An Overview of the Respondent-Generated Intervals (RGI): Approach to Sample Surveys

S. James Press and Judith M. Tanur

Abstract

This article brings together many years of research on the Respondent-Generated Intervals (RGI) approach to recall in factual sample surveys. Additionally presented is new research on the use of RGI in opinion surveys and the use of RGI with gamma-distributed data. The research combines Bayesian hierarchical modeling with various cognitive aspects of sample surveys.

Keywords: anchoring, Bayesian, confidence scale, recall, surveys

1. Introduction

This work provides an overview of the research to date on Respondent-Generated Intervals, or RGI, protocol for asking questions in sample surveys. It brings together a body of research that started in 1996 with some theoretical ideas about how survey questionnaire design might be improved by asking respondents (Rs) for more than just a basic answer to a question, but also eliciting information about how certain the Rs might be about their answers. Over the years we developed various theoretical models for analyzing such RGI data from a survey, culminating in the current Bayesian hierarchical model detailed in Section II. With the development of a theoretical model came the need to explore how well the model might work in practice. We examined pencil-and-paper classroom surveys, and a telephone survey using Census data. Various surveys we carried out under the different survey protocols are described in Section III. Our conclusions so far can be found in Section IV.

The RGI protocol was originally developed to deal with questions requiring recall of numerical facts; it has been extended to address questions of opinion as well. We discuss this extension below. In its original form, the RGI protocol involves asking each R not only for a basic answer to a recall-type question (an answer we call a "usage quantity") but also, for a smallest value his/her true answer could be, and a largest value. We call these values the lower and upper bounds. The result of the RGI protocol is that the Rs themselves generate intervals in which their true beliefs lie, instead of having their quantitative beliefs forced into intervals pre-assigned by a survey designer, as is often done in other survey protocols. (For a discussion of survey protocols using intervals or brackets, see Press, 2004).

1.2 Genesis of RGI

The RGI protocol has its origins in Bayesian assessment procedures. In that context, for a specific individual, we might assess an entire prior distribution about an unknown parameter. That prior distribution represents the individual's degrees of uncertainty about that unknown parameter. We might assess many points on the individual's subjective probability distribution for that parameter by means of a sequence of elicitation questions, and then connect those points by a smooth curve that represents the underlying distribution. In the RGI protocol, because of concern for respondent burden, we ask for only three points on the recall distribution.

For example, suppose an individual has a normal subjective probability distribution representing "θ_0", the true (but unknown) change in the number of doctor visits he/she believes he/she made last year, compared with the previous year, so that $\theta_0 \sim N(4,1)$. (We use "change" in doctor visits as our illustrative variable in order to provide for both positive and negative values; thus we make the assumption of normality more plausible.) In such a case, the individual believes that it is most likely that he/she visited a doctor 4 more times last year than the previous year, with a standard deviation of 1. So this individual equivalently believes that there is a 99.7% chance that he/she visited a doctor between 1 and 7 more times last year, or that there is really almost no chance that the true number of additional times was less than 1 or greater than 7. This probability distribution is subjective, in that it represents a specific individual's degrees of belief about his/her uncertainty about the underlying quantity.

We postulate that: in a factual survey each R has a distinctive recall distribution, and in an attitude or opinion survey he/she has an underlying probability distribution for his/her opinion or attitude about some issue. In the case of a recall-type question, we assume that the R knew the true value at some time in the past (or knew enough to construct an accurate answer) but because of imperfect recall, he/she is now not certain of the true value. We also assume that R is not purposely trying to deceive. In the case of opinions or attitudes, R may have a very fuzzy attitude about an issue, or he/she may feel quite strongly and specifically about it.

2. Theoretical Development

2.1 Normal Data for Recall Questions

Suppose Rs answer independently and suppose R_i gives a point response, y_i, and bounds (a_i, b_i), $a_i \le b_i$, $i = 1,\dots, n$, as his/her answers to a factual recall question. We'll refer to y_i as R_i's "usage quantity" (the term "usage quantity" was introduced originally to reflect estimated frequency of a behavior). The random quantities (y_i, a_i, b_i) are jointly distributed. Assume:

$$(y_i | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2). \qquad \text{(A1)}$$

The normal distribution will often be appropriate in situations for which the usage quantity corresponds to a change in some quantity of interest. In other situations the gamma or another sampling distribution might be more appropriate. In such a case, we assume the $y_i\text{'s}$ (and the (a_i, b_i)) have been pre-transformed, so that after the transformation, the resulting variables are approximately normally distributed. Assume the means of the usage quantities are themselves exchangeable, and normally distributed about some unknown population mean of fundamental interest, θ_0:

