

**DRAFT: DO NOT QUOTE OR CITE WITHOUT PERMISSION OF THE
AUTHORS**

**Relating RGI Questionnaire Design to Survey Accuracy and
Response Rate**

by

S. James Press
University of California
Riverside

and Judith M. Tanur
State University of New York,
Stony Brook, New York

Abstract

This paper 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. The 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 involves a confluence of methodology from cognitive psychology and statistics and addresses the problem that respondents' memory lapses may lead to large non-sampling (bias) errors. The novelty of this question protocol is asking respondents for both an answer to the recall question, and also, asking them for the smallest and largest possible values they think the true answer might be. Each respondent thus generates an interval in which he/she believes his/her true value lies, as well as a basic answer. We find that a Bayesian estimator of the population mean is given by a weighted average of the basic responses, where the weight assigned to a given respondent's estimate of his/her true value depends upon ~~the~~the length of the interval the respondent gives; we assume that length is related to the respondent's confidence in his/her answer. We find 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. We show with a toy example that for a case in which half of the respondents give very short bounding intervals and very accurate usage quantities, and the

other half give relatively long intervals and inaccurate usage quantities, the bias of the RGI estimator is smaller than that of the sample mean. We summarize four experiments in which the RGI protocol has been and is being applied. These are record check surveys, so true values of the quantities respondents are recalling are available for verification.

Key Words: Accuracy, Bayesian, Brackets, Recall, Respondent-Generated Intervals, Item Non-response.

1. Introduction

We are concerned in this paper with relating questionnaire design to survey accuracy and response rate. We take up these questions with respect to a new survey questioning protocol called Respondent-Generated Intervals (RGI), see Press (1999, 2002). In this procedure we ask respondents to provide a basic answer (we call the basic answer the *usage quantity*) 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 s/he 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, 6 times. But then the respondent might also indicate that the true value is surely no greater than 7 times, but also, it is surely greater than 4 times. So the respondent has now also provided an interval of confidence (see Press and Tanur, 2002) for his/her response. It is assumed that the respondent knew the true value at some point but because of imperfect recall, he/she is not certain of the true value. We must now address the important issues of how the population parameters should be estimated, 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 usually of greatest interest in a survey is the population mean. We derive this estimator 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 paper. We discuss the characteristics of the Bayesian point estimator of

the population mean in Section 2. In Section 3 we will relate the cognitive aspects of the design of the questions to the properties of the estimator. In Section 4 we present a toy example in which the RGI estimator reduces the bias over that of the sample mean. In Sect. 5 we describe some experiments in which we have used, and are currently using, the RGI protocol, and summarize some results from them. In Section 6 we discuss some remaining outstanding questions. A detailed derivation of the RGI estimator is given in the Appendix.

2. The Bayesian Point Estimator

Let y_i, a_i, b_i denote the basic usage quantity response, the lower bound response, and the upper bound response, respectively, of respondent $i, i = 1, \dots, n$. Suppose that the y_i 's are all normally distributed. It is shown (in the Appendix), using a hierarchical Bayesian model, that in such a situation, the posterior distribution of the population mean, say θ_0 , is given by:

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

where the posterior mean, $\tilde{\theta}$, is expressible as a weighted average of the usage quantities, the y_i 's, and the weights are expressible approximately as simple algebraic functions of the bounds. ω^2 denotes the posterior variance.

The posterior mean is given by:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i, \quad (2.2)$$

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

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(\bar{b} - \bar{a})^2}{k_2^2}} \right)}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(\bar{b} - \bar{a})^2}{k_2^2}} \right)}, \quad \sum_1^n \lambda_i = 1. \quad (2.3)$$

Here, \bar{a} denotes the average of all of the a_i 's, and \bar{b} denotes the average of all of the b_i 's. 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 3 standard deviations below, and above, the mean, respectively. With this interpretation, we would take $k_1 = k_2 = 6$, 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 $k_1 = k_2 = k = 4, 5, 6, 7, 8$, and study how the estimate of the population mean varies with k . This issue relates to questionnaire design and is discussed further in Section 3.

We note the following characteristics of this estimator:

- 1) The weighted average is simple and quick to calculate, without requiring any computer-intensive sampling techniques. The weighted average point estimator may be used non-parametrically, 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 eqn. (2.3) are all the same, the weighted average reduces to the sample mean, \bar{y} , where the

weights all equal (1/n). This result suggests that the weighted average estimator and the sample mean estimates of the population mean will differ most when the respondents provide the most disparate interval lengths from one another.

