Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 14

11-1-2004

Confidence Elicitation And Anchoring In The Respondent-Generated Intervals (RGI) Protocol

LiPing Chu
University of California, Riverside, Liping13@netzero.com

S. James Press
University of California, Riverside, jpress@ucr.edu

Judith M. Tanur
SUNY, Stony Brook, jtanur@notes.cc.sunysb.edu

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 417-431

1538 – 9472/04/$95.00

Confidence Elicitation And Anchoring In The Respondent-Generated Intervals (