

$$p(\utilde{y} | \utilde{\theta}, \utilde{\sigma}^2) \propto \exp\left\{\left(-\frac{1}{2}\right)\sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2\right\} \tag{A3}$$

The joint density of the θ_i's is given by:

$$p(\utilde{\theta} | \theta_0, \tau^2) \propto \exp\left\{\left(-\frac{1}{2}\right)\sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2\right\} \tag{A4}$$

So the joint density of $(\utilde{y}, \utilde{\theta})$ is given by:

$$p(\utilde{y}, \utilde{\theta} | \theta_0, \tau^2, \utilde{\sigma}) = p(\utilde{y} | \utilde{\theta}, \utilde{\sigma}^2) p(\utilde{\theta} | \theta_0, \tau^2)$$

or, multiplying (A3) and (A4) gives:

$$\begin{aligned} p(\utilde{y}, \utilde{\theta} | \theta_0, \tau^2, \utilde{\sigma}^2) &\propto \exp\left\{\left(-\frac{1}{2}\right)\left[\sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2\right]\right\} \\ &\propto \exp\left\{\left(-\frac{A(\utilde{\theta})}{2}\right)\right\} \end{aligned} \tag{A5}$$

where:

$$A(\utilde{\theta}) \equiv \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2 \tag{A6}$$

Expand (A6) in terms of the θ_i's by completing the square. After some algebra, we find:

$$A(\utilde{\theta}) = \sum_1^n \left\{\alpha_i \left[\left(\theta_i - \frac{\beta_i}{\alpha_i}\right)^2 + \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2}\right)\right]\right\} \tag{A7}$$

where:

$$\alpha_i = \frac{1}{\sigma_i^2} + \frac{1}{\tau^2}, \quad \beta_i = \frac{y_i}{\sigma_i^2} + \frac{\theta_0}{\tau^2}, \quad \gamma_i = \frac{\theta_0^2}{\tau^2} + \frac{y_i^2}{\sigma_i^2} \tag{A8}$$

Now find the marginal density of $\utilde{y}$ by integrating (A5) with respect to $\utilde{\theta}$. After rearranging terms, the required integral is given in (A9), below. We find:

$$p(\underset{\sim}{y}|\theta_0, \tau^2, \underset{\sim}{\sigma}^2) \propto J(\theta_0)\exp\left\{\left(-\frac{1}{2}\sum_1^n \alpha_i\delta_i\right)\right\} \tag{A9}$$

where:

$$J(\theta_0) \equiv \int \exp\left\{\left(-\frac{1}{2}\right)\sum_1^n \alpha_i\left(\theta_i - \frac{\beta_i}{\alpha_i}\right)^2\right\}d\underset{\sim}{\theta}, \quad \delta_i = \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2}\right)$$

Rewriting (A9) in vector and matrix form, to simplify the integration, we find that if

$$f \equiv \left(\frac{\beta_i}{\alpha_i}\right), \quad K^{-1} \equiv diag(\alpha_1, \ldots, \alpha_n)$$

$$(\underset{\sim}{\theta} - \underset{\sim}{f})'K^{-1}(\underset{\sim}{\theta} - \underset{\sim}{f}) = \sum_1^n \alpha_i\left(\theta_i - \frac{\beta_i}{\alpha_i}\right)^2 \tag{A10}$$

Carrying out the (normal) integration gives:

$$p(\underset{\sim}{y}|\theta_0, \tau^2, \underset{\sim}{\sigma}^2) \propto \frac{1}{|K^{-1}|^{1/2}}\exp\left\{\left(-\frac{1}{2}\sum_1^n \alpha_i\delta_i\right)\right\} \tag{A11}$$

Now note that $|K^{-1}| = \prod_1^n \alpha_i =$ constant and the constant can be absorbed into the proportionality constant, but δ_i depends on θ_0. So:

$$p(\underset{\sim}{y}|\theta_0, \tau^2, \underset{\sim}{\sigma}^2) \propto \exp\left\{\left(-\frac{1}{2}\sum_1^n \alpha_i\delta_i\right)\right\} \tag{A12}$$

Now applying Bayes' theorem to θ_0 in (A12) gives:

$$p(\theta_0|\underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto p(\theta_0)\exp\left\{\left(-\frac{1}{2}\sum_1^n \alpha_i\delta_i\right)\right\} \tag{A13}$$

where $p(\theta_0)$ denotes a prior density for θ_0. We consider two cases for the prior distribution for θ_0, a vague prior and a normal prior.