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 seems intuitively to be 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 non-negative, 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) From eqn. (2.3) we see that if we take $k_1 = k_2$, 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

$$\left\{ \frac{(b_i - a_i)^2}{k_1^2} + \frac{(\bar{b} - \bar{a})^2}{k_2^2} \right\}$$

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:

$$\text{total conditional posterior precision} = \sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(\bar{b} - \bar{a})^2}{k_2^2}} \right). \quad (2.4)$$

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

3. Cognitive Aspects of the Design of 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 how to determine the effects and implications of alternative wordings is difficult. In an RGI-based survey, however, 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, Chap. 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 greater 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 greater 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 greater 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 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 pdf. 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. Most ordinary respondents would probably throw up their hands in total confusion 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. But while this alternative does provide analogous information to the analyst, it is not itself without some remaining interpretive difficulties, as explained below.

We see from eqn. (2.3) that the posterior mean estimator of the population mean is a weighted average of the usage quantities. 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, and in addition, the weights depend upon two *interpretation constants*, k_1, k_2 , which relate to how the respondents interpret the bounds questions. For simplicity, we will be taking $k_1 = k_2 = k$ throughout this paper.

As statisticians we know that as long as the data are normally distributed, for example, it is unlikely that we would find any observations beyond 3 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 3 standard deviations and conclude that the true value the respondent is being questioned about lies in an interval of length 6 standard deviations? In some of our earlier work using a paper and pencil instrument on college campuses, we phrased the bounds question, e.g. “Please fill in the blank – “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 $k = 6$. We seek to refine this process of assigning values to k .

Perhaps the ways the bounds questions are 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%.. The wording of the bounds questions can be studied to determine how to relate the interpretation of the length of the interval to the question wording; it is an empirical issue. That will be the focus of a future empirical study to provide some basis for making this critical evaluation. In the meantime, we plan to use normality when we can, to interpret the interval defined by the bounds as a 2×3 standard deviation interval, and to study how the variation in values of k around $k = 6$ affects the estimation results.

A further complication: What if the respondent's recall distribution is not normally distributed? For example, if 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 for us readily (under "Stat" → "Control Chart").

While for $k_1 = k_2 = k$, the value of a Bayesian point estimator does not depend upon k , regardless of whether $k_1 = k_2$ or not, we see from eqn. (A24) that the posterior variance of our estimator (and hence the length 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. A Toy Example

We illustrate the behavior of the RGI Bayesian estimator using a toy example. It will be seen that for this toy example, the way this estimator works is to assign greater weight to the usage quantities of respondents who give relatively short bounding intervals, and less weight to the usage quantities of those who give relatively long intervals. If the

respondents who give short intervals are also the more accurate ones, RGI will tend to give an estimate of the population mean that has smaller bias than the sample mean. Otherwise, the reverse may be true.

Suppose we have a small sample survey of size $n = 100$, in which the RGI protocol has been used. Suppose also that the true population mean of interest that we are trying to estimate is given by $\theta_0 = 1000$. Define the lower and upper bounds given by the i th respondent as (a_i, b_i) , and the corresponding interval length as: $r_i = b_i - a_i$, $i = 1, \dots, n$. Define $r_0 = \bar{b} - \bar{a}$. This quantity will be used as an assessment for τ , the common standard deviation of the respondent's means.

Assume that the first 50 respondents all have excellent memories and are quite accurate. Suppose the intervals these accurate respondents give are:

$$(a_1, b_1), \dots, (a_{50}, b_{50}) = (975, 975), \dots, (975, 975) .$$

That is, they are all not only pretty accurate, but they all believe that they are accurate, so they respond to the bounds questions with degenerate intervals whose lower and upper bounds are the same. Accordingly, these accurate respondents all report intervals of length $r_i = 0$, and usages of equal amounts, $y_i = 975$ (compared with the true value of 1000).

Next suppose that the last 50 respondents all have poor memories and are inaccurate.

They report the intervals:

$$(a_{51}, b_{51}), \dots, (a_{100}, b_{100}) = (500, 1500), \dots, (500, 1500) .$$

Accordingly, they all report interval lengths of $r_i = 1000$, and usages of equal amounts,

$y_i = 550$. We now find: $\bar{a} = 737.5$, and $\bar{b} = 1237.5$, so $r_0 = 500$.