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

Thus, R$_i$ has a recall distribution whose true mean value is θ_i (e.g., each R is attempting to recall his/her particular number of visits to the doctor last year). It is desired to estimate θ_0. Assume $(\sigma_1^2, \ldots, \sigma_n^2, \tau^2)$ are known; they will be assigned later. Denote the column vector of usage quantities by $\underset{\sim}{y} = (y_i)$, and the column vector of means by $\underset{\sim}{\theta} = (\theta_i)$. Let $\underset{\sim}{\sigma}^2 = (\sigma_i^2)$ denote the column vector of data variances. The joint density of the $y_i\text{'s}$ is given in summary form by:

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

(A3)

The joint density of the $\theta_i\text{'s}$ is given by:

$$p(\underset{\sim}{\theta} | \theta_0, \tau^2) \propto \exp\left\{(-\frac{1}{2})\sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2\right\}. \qquad \text{(A4)}$$

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

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

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

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

$$\propto \exp(-\frac{1}{2})\left[\sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2\right]$$

$$\exp(-\frac{1}{2})\left[\sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2\right] \qquad \text{(A5)}$$

$$\propto \exp\left\{(-\frac{A(\underset{\sim}{\theta})}{2})\right\},$$

where: $A(\underset{\sim}{\theta}) \equiv \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau}\right)^2. \qquad \text{(A6)}$

Expand Eq. (A6) in terms of the $\theta_i\text{'s}$ by completing the square. This takes some algebra. Then find the marginal density of $\underset{\sim}{y}$ by integrating Eq. (A5) with respect to $\underset{\sim}{\theta}$.

Next apply Bayes' theorem to θ_0, where $p(\theta_0)$ denotes a prior density for θ_0. Prior belief (prior to observing the point and bound estimates of the Rs) 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 R, and b_0 denotes the largest upper bound. So adopt a uniform prior distribution on (a_0, b_0). To be fully confident of covering all possibilities, however, adopt an (improper) prior density. Therefore adopt a prior density of the form:

$$p(\theta_0) \propto \text{constant}, \qquad \text{(A7)}$$

for all θ_0 on the entire real line. (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 such cases we could improve on our estimator by using a proper prior distribution for θ_0 instead of the one given in Eq. (A7).) The development for a normal (rather than a vague) prior distribution on the population mean is simple and analogous.

Next, complete the square in θ_0 to get, after some algebra, the final result that if:

$$\lambda_i \equiv \frac{\left(\frac{1}{\sigma_i^2 + \tau^2}\right)}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2}\right)}, \quad \sum_1^n \lambda_i = 1,\ 0 \le \lambda_i \le 1, \qquad \text{(A8)}$$

the conditional posterior density of θ_0 is seen to be expressible as:

$$(\theta_0 | \underset{\sim}{y}, \tau^2, \underset{\sim}{\sigma}^2) \sim N(\tilde{\theta}, \omega^2), \qquad \text{(A9)}$$

where:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i, \qquad \text{(A10)}$$

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2}\right)}. \qquad \text{(A11)}$$

Thus, the mean, $\tilde{\theta}$, of the conditional posterior density of the population mean, θ_0, is a convex combination of the Rs' 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, $\bar{y}$. Interpret $(\sigma_i^2 + \tau^2)^{-1}$ as the precision attributable to R_i's response, and $\sum_1^n (\sigma_i^2 + \tau^2)^{-1}$ as the total precision attributable to all Rs; then, λ_i is interpretable as the proportion of total precision attributable to R_i. Thus, the greater his/her precision proportion, the greater the weight that is automatically assigned to R_i.'s usage response. We must still assess the variances $(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2, \tau^2)$.

2.2 Assessing the Variances

Suppose that in addition to Eq. (A1):

$$a_i \left| a_{i0}, \psi_{ai}^2 \sim N(a_{i0}, \psi_{ai}^2); \right. \tag{A12}$$

$$b_i \left| b_{i0}, \psi_{bi}^2 \sim N(b_{i0}, \psi_{bi}^2), \right. \tag{A13}$$

where θ_i in Eq. (A1) denotes the true population value for the mean usage for R_i.; a_{i0}, b_{i0} denote the true population values for R_i.'s lower and upper bounds, respectively; and $(\sigma_i^2, \psi_{ai}^2, \psi_{bi}^2)$ denote the corresponding population variances, respectively. Next, using the structure of the normal distribution, assume the approximate bounds for all subjects in the population are approximately 2 standard deviations on either side of the respective means. Accordingly, take approximately,

$4\sigma_i \approx b_i - a_i,\ i = 1, ..., n,$ as our assessments for σ_i's.