Vague prior for θ_0

Our prior belief (prior to observing the point estimates of the respondents) is that for the large sample sizes typically associated with sample surveys, the population mean, θ_0, might lie, with equal probability, anywhere in the interval (a_0, b_0), where a_0 denotes the smallest lower bound given by any respondent, and b_0 denotes the largest upper bound. So we could reasonably adopt a uniform prior distribution on (a_0, b_0). To be fully confident that we are covering all possibilities, however, we adopt the (improper) prior density on the entire real line. We therefore adopt a (vague) prior density of the form:

$$p(\theta_0) \propto \text{constant} \tag{A14}$$

for all θ_0 on the real line. Inserting (A14) into (A13), and noting that $p(\theta_0) \propto$ constant, gives:

$$p(\theta_0|\underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto \exp\left\{\left(-\frac{1}{2}\sum_{1}^{n} \alpha_i \delta_i\right)\right\} \tag{A15}$$

Next substitute for δ_i and complete the square in θ_0 to get:

$$p(\theta_0|\underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto \exp\left\{\left(-\frac{u}{2}\right)\left(\theta_0 - \frac{v}{u}\right)^2\right\} \tag{A16}$$

where:

$$u = \sum_{1}^{n}\left(\frac{1}{\tau^2} - \frac{1}{\alpha_i \tau^4}\right), \quad v = \sum_{1}^{n}\left(\frac{y_i}{\alpha_i \sigma_i^2 \tau^2}\right) \tag{A17}$$

Thus, the conditional posterior density of θ_0, under a vague prior for θ_0, is seen to be expressible as:

$$(\theta_0|\underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \sim N(\tilde{\theta}, \omega^2) \tag{A18}$$

where:

$$\tilde{\theta} \equiv \frac{v}{u}, \text{ and } \omega^2 \equiv \frac{1}{u} \tag{A19}$$

But these terms may be simplified, as shown below.

Conditional posterior mean of θ_0, under a vague prior, as a convex mixture of usages
The appropriate measure of location of the posterior distribution in Equation (A18) to use in any given situation depends upon the loss function that is appropriate. For many cases of interest the quadratic loss function (mean squared error) is appropriate. For such situations, we are interested in the posterior mean (under the normality assumptions in the current model, the conditional posterior distribution of θ_0 is also normal, so the posterior mean, median, and mode are all the same). It can be readily found by simple algebra that if:

$$\lambda_i \equiv \frac{\left(\dfrac{1}{\sigma_i^2 + \tau^2}\right)}{\sum_{1}^{n}\left(\dfrac{1}{\sigma_i^2 + \tau^2}\right)}, \quad \sum_{1}^{n}\lambda_i = 1 \tag{A20}$$

then:

$$\tilde{\theta} = \sum_{1}^{n}\lambda_i y_i \tag{A21}$$

Equation (A13) may now be reexpressed as:

$$p(\theta_0|\underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto p(\theta_0)\exp\left\{\left(-\frac{1}{2\omega^2}(\theta_0 - \tilde{\theta})^2\right)\right\} \tag{A22}$$

Thus, the mean of the conditional posterior density of the population mean, under a vague prior, is a convex combination of the respondents' point estimates, that is, their usage quantities. It is an unequally weighted average of the usage quantities, as compared with the sample estimator of the population mean, which is an equally weighted estimator, $\bar{y}$. If we interpret $\left(\sigma_i^2 + \tau^2\right)^{-1}$ as the precision attributable to respondent i's response, and $\sum_1^n \left(\sigma_i^2 + \tau^2\right)^{-1}$ as the total precision attributable to all respondents, λ_i is interpretable as the proportion of total precision attributable to respondent i. Thus, the larger his/her precision proportion, the larger the weight that is automatically assigned to respondent i's usage response.