RGI Bayesian Point Estimate of the Population Mean

We may now calculate that the weights are given by:

$$\lambda_i = \begin{cases} .0167, & i = 1, \dots, 50 \\ .0033, & i = 51, \dots, 100 \end{cases}$$

It is easy to check that: $\sum_1^{100} \lambda_i = 1$. We may now readily find that the conditional

posterior mean RGI estimator of the population mean, θ_0 , is given by:

$$\tilde{\theta} = \sum_{i=1}^{100} \lambda_i y_i = 904.875.$$

The corresponding sample mean may readily be found to be given by:

$$\bar{y} = 762.5.$$

The numerical error (bias) of the posterior mean is given by $1000 - \tilde{\theta} = 1000 - 904.875 = 95.125$.

The numerical error (bias) of the sample mean is given by $1000 - \bar{y} = 1000 - 762.5 = 237.5$. The RGI estimator has reduced the bias error by

$$237.5 - 95.125 = 142.375, \text{ or about } 60\%,$$

compared with that of the sample mean.

Interval Estimates

It is also interesting to compare interval estimates of the population mean by comparing the standard error of \bar{y} , with ω , the standard deviation of the posterior distribution of $\tilde{\theta}$. These estimates give rise to the corresponding confidence and credibility intervals for θ_0 , respectively.

We may readily find that for the data in our toy example, $\omega = 10.76$. It is also easy to check that for our data, the standard deviation of the data is 213.56. So the standard

error for a sample of size 100 is $213.56/10$, or 21.36. Thus, the RGI estimate of standard deviation is half of the standard error..

Correspondingly, the length of the 95% credibility interval $2(1.96)\omega = 42.18$, while the length of the 95% confidence interval is $2(1.96)(21.36) = 83.74$. The confidence interval is about twice as long as the credibility interval.

The 95% credibility interval is given by: (883.785, 925.965).

The 95% confidence interval is given by: (720.63, 804.37).

We note in this toy example that:

- 1) neither the RGI credibility interval nor the confidence interval covers the true value of 1000;
- 2) the intervals do not even overlap (the entire credibility interval is closer to the true value);
- 3) we expect to find many situations for which the bias error of the RGI estimator is smaller than that of the sample mean; however, the differences may be more, or less, dramatic compared with their values in this example;
- 4) in large samples, we expect that the sample mean and the RGI estimator will converge to one another, and to the true population mean.

Comment

We note that had there been only 30 accurate respondents (instead of the 50 assumed in the above example), responding in exactly the same way, and 70 inaccurate respondents (instead of the 50 assumed in the above example), the RGI estimate would still have been an improvement in bias error over that of the sample mean, although the improvement would have been smaller (31.5%).

5. Some Empirical Studies

5.a. Description of the experiments

In this section we describe four experiments we have carried out to explore the usefulness of the RGI protocol.

5 a. 1. The Two Campus Surveys

At each of our campuses, the University of California at Riverside (UCR) and the State University of New York at Stony Brook (SUNYSB), we carried out a paper-and-pencil survey. We asked students questions about the amounts of fees they paid, their SAT math and verbal scores, number of on-campus traffic tickets, number of library fines, grade point average, number of credits earned, number of grades of C or less, and 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. The form of the bounds question, as stated above, was “Please fill in the blanks – There is almost no chance that the number of credits I earned by the beginning of this

quarter [semester] was less than _____, and almost no chance that it was more than _____.” Because at that time we were thinking of using the bounds information only to assess respondents’ uncertainty, we also tested another form of question – “Please fill in the blank – I would be surprised if I had earned more than _____ credits by the beginning of this quarter [semester].” For most questions a random half of the respondents was given this “surprise” form, and thus their data had to be discarded when we realized that we wanted to deal only with the “interval” form, as it did indeed provide a Respondent-Generated Interval for each respondent. For the questions about fees (2 at UCR and 2 at SUNYSB), however, we asked all respondents the question in both forms, and so we were able to retain data on all respondents. These fees were uniform for all full time students, and thus both because they were asked in the interval form for all respondents and because no individual verification was needed for the data, we have much bigger sample sizes to work with for these questions than for the other ones on the student surveys.