Then, define:

$$a^* = \frac{1}{N}\sum_1^N a_{i0}; \qquad b^* = \frac{1}{N}\sum_1^N b_{i0};$$

$$\bar{a} = \frac{1}{n}\sum_1^n a_i; \qquad \bar{b} = \frac{1}{n}\sum_1^n b_i,$$

where: a^*, b^* are averages of the *true* (unobserved) values of these bounds over the entire population; $\bar{a}, \bar{b}$ are the averages of the *observed* values of the bounds over the sample.

Assume approximately:

$$\psi_{a1}^2 = \psi_{a2}^2 = ... = \psi_a^2; \quad \psi_{b1}^2 = \psi_{b2}^2 = ... = \psi_b^2.$$

Then,

$$\bar{a} \sim N(a^*, \frac{\psi_a^2}{n}); \qquad \bar{b} \sim N(b^*, \frac{\psi_b^2}{n}). \tag{A14}$$

Next note that the true population mean value for R_i. must be between its bounds,

$$a^* \le \theta_0 \le b^*. \tag{A15}$$

2.2.1 Case 1—Extended Average Estimator

For 95% credibility on a^* with respect to a vague prior we have (approximating 1.96 by 2, here and throughout, for convenience):

$$\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \le a^* \le \bar{a} + 2\frac{\psi_a}{\sqrt{n}}; \tag{A16}$$

and for 95% credibility on b^* with respect to a vague prior:

$$\bar{b} - 2\frac{\psi_b}{\sqrt{n}} \le b^* \le \bar{b} + 2\frac{\psi_b}{\sqrt{n}}. \tag{A17}$$

Then, we get:

$$\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \le a^* \le \theta_0 \le b^* \le \bar{b} + 2\frac{\psi_b}{\sqrt{n}},$$

or:

$$\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \le \theta_0 \le \bar{b} + 2\frac{\psi_b}{\sqrt{n}}. \tag{A18}$$

From the normality and 95% credibility,

$$4\tau = \left(\bar{b} + 2\frac{\psi_b}{\sqrt{n}}\right) - \left(\bar{a} - 2\frac{\psi_a}{\sqrt{n}}\right) = \left(\bar{b} - \bar{a}\right) + \frac{2}{\sqrt{n}}\left(\psi_a + \psi_b\right). \tag{A19}$$

But ψ_a and ψ_b are unknown. Estimate them by their sample quantities:

$$s_a^2 \equiv \hat{\psi}_a^2 \equiv \frac{1}{n}\sum_1^n (a_i - \bar{a})^2; \qquad s_b^2 \equiv \hat{\psi}_b^2 \equiv \frac{1}{n}\sum_1^n (b_i - \bar{b})^2. \tag{A20}$$

Then, the assessment procedure for τ becomes:

$$4\tau \doteq \left(\bar{b} - \bar{a}\right) + \frac{2}{\sqrt{n}}\left(s_a + s_b\right). \tag{A21}$$

There is a Minitab 13 macro for computing the Bayesian RGI extended average estimator.

2.2.2 Case 2—Extended Range Estimator

From Eq. (A24), since $a_0 < \bar{a}$, and $\bar{b} < b_0$, we can consider for an alternative assessment procedure,

$$a_0 - 2\frac{\psi_a}{\sqrt{n}} \le \theta_0 \le b_0 + 2\frac{\psi_b}{\sqrt{n}}. \tag{A22}$$

Then, we find:

$$4\tau = \left(b_0 + 2\frac{\psi_b}{\sqrt{n}}\right) - \left(a_0 - 2\frac{\psi_a}{\sqrt{n}}\right) = \left(b_0 - a_0\right) + \frac{2}{\sqrt{n}}\left(\psi_a + \psi_b\right). \tag{A23}$$

Using Eq. (A2) gives:

$$4\tau \doteq \left(b_0 - a_0\right) + \frac{2}{\sqrt{n}}\left(s_a + s_b\right). \tag{A24}$$

Note that the second term disappears for large sample sizes, leaving us with just the average or range of the bounds, but for smaller sample sizes, the second term can have a substantial effect.