Normal prior for θ_i

In some survey situations the same survey is carried out repeatedly so that there is strong prior information available for providing a realistic finite range for θ_0; in other situations we may have substantial information about the hyperparameters from other sources. In such cases we could improve on our estimator by using a proper prior distribution for θ_0 instead of the one given in Equation (A14). This is done explicitly below using a normal prior.

Suppose that for preassigned hyperparameters (θ^*, ρ^2), a priori,

$$\theta_0 \sim N(\theta^*, \rho^2) \tag{A23}$$

Substituting into Equation (A22) gives:

$$p(\theta_0 | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto \exp\left\{-\frac{1}{2\rho^2}(\theta_0 - \theta^*)^2\right\}\exp\left\{\left(-\frac{1}{2\omega^2}(\theta_0 - \tilde{\theta})^2\right)\right\} \tag{A24}$$

Expanding the quadratic terms in the exponents, combining terms, and completing the square in the form of a single quadratic term gives:

$$p(\theta_0 | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \propto \exp\left\{-\frac{1}{2\eta^2}(\theta_0 - g)^2\right\} \tag{A25}$$

where:

$$g \equiv \frac{\dfrac{\theta^*}{\rho^2} + \dfrac{\tilde{\theta}}{\omega^2}}{\dfrac{1}{\rho^2} + \dfrac{1}{\omega^2}}, \quad \eta^2 \equiv \frac{1}{\dfrac{1}{\rho^2} + \dfrac{1}{\omega^2}} \tag{A26}$$

Equivalently,

$$(\theta_0 | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \sim N(g, \eta^2) \tag{A27}$$

Just as under a vague prior the posterior mean could be written as a convex mixture (Equation (A21)), similarly, under the normal prior the posterior mean, g, in Equation (A26) may also be rewritten as the convex mixture:

$$g \equiv \xi\tilde{\theta} + (1 - \xi)\theta^* \tag{A28}$$

where:

$$\xi \equiv \frac{\frac{1}{\omega^2}}{\frac{1}{\omega^2}+\frac{1}{\rho^2}}, \quad (1-\xi) \equiv \frac{\frac{1}{\rho^2}}{\frac{1}{\omega^2}+\frac{1}{\rho^2}} \tag{A29}$$

$0 \leq \xi \leq 1$, with $\tilde{\theta}$ as given in (A20) and (A21). The posterior precision, η^{-2}, is the sum of the precision of the data information under a vague prior, ω^{-2}, and the precision of the prior information, ρ^{-2}.

Assessing the variance parameters

To assess the variance parameters when they may all be dissimilar, define $k_1 = k_2 = k$ and obtain σ_i and τ from:

(a) $k_1\sigma_i = (b_i - a_i)$, for all $i = 1, \ldots, n$; for some k_1, such as $k_1 = 2, 3, 4, 5, 6, 7, 8$. Typically, we would take $k = 4$ (2 standard deviations on either side of the mean). Define, as above:
(b) $\bar{a} = \frac{1}{n}\sum_1^n a_i$, and $\bar{b} = \frac{1}{n}\sum_1^n b_i$. Then,
(c) $k_2\tau = b_0 - a_0$ for some preassigned k_2, $a_0 \equiv \min_{1\leq i\leq n}(a_i)$, and $b_0 \equiv \max_{1\leq i\leq n}(b_i)$. τ is the same for all respondents. We use an interval of 2 standard deviations on either side of the (normal) mean of the individual recall distribution means for the respondents. We need an assessment that will be reasonable for all respondents. We use the sample range over all respondent's intervals.