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 a few students were not completely cooperative in terms of giving serious or truthful answers, and only about half the students were willing to let us to 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.

5.a.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 procedure for sensitive questions, such as a respondent’s income.

This experiment used extensive cognitive pretesting for the form of the interval question. Further, 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.

A frame of households was developed from the 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 4 states in which the American Community Survey (ACS) held its first pilot tests. A sample of about 2000 households was drawn from this frame, and each household was assigned to an experimental interviewing treatment. From this sample the 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.

5.a.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 3000 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, mamogram, 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 are 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).

5. b. Some Results

5.b.1. Accuracy of Estimation

While we have done considerable work in our record check surveys to gauge the accuracy of the point and interval estimates derived from the RGI protocol, all those analyses were based on an earlier model that we now consider much less satisfactory than the one described above, which has only recently been derived. There has been no time to update the findings on accuracy using the new estimation model, and so we do not report any

model-based results on accuracy here. We do touch on the accuracy of some ad hoc estimators when respondents offered bounds but no usage quantity.

5. b. 2. Wording Variations Already Explored

We have some empirical evidence that the form of the question affects the respondents' intervals.

In our paper and pencil campus experiments we phrased the additional bounding question (surprise version) in a split ballot design e.g. "Please fill in the blank – "I would be surprised if I had earned more than _____ credits by the beginning of this quarter". As noted above, we abandoned this approach primarily because we realized that asking for two bounds rather than just one offered us the advantages of a respondent-generated interval without the necessity of making any symmetry assumptions. But we also abandoned this approach when we found that respondents were being very inconsistent in their interpretation of what "surprise" meant, so that we were not able to assign a constant k across respondents.

In the cognitive testing used to prepare for a telephone interview to gather data for the Census experiment, interviews with paid respondents were videotaped. These videotapes included the parts of the interview involving asking the respondents to talk aloud about any problems they had in trying to answer the questions. This process was of great help to the designers of the question wording.

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 was 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 face-to-face mock Consumer Expenditure Quarterly Survey interviews that compared RGI with unfolding brackets and conventional instrument-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.”

In addition, 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.

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.

5.b.3 Reduction of Item Non-response

To investigate whether the RGI procedure reduces item non-response we used data from the paper-and-pencil campus experiments. Those *Rs* 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 non-responders 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, ARG I) 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 ARG I cover the true value in all cases.

Thus in the Campus Experiments in a substantial proportion of cases, *Rs* who do not supply an estimate of usage quantities do supply intervals which are reasonably accurate, thus reducing the amount of item non-response appreciably. Schwartz and Paulin (2000) found that the use of an interval technique reduced item non-response 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 non-response came exclusively from those whose response to the usage quantity question was “don’t know” rather than a refusal.

In the Census Experiment, although many *Rs* did not supply usage quantities, in only a few such cases did they supply bounds information. Why these differences? 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 Rs, students distributing questionnaires, and an “academic” survey. The Census Experiment interviewed Rs from established households, who were presented with questions from professional interviewers representing the US Census Bureau. Overall, there was greater respondent cooperation in this government survey by telephone than we found in our earlier campus-based experiments.

6. Some Remaining Questions

Why do respondents using the RGI protocol prefer to answer one type of question format instead of another? Which type of questioning format do they feel will yield the most accurate results? Which type of question format will result in the greatest reduction in non-response? Some of these questions are addressed in Schwartz and Paulin (2000). Which type of question format will result in the greatest reduction in response error bias? Will RGI result in greater item response rate for both sensitive and less sensitive questions when asked via the same mode of questioning?

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. How will this affect the results? Respondent burden would certainly be reduced. A reasonable estimator of the population mean in such a case 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 could be expressed as

$$\tilde{\theta} = t\bar{a} + (1-t)\bar{b}, \quad 0 \leq t \leq 1.$$

But how should the t -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. (The same result is obtained by choosing the midpoints of all ranges given, and averaging these midpoints.) Another choice would be to select t to depend on the saliences of the question 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 t -weights could be taken to be:

$$t = \frac{\frac{1}{\hat{\sigma}_a^2}}{\frac{1}{\hat{\sigma}_a^2} + \frac{1}{\hat{\sigma}_b^2}}, \quad (1-t) = \frac{\frac{1}{\hat{\sigma}_b^2}}{\frac{1}{\hat{\sigma}_a^2} + \frac{1}{\hat{\sigma}_b^2}}.$$