2.3 Non-Normal Data for Recall Questions

Suppose the usage quantity data, the y_i 's , follow a 2-parameter gamma distribution instead of the normal distribution assumed in Section IIA. Adopt the probability density structure:

$$f(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-\frac{y}{\beta}} \quad y > 0,$$
$$\alpha > 0,\ \beta > 0, \qquad \text{(B1)}$$

With:

$$E(Y) = \alpha\beta \equiv \mu,\ mode(Y) = (\alpha-1)\beta,$$
$$var(Y) = \alpha\beta^2. \qquad \text{(B2)}$$

Define a new transformation parameter μ by:

$$\beta = \frac{\mu}{\alpha}$$

We can rewrite the gamma distribution in terms of μ as:

$$f(y|\alpha,\mu) = \frac{1}{\Gamma(\alpha)\left(\frac{\mu}{\alpha}\right)^{\alpha}} y^{\alpha-1} e^{-\frac{\alpha y}{\mu}}, \qquad \text{(B3)}$$

with: mean $E(Y) = \mu$,

$$mode(Y) = \mu - \frac{\mu}{\alpha},$$

$$var(Y) = \frac{\mu^2}{\alpha}.$$

Now make the normalizing transformation (see McCullagh and Nelder, 1983, Chapter 7.2):

$$Z = 3\left[\left(\frac{y}{\mu}\right)^{1/3} - 1\right], \qquad \text{(B4)}$$

so that now, the transformed variable is approximately a standard normal variable; i.e., $Z \sim N(0,1)$. Under this transformation the precision parameter $\alpha = \sigma^{-2}$ is assumed constant for all observations. Applying this transformation to all the variables creates a new set of standard normal variables. Modifying their locations and scales, as shown below, reduces the problem, approximately, to the one discussed in IIA.

Applying the transformation in (B4) to all the usage quantities gives:

$$Z_i = 3\left[\left(\frac{y_i}{\theta_i}\right)^{1/3} - 1\right]. \qquad \text{(B5)}$$

Now, the Z_i 's are independent, and approximately,

$$Z_i \sim N(0,1). \qquad \text{(B6)}$$

Next define the new variables, Z_i^* by:

$$Z_i^* \equiv \theta_i + \sigma_i Z_i. \qquad \text{(B7)}$$

Now we have the Z_i^* 's mutually conditionally independent, and

$$(Z_i^* | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2),\ i = 1,\ldots,n. \qquad \text{(B8)}$$

Suppose the θ_i 's are exchangeable, with

$$\theta_i \sim N(\theta_0, \tau^2). \qquad \text{(B9)}$$

Assume

$$p(\theta_i) \propto \text{ constant}. \qquad \text{(B10)}$$

We would like to find a Bayesian estimator of the population mean, θ_0.

We already know that for given $(\sigma_1^2,\ldots,\sigma_n^2,\tau^2)$, by Bayes' theorem,

$$(\theta_0 | Z_1^*,\ldots,Z_n^*,\sigma_1^2,\ldots,\sigma_n^2,\tau^2) \sim N(\tilde{\theta},\omega^2), \qquad \text{(B11)}$$

where the posterior mean of θ_0 is given by:

$$\tilde{\theta} = \sum_1^n \lambda_i Z_i^*, \qquad \text{(B12)}$$

$$\lambda_i \equiv \frac{\left(\frac{1}{\sigma_i^2+\tau^2}\right)}{\sum_1^n \left(\frac{1}{\sigma_i^2+\tau^2}\right)}, \qquad \sum_1^n \lambda_i = 1,\ 0 \le \lambda_i \le 1,$$

and

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2+\tau^2}\right)}. \qquad \text{(B13)}$$

Now substitute approximations for the unknown parameters.

3. RGI and Opinion Questions

Suppose there is a population of opinions about some issue. Perhaps the analyst would like to establish the mean of the opinions of all people living in New York City about the issue. There is no "correct" answer for an opinion or for an attitude for a given R, as there would be for a recall-type of question. Similarly, response bias does not have the same meaning as in recall. (For a recall-type of question, one reason for response bias is faulty memory.)

When using RGI for attitudes or opinions we can find both point and interval estimators. The RGI point estimator provides some information about the intensity of opinions of New Yorkers about the issue more than would a mere traditional sample mean that includes some people with very fuzzy opinions, and some people who have very firm opinions. RGI can provide various measures of strength-of-opinion. One such is the average range of the bounds supplied by all Rs , $(\bar{b} - \bar{a})$. It can also supply a credibility interval measure of belief. Of course, a confidence interval can also supply an interval measure of belief, but the confidence interval only reflects sampling uncertainty, whereas the RGI credibility interval also reflects individual fuzziness of opinion. The range also available with RGI, r_i o $(b_0 - a_0)$, is somewhat different; it measures the distance between extremes of opinion.