Different analysts might interpret the k's somewhat differently. Using these variance assessments, the weights become approximately:

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i-a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}}\right)}{\sum_1^n \left(\frac{1}{\frac{(b_i-a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}}\right)}, \quad \sum_1^n \lambda_i = 1 \tag{A30}$$

where: $r_0 \equiv b_0 - a_0$. Note that in the special case that $k_1 = k_2$, the k's cancel out in numerator and denominator of (A30), so that the weights do not depend upon the k's. Then, the weights become:

$$\lambda_i \doteq \frac{\left(\frac{1}{(b_i-a_i)^2+r_0^2}\right)}{\sum_1^n \left(\frac{1}{(b_i-a_i)^2+r_0^2}\right)} \tag{A31}$$

Conditional posterior variance of θ_0

It is straightforward to check that the conditional posterior variance of θ_0 under a vague prior is given by:

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2+\tau^2}\right)} \doteq \frac{1}{\sum_1^n \left(\frac{1}{\frac{(b_i-a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}}\right)} \tag{A32}$$

the reciprocal of the total precision for all respondents in the sample. For $k_1 = k_2 = k$,

$$\omega^2 \doteq \frac{1}{\sum_1^n \left(\frac{k^2}{(b_i - a_i)^2 + r_0^2}\right)} \tag{A33}$$

so that in this case, while the conditional posterior mean does not depend upon k, the conditional posterior variance does. So the conditional posterior distribution of the population mean, under a vague prior, is given approximately by:

$$(\theta_0 | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \sim N(\tilde{\theta}, \omega^2) \tag{A34}$$

where $\tilde{\theta}$ and ω^2 are given in (A21), (A30) and (A31), and (A32) or (A33). Results are analogous for the normal prior.

Credibility intervals

Let $z_{\varepsilon/2}$ denote the $\varepsilon/2$-percentile of the standard normal distribution. Then, for a vague prior on the population mean, a (100-ε)% credibility interval for the population mean, θ_0 is given by:

$$(\tilde{\theta} - z_{\varepsilon/2}\omega, \tilde{\theta} + z_{\varepsilon/2}\omega) \tag{A35}$$

That is,

$$P\{\tilde{\theta} - z_{\varepsilon/2}\omega \leq \theta_0 \leq \tilde{\theta} + z_{\varepsilon/2}\omega | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2\} = (100 - \varepsilon)\% \tag{A36}$$

For a normal prior $N(\theta^*, \rho^2)$ on the population mean, a $(100 - \varepsilon)\%$ credibility interval for the population mean, θ_0 is given analogously by:

$$P\{g - z_{\varepsilon/2}\eta \leq \theta_0 \leq g + z_{\varepsilon/2}\eta | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2\} = (100 - \varepsilon)\% \tag{A37}$$

6. References

Berry, D.A. (1996). Statistics: A Bayesian Perspective. Belmont, CA: Wadsworth Pub. Co.

Bothwell, R.K., Deffenbacher, K.A., and Brigham, J.C. (1987). Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited. Journal of Applied Psychology, 72, 691–695.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society, Series B, 26, 211–252.

Conrad, F.G., Brown, N.R., and Cashman, E.R. (1998). Strategies for Estimating Behavioral Frequency in Survey Interviews. Memory, 6, 339–366.

Gentner, D. and Collins, A. (1981). Studies of Inference from Lack of Knowledge. Memory and Cognition, 9, 434–443.

Hogarth, R. (1980). Judgment and Choice. New York: John Wiley and Sons, Inc.

Marquis, K.H. and Press, S.J. (1999). Cognitive Design and Bayesian Modeling of a Census Survey of Income Recall. Proceedings of the Federal Committee on Statistical Methodology Conference, Washington, DC, Nov. 16, 51–64 (see http://bts.gov/fcsm).

Miller, D. (2004). Can We Ask Better Questions: An Examination of the Respondent-Generated Intervals Protocol. Ph.D. Thesis, Department of Statistics, University of California, Riverside.

Miller, D. and Press, S.J. (2002). An Experiment Embedded in a Health Survey with Respondent-Generated Intervals. Paper presented at the Annual Meetings of the American Statistical Association, August.

Press, S.J. (1996). Bayesian Recall: A Cognitive Bayesian Modeling Approach to Surveying A Recalled Quantity. Technical Report No. 236, Department of Statistics, University of California, Riverside, August.