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

Should RGI questions be ordered so that the bounds questions are asked first, followed by the request for the usage quantity, or should the questions be posed in reverse order? Does it make a difference in accuracy and/or item response rate? Tversky and Kahneman (1974) comment on the heuristics used by respondents to surveys. If the usage quantity question is asked first, the “anchoring heuristic” may be used by the respondent to

determine the bounds by adding some small percentage in either direction to the usage quantity. If the bounds question is asked first, as it was for 75% of the respondents in the Census Experiment, we have found that there is a relationship between respondents' confidence in the accuracy of their recall and their placement of the usage quantity within those bounds. Those respondents who were most confident tend to place their usage quantity towards the middle of the interval; those least confident tend to place it towards the extremes of the interval (Press and Tanur, 2002). Which ordering offers more accuracy will be explored with the reanalysis of the data from the Census experiment.

We are also eager to see whether respondents are more likely to choose whether to answer the bounds questions or the usage question when given the choice, as they are in two experimental groups in the HMO experiment. Will the order of presentation of the choices matter? And what are the effects on accuracy?

References

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

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

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

Marquis, Kent H., and Press, S. James (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, 1999, pp.51-64 (see <http://bts.gov/fcsm>).

Miller, Diane and Press, S. James (2002). "An Experiment Embedded in a Health Survey With Respondent-Generated Intervals", paper presented at the Annual Meetings of the American Statistical Association., Aug., 2002, and to appear in the *Proceedings of the Survey Research Methods Section* of that conference.

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

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

Press, S. James, and Marquis, Kent 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. James, and Marquis, Kent 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, 2000, Crete, Greece; Amsterdam: Eurostat.

Press, S. James, and Tanur, Judith 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.

----- and -----.(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.

Ramsey, Frank Plumpton (1931). "Truth and Probability", in *The Foundations of Mathematics and Other Logical Essays*, Edited by R. B. Braithwaite, London and New York, p.71. Reprinted in *Studies in Subjective Probability*, edited by Henry E. Kyburg, Jr. and Howard E. Smokler, Huntington, New York: Robert E. Krieger Pub. Co., 25-52, 1980.

Schwartz, Lisa K. and Paulin, Geoffrey D. (2000) "Improving Response Rates to Income Questions", *Proceedings of the American Statistical Association Section on Survey Research Methods*, pp. 965-969.

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

Appendix

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

Suppose respondent i gives a point response y_i , and bounds (a_i, b_i) , $a_i \leq b_i$, $i = 1, \dots, n$, as his/her answers to a factual recall question. Assume:

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

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

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

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

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

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

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

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

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

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

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

where:

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

Expand (A6) in terms of the θ_i 's by completing the square. This takes some algebra. We find:

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

where:

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

Now find the marginal density of y by integrating (A5) with respect to $\underline{\theta}$. We find:

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

where:

$$J(\theta_0) \equiv \int \exp \left\{ \left(-\frac{1}{2} \right) \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2 \right\} d\underline{\theta}, \quad \delta_i = \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2} \right). \quad (\text{A9})$$

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

$$f \equiv \begin{pmatrix} \beta_i \\ \alpha_i \end{pmatrix}, \quad K^{-1} \equiv \text{diag}(\alpha_1, \dots, \alpha_n),$$

$$(\underline{\theta} - f)' K^{-1} (\underline{\theta} - f) = \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2. \quad (\text{A10})$$

Carrying out the (normal) integration gives:

$$p(\underline{y}|\underline{\theta}_0, \tau^2, \underline{\sigma}^2) \propto \frac{1}{|K^{-1}|^{1/2}} \exp\left\{\left[-\frac{1}{2} \sum_1^n \alpha_i \delta_i\right]\right\}. \quad (\text{A11})$$

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

$$p(\underline{y}|\underline{\theta}_0, \tau^2, \underline{\sigma}^2) \propto \exp\left\{\left[-\frac{1}{2} \sum_1^n \alpha_i \delta_i\right]\right\}. \quad (\text{A12})$$

Now apply Bayes' theorem to θ_0 in (A12).