Another measure of strength-of-opinion is one we call "fuzziness." There is certainly no unique way to define such a quantity. One way might be to measure it using the following scale. Recall that the ith R's bounds are given by (a_i, b_i), and the usage quantity for R$_i$. is given by y_i. Now define the fuzziness of R$_i$.'s opinion as:

$$f_i = r_i \left[1 - \exp\left\{ -\frac{r_i}{|y_i|} \right\} \right]. \qquad \text{(C1)}$$

As y_i varies, this measure varies between 0 and $(b_i - a_i)$. It is a monotone increasing function of the range, ($b_i - a_i$). So the greater the range, the greater the degree of fuzziness, and conversely. Moreover, when $r_i = 0$, $f_i = 0$. This definition is driven by the need to avoid mathematical difficulties using $(b_i - a_i)/y_i$ when y_i is near the origin.

4. Empirical Studies of RGI

During the time that we have worked on RGI, our thinking has evolved. We have improved our modeling, the way we assess parameters (the population variance and the prior mean), and the form of our questioning. These changes are reflected in the design, analyses, and findings of our empirical work.

In our first empirical effort we ran parallel record-check surveys on our campuses, asking students questions about their life on campus. If the student-Rs gave their consent, their answers were verified through the appropriate campus offices. On both campuses we asked about the number of credits earned (CREDITS), SAT math and verbal scores (SATM, SATV), GPA, the number of grades of C or below (Cs), and the number of parking tickets (TICKETS). At the University of California at Riverside (UCR) we also asked about the registration fee (REGFEE) and the recreation center fee (RECFEE) the student had paid. At the State University of New York at Stony Brook (SUNY-SB) we also asked about the student activities fee (SAFEE) and the health fee (HEALTH) the student had paid, as well as the amount spent on food via the food plan (FOOD) and the number of library fines (FINES) assessed. In the campus surveys we asked the bounds question thus:

Please fill in the blanks – "There is almost no chance that the number of credits that I had earned by beginning of this quarter was less than ______ and almost no chance that it was more than ______."

In an attempt to estimate the population mean, our initial estimation procedure for these experiments compared:

(1) the usual sample mean;
(2) the average of the midpoints of the intervals given by the Rs, designated the midpoint estimator; as well as
(3) a Bayesian point estimator.

That Bayesian estimator was the mean of the posterior distribution of the true population mean value obtained from a two-stage hierarchical model using an assumed normal likelihood, exchangeable normal priors for the means of each R's data distribution, and an exponential distribution for the common precision parameter of the Rs' exchangeable normal priors. In addition, we adopted a normal prior for the population mean, centered at the midpoint estimator, $(\bar{a} + \bar{b})/2$.

The posterior distribution for the population mean was complicated (the ratio of multiple integrals), but was evaluated numerically by Gibbs sampling Markov Chain Monte Carlo (MCMC). (See Press, 1997 for a derivation of the estimator, and Press and Tanur, 2000, for further details about the campus experiments.) The results given by these estimators were compared in terms of their closeness to the true means found in record checks.

For the 18 items tested in the campus experiments, this initial analysis found that the posterior mean was always very close to the midpoint estimator. This similarity was not surprising as we chose deliberately to use a sharp (non-vague) prior. The Bayesian estimator looked relatively good; but it was difficult to compute. Of the three estimates, the Bayesian estimate was least accurate for just one item, the midpoint estimate least accurate 7 times, and the sample mean least accurate 10 times.

Our next empirical study was carried out during a fellowship at the Bureau of the Census held by S. J. Press. Census Bureau interviewers carried out telephone interviews with Rs from 500 households, asking about the household's economic situation. Rs were asked about their income from salary and wages for the most recent year and the previous year and about the change in their income from these sources over the previous 5 years. They were also asked similar questions about their income from interest and dividends. This study involved extensive cognitive testing of the question form (see Marquis and Press, 1999), and finally settled on asking 25% of the Rs the usage quantity first, followed by questions about the bounds (e.g.):

a) What is your best estimate of your household's income from salary and wages in 1997?
b) What is the lowest the correct value could be?
c) What is the highest the correct value could be?

Thus the bounds question was broken into two separate questions. For the remaining 75% of the Rs, the bounds questions was asked before the usage quantity question, to see if the order of the questions would make any difference. Bayesian estimation was again carried out using normal priors, and MCMC. In this work, however, two versions of the estimation were carried out. One used the sample median of the usage quantities as the mean of the prior distribution and the other used the midpoint estimator as in the campus experiments.

There were 12 comparisons possible between the sample mean and the two Bayesian estimators. Of these, the sample mean was closest to the truth 4 times, the Bayesian estimator using the median closest 4 times, the Bayesian estimator using the midpoint estimator closest 3 times, and there was one tie between the Bayesian estimators. In a "head to head contest" between the two Bayesian estimators, the one using the median as the prior mean was closer to truth 5 times, the one using the midpoint estimator closer 6 times, and there was one tie. The order in which the usage and bounds questions were asked did not seem to make any difference in the accuracy of estimation.

Meanwhile, other progress was being made. Schwartz and Paulin (2000) did a study at the Bureau of Labor Statistics comparing several techniques using bounds/interval questions. They found that Rs liked the RGI technique because they felt it gave them some control over their disclosures of income. They also found that the intervals offered by Rs tended to be smaller than those generated by the investigators themselves in another condition of the experiment. And intervals generated by the Rs had been used in several other contexts. Earlier rounds of the Survey of Consumer Finances used interval estimates to elicit answers from reluctant Rs (Kennickell, 1997) and the 2004 round was planning to put more emphasis on letting Rs who can't or won't give exact amounts determine their own ranges--rather than falling back on a range card or a decision tree (Kennickell, 2004).

Further, Lusinchi (2003) had encouraged Rs on a web survey to use such intervals when they were not sure of their answers. We ourselves (Press and Tanur, 2001) showed that in the early campus experiments up to 41% of Rs who did not choose to give a point estimate of a usage quantity did give a set of bounds. If we use the midpoint of the bounds as an approximation of what the R might have answered for the usage quantity, we see that the RGI protocol has the potential to reduce item nonresponse considerably. Clearly, RGI was useful, but we needed to work on the estimation strategy and the question format.

As our thinking evolved, we went on to develop a new model that allowed a closed form solution rather than the MCMC computer intensive numerical evaluation. That new model was presented in Section II above. The new modeling develops results for a vague prior for the population mean, but results for a proper (normal) prior are analogous. We tried this model out on the data from the campus experiments described above. In order to assess the hyperparameters for a proper prior distribution we needed demographic information about Rs. (For example, we needed to know the composition of the sample in terms of year in school in order to derive a prior mean of the number of credits students would have earned. For a description of the how the prior means were derived see Press and Tanur, 2004, p. 272.)

Unfortunately, over the years some of those demographic data for the campus experiments became separated from Rs' reports on the items using the RGI questioning protocol. Hence our reanalysis of the campus experiments could use only 6 variables at SUNY-SB and only 4 at UCR. We found that the posterior mean, using a proper prior and the range of the bounds to estimate the population variance was closer to truth than the sample mean for 8 of the 10 items. Moreover, the Bayesian credibility interval covered truth for all 10 items, while the traditional confidence interval covered truth only for 6.

Clearly, the closed form estimation procedure was doing better than the MCMC procedure, but there was still room for improvement. We turned to issues of assessing the hyperparameters and to the questioning format to attempt further improvement.

We moved to expressing the hyperparameter τ according to Eqs. A21 (for the extended average estimator) and A24 (for the extended range estimator). (Earlier we had taken 4τ to be equal to $\overline{a} - \overline{b}$ for the average estimator or equal to $b_0 - a_0$ for the range estimator.)

From Eq. (A8) it is clear that sample usage quantities that are coupled with narrow intervals receive greater weight in the Bayesian estimation than do sample usage quantities that are coupled with wide intervals. Hence estimation would improve if Rs who give accurate usage quantities also gave narrow intervals and Rs who give inaccurate usage quantities gave wide intervals. We had found earlier that there is a correlation between interval length and accuracy (see Press and Tanur, 2003); we set out to improve that correlation via our questioning.

We designed a new UCR classroom survey administered to a large undergraduate statistics class in spring, 2003. We worked through Rs' confidence, having earlier found a correlation between confidence and accuracy (see Press and Tanur, 2002). In the questionnaires, we prompted confident Rs to give narrow intervals and less confident Rs to give wide intervals. Again, we asked students about everyday facts of their life on campus that we could verify – the score earned in the midterm for that class, the score on the second homework, and the registration fee (for details about this experiment, see Chu, Press, and Tanur, 2004). But before the R answered each question, s/he responded to a confidence scale we devised, and confident Rs were directed to a question that asked them to provide narrow intervals, while less confident Rs were directed to a question asking for wide intervals.

Because we varied the amount of guidance we gave the R on how wide or narrow the intervals should be, we had 3 conditions for each of the 3 items we inquired about. Thus we had 9 chances to measure the accuracy of the extended range and extended average estimators against the sample mean. Using a vague prior, we found that in 6 of these cases the extended average estimate was closest to truth (and in all these cases, the extended range estimate was in second place), in one case the extended range estimate was closest to truth, and in the remaining two cases the sample mean "won." Using a normal prior (see Chu, 2005) the results are even more encouraging. For the question about the midterm grade the extended average estimate was closest to truth in all 3 cases, and for the other 2 questions the extended range estimate was closest to truth in all 6 cases. In both this classroom experiment and in another that followed (described below), we varied the amount of guidance we gave the Rs about how wide their intervals should be if they were not confident about the accuracy of their recall. This manipulation worked – those instructed to give a wider interval gave a wider one on average than those instructed to give a less wide interval. Thus the results given above and those to be presented below use only "obedient" Rs – those who followed our guidelines on how wide their intervals should be. For details on these guidelines and results for all Rs, see Chu, Press, and Tanur (2004).

Because the sample sizes in the classroom experiment of spring, 2003 were small, we ran a similar experiment later (Nov. 5, 2003; see Chu, 2005). The questions were the same as in the spring, 2003 experiment (including the confidence scale), with the exception that instead of asking about scores on homework the student-Rs were asked for the number of movie videos they owned. (Verification data consisted of an earlier report these students had given to the professor.) We again used both a vague prior and a normal prior and the extended range and extended average estimators. Again we had 9 cases for which we could compare the estimators. Using a vague prior we found that the extended average estimate was closest to truth in 3 cases, the extended range estimate closest once, and the sample mean closest for 5 cases. With a normal prior, the results were somewhat more encouraging,: the extended average estimate, the extended range estimate, and the sample mean each was closest to truth in 3 cases.

In the November, 2003 survey, almost exactly one year before the 2004 US presidential election we also asked our student Rs an opinion question: "In your opinion, what percentage of the total vote will Mr. George W. Bush receive in the 2004 presidential election (0-100%)?"

We found that the modal response was 40% (the actual percent of the popular vote achieved by President Bush on November 2, 2004 was 51%). A graph of the Rs' bounds plotted against their usage quantities is shown in Figure 1, in which Rs have been ordered first by their usage quantity, then by their lower bounds, and then by their upper bounds to smooth the graph as much as possible.

Nevertheless, the many spikes in the graphs, and the wide variations in bounds from one R to another show that about a year before the actual presidential election of 2004, these Rs were very uncertain (fuzzy) about how strong the support for President Bush would be. It is also interesting to note from Figure 1 that as usages increase beyond about 40%, the spikiness of the graphs tends to decrease, and the lower and upper bounds tend to get closer. For the opinion data in this example, we have calculated f_i (see Eq. C1) for all Rs, and show a histogram of the distribution of the f_i's in Figure 2.

The mean fuzziness for this group of Rs on this question = 18.37; the corresponding standard deviation is 18.31 Note that these data are not available in a traditional survey of opinion without bounds information. So there is an additional "intensity of belief " (or degree of fuzziness of belief) that is being provided by an RGI survey.

This more recent classroom experiment presented an opportunity to refine our modeling. Note that the derivation in Section II assumes that the recall distribution for each R_i is normal. Of course this assumption is untestable, but evidence of possible violations of the normality assumption for the recall distributions might be reflected in a lack of normality in the sample distribution of recall quantities. Chu (2005) studied the sample distributions for each of the items in the questionnaire and applied the Wilson-Hillferty transformation (McCullagh and Nelder, 1983, see Eq. B4), to transform the distribution of these data to approximate normality. After the transformation and using a vague prior, the extended average estimator was closest to truth for 5 items, the extended range estimator closest for 2, and the sample mean closest for 2.

5. Conclusions

The RGI approach has several advantages over more conventional methods of fielding and analyzing surveys.

1)It provides a method for getting Rs to give an answer to sensitive questions which they might not otherwise answer. Rs generally feel that providing merely bounds to a question that has a numerical answer is less revealing to the interviewer than is answering a question that requires a specific answer. Hence RGI can reduce item nonresponse.

2) Many Rs feel more comfortable giving their own point estimate and range that their true value could possibly be than merely giving a point estimate, because they feel it is more accurate.

3) Rs answering using the RGI protocol are able to provide bounds for their responses as long as the bounds questions are carefully worded, and Rs are prompted with examples.

4) It is helpful to have Rs provide confidence scores for how sure they are of their answers.

5) Providing Rs with guidance in the width of intervals to use permits the analyst to focus attention on the answers of those Rs who are most confident of their responses.

6) To improve accuracy it is helpful to study a measure of the distribution of the sample data. If the data are non-normal it is likely that a transformation of the data to approximate normality followed by an RGI estimation of the transformed data will generate accurate point and interval estimates of the population parameter.

7)When the RGI protocol is used with opinion questions it can provide various measures of intensity-of-belief in the opinions of a group.

References

Chu, LiPing, (2005). Robustness of RGI Survey Estimates. Ph.D. Dissertation, Department of Statistics, University of California, Riverside.

Chu, LiPing,., Press, S. J., & Tanur, J. M. (2004). Anchoring Questions in the Respondent-Generated Intervals Protocol. *J. Modern Applied Statistical Methods,* 3, 2.

Kennickell, A. B. (1997, January). *Using range techniques with CAPI in the 1995 Survey of Consumer Finances*. Board of Governors of the Federal Reserve System, Washington, D.C. 20551.

Kennickell, A. B. (2004). Personal Communication, September 9, 2004

Lusinchi, D. (2003). Respondent-Generated Intervals: Do They Help in Collecting Quantitative Data? Paper presented at the AAPOR Annual Conference.

Marquis, K. H., & Press, S. J. (1999). Cognitive Design and Bayesian Modeling of a Census Survey of Income Recall, in *Federal Committee on Statistical Methodology Research Conference*, Wash. DC, pp. 51-64.

McCullagh, P., & Nelder, A. J. (1983). *Generalized Linear Models*, London: Chapman and Hall, Chapter 7.2, p. 152.

Press, S. J. (1997). *Bayesian Recall: A Cognitive Bayesian Modeling Approach to Surveying a Recalled Quantity*. Technical Report No. 239, Jan. 1997, Department of Statistics, University of California, Riverside, Revision of Tech. Rept. #236.

Press, S. J. (2004). Respondent-Generated Intervals for Recall in Sample Surveys. *J. Modern Applied Statistical Methods,* 3, 1, 104-116.

Press, S. J., & Marquis, K. H. (2001). Bayesian Estimation in a U.S. Census Bureau Survey of Income Recall Using Respondent-Generated Intervals. *Research in Official Statistics*, 4, 1, 151-168.

Press, S. J., & Tanur, J. M. (2000). Experimenting with Respondent-generated Intervals in Sample Surveys, with discussions, in *Survey Research at the Intersection of Statistics and Cognitive Psychology*, Working Paper Series #28, M.G. Sirken, Ed., National Center for Health Statistics, Center for Disease Control and Prevention, pp. 1-18.

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

Press, S. J., & Tanur, J. M. (2002). Decision Making of Survey Respondents. *Proceedings of the Annual Meetings of the American Statistical Association, Survey Research Methods Section, 2002*. Washington, D. C.: American Statistical Association.

Press, S. J., and Judith M. Tanur (2003). The Relationship between Accuracy and Interval Length in the Respondent-Generated Interval Protocol. Paper presented at the AAPOR Annual Conference and *Proceedings of the Survey Research Methods Section of the American Statistical Association.* Washington, D.C.: American Statistical Association.

Press, S. J., & Tanur, J. M. (2004). Relating RGI Questionnaire Design to Survey Accuracy and Response Rate. *Journal of Official Statistics*, 20, 2, 265-287. Special issue on Questionnaire Development, Evaluation, and Testing Methods.

Schwartz, L. K., & Paulin, G. D. (2000). Improving Response Rates to Income Questions. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 965-969. Washington, D.C.: American Statistical Association.

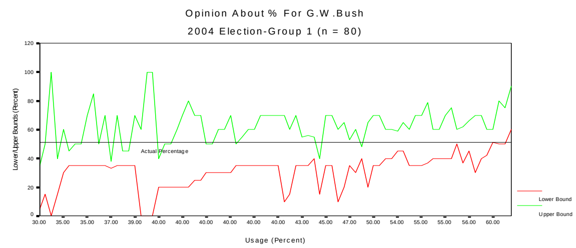

Figure 1. Opinion about Percent for Bush, 2004 Election. (Group 1, N=80)

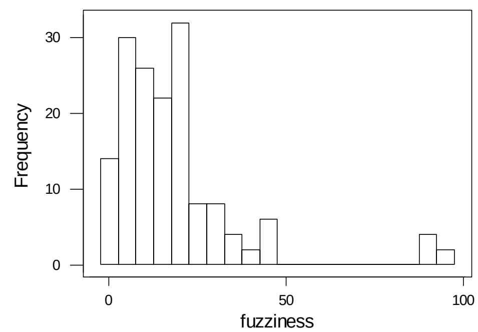

Figure 2. Histogram of Fuzziness