Press, S.J. (1999). Respondent-Generated Intervals for Recall in Sample Surveys. Manuscript, Department of Statistics, University of California, Riverside, CA 92521-0138, Jan. http://cnas.ucr.edu/ ~ stat/press.htm

Press, S.J. (2002). Respondent-Generated Intervals for Recall in Sample Surveys. Technical Report No. 272, July. Department of Statistics, University of California, Riverside.

Press, S.J. (2004). Respondent-Generated Intervals for Recall in Sample Surveys. Journal of Modern Applied Statistical Methods, Vol. 3, No. 1.

Press, S.J. and Marquis, K.H. (2001). Bayesian Estimation in a U.S. Census Bureau Survey of Income Recall Using Respondent-Generated Intervals. Journal of Research in Official Statistics, Amsterdam: Eurostat.

Press, S.J. and Marquis, K.H. (2002). Bayesian Estimation in a U.S. Government Survey of Income Using Respondent-Generated Intervals. Proceedings of the Sixth World Meeting of the International Society for Bayesian Analysis, May, Crete, Greece. Amsterdam: Eurostat.

Press, S.J. and Tanur, J.M. (2000). Experimenting with Respondent-Generated Intervals in Sample Surveys, with Discussion. Pages 1–18 in Monroe G. Sirken (ed.) Survey Research at the Intersection of Statistics and Cognitive Psychology. Working Paper Series #28, National Center for Health Statistics, U.S. Department of Health and Human Services, Center for Disease Control and Prevention.

Press, S.J. and Tanur, J.M. (2001). Can Respondent-Generated Interval Estimation in Sample Surveys Reduce Item-Nonresponse? In V.M. Ahsanullah, J. Kenyon, and S.K. Sarkar (eds) Applied Statistical Science. Huntington, NY: Nova Science Publishers, Inc., 39–49.

Press, S.J. and Tanur, J.M. (2002). Cognitive and Econometric Aspects of Responses to Surveys as Decision-Making. Technical Report #271, Department of Statistics, University of California at Riverside, Riverside, CA 92521-0138.

Press, S.J. and Tanur, J.M. (2003). The Relationship between Accuracy and Interval Length in the Respondent-Generated Interval Protocol. Paper presented at the Annual Conference of the American Association for Public Opinion Research, May. Washington, D.C.: American Statistical Association.

Ramsey, F.P. (1931). Truth and Probability. In R.B. Braithwaite (ed.), The Foundations of Mathematics and Other Logical Essays, London and New York, p.71. Reprinted in Henry E. Kyburg, Jr. and Howard E. Smokler (eds) (1980), Studies in Subjective Probability. Huntington, New York: Robert E. Krieger Pub. Co., 25–52.

Schlaifer, R. (1959). Probability and Statistics for Business Decisions. New York: McGraw Hill Book Co, Inc.

Schwartz, L.K. and Paulin, G.D. (2000). Improving Response Rates to Income Questions. Proceedings of the American Statistical Association, Section on Survey Research Methods, 965–969.

Schwarz, N. and Sudman, S. (eds) (1994). Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer-Verlag.

Sirken, M.G., Herrmann, D.J., Schechter, S., Schwarz, N., Tanur, J.M., and Tourangeau, R. (eds) (1999). Cognition and Survey Research. New York: John Wiley and Sons.

Sudman, S., Bradburn, S., and Schwarz, N. (1996). Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge: Cambridge University Press.

Tversky, A. and Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. Science, 185, 1124–1131.

Tversky, A. and Koehler, D.J. (1994). Support Theory: A Nonextensional Representation of Subjective Probability. Psychological Review, 101, 547–567.

Wells, G.L. (1993). What Do We Know about Eye Witness Identification? American Psychologist, 35, 553–571.

Wells, G.L. and Murray, D.M. (1984). Eyewitness Confidence. In G.L. Wells and E.F. Loftus (eds). Eyewitness Testimony: Psychological Perspectives. New York: Cambridge University Press, 155–170.

Received February 2003
Revised January 2004