$$p(\theta_0|\underline{y}, \tau^2, \underline{\sigma}^2) \propto p(\theta_0) \exp\left\{\left[-\frac{1}{2} \sum_1^n \alpha_i \delta_i\right]\right\}, \quad (\text{A13})$$

where $p(\theta_0)$ denotes a prior density for θ_0 . Our prior belief (prior to observing the point and bound 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. 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 positive real line. We therefore adopt a prior density of the form:

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

for all θ_0 on the positive half 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 eqn. (A14).) Inserting (A14) into (A13), and noting that $p(\theta_0) \propto \text{constant}$, gives:

$$p(\theta_0|\underline{y}, \tau^2, \underline{\sigma}^2) \propto \exp\left\{\left[-\frac{1}{2} \sum_1^n \alpha_i \delta_i\right]\right\}. \quad (\text{A15})$$

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

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

where:

$$u = \sum_1^n \left(\frac{1}{\tau^2} - \frac{1}{\alpha_i \tau^4} \right) \quad v = \sum_1^n \left(\frac{y_i}{\alpha_i \sigma_i^2 \tau^2} \right) \quad (\text{A17})$$

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

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

$$\text{where: } \tilde{\theta} \equiv \frac{v}{u}, \text{ and } \omega^2 \equiv \frac{1}{u}. \quad (\text{A 19})$$

Conditional Posterior Mean Of θ_0 As A Convex Mixture Of Usages

The appropriate measure of location of the posterior distribution in eqn. (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:

$$\lambda_i \equiv \frac{\left(\frac{1}{\sigma_i^2 + \tau^2} \right)}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}, \quad \sum_1^n \lambda_i = 1, \quad (\text{A20})$$

then:

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

Thus, the mean of the conditional posterior density of the population mean 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 $(\sigma_i^2 + \tau^2)^{-1}$ as the precision attributable to respondent i's response, and

$\sum_1^n (\sigma_i^2 + \tau^2)^{-1}$ as the total precision attributable to all respondents, λ_i is interpretable as the proportion of total precision attributable to respondent i . Thus, the greater his/her precision proportion, the greater the weight that is automatically assigned to respondent i 's usage response.

Assessing the Variance Parameters

Take:

a) $k_1\sigma_i = (b_i - a_i)$, for all $i = 1, \dots, n$; for some k_1 , such as $k_1 = 4, 5, 6, 7, 8$. Typically, we would take $k = 6$ (3 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, take

c) $k_2\tau = \bar{b} - \bar{a}$ for some pre-assigned k_2 . τ is the same for all respondents. We use an interval of 3 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 average respondent interval.

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

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}, \quad \sum_1^n \lambda_i = 1, \quad (\text{A21})$$

where: $r_0 \equiv \bar{b} - \bar{a}$. Note that in the special case that $k_1 = k_2$, the k 's cancel out in numerator and denominator, so that the weights do not depend upon the k 's. Then, the weights become:

$$\lambda_i \doteq \frac{\left(\frac{1}{(b_i - a_i)^2 + r_0^2} \right)}{\sum_1^n \left(\frac{1}{(b_i - a_i)^2 + r_0^2} \right)}. \quad (\text{A22})$$

Conditional Posterior Variance Of θ_0

It is straightforward to check that the conditional posterior variance of θ_0 is given by:

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)} \doteq \frac{1}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}, \quad (\text{A23})$$

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

$$\omega^2 \doteq \frac{1}{\sum_1^n \left(\frac{k^2}{(b_i - a_i)^2 + r_0^2} \right)}, \quad (\text{A24})$$

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 is given by:

$$(\theta_0 | y, \tau^2, \sigma^2) \sim N(\tilde{\theta}, \omega^2), \quad (\text{A25})$$

where $\tilde{\theta}$ and ω^2 are given, respectively, in (A19), (A20), and (A23) or (A24).

Credibility Intervals

Let z_γ denote the $\gamma/2$ -percentile of the standard normal distribution. Then, from (A25), a $(100-\gamma)\%$ credibility interval for the population mean, θ_0 is given by:

$$(\tilde{\theta} - z_\gamma \omega, \tilde{\theta} + z_\gamma \omega). \quad (\text{A26})$$

That is,

$$P\{\tilde{\theta} - z_\gamma \omega \leq \theta_0 \leq \tilde{\theta} + z_\gamma \omega | y, \tau^2, \sigma^2\} = (100 - \gamma)\%. \quad (\text{A27})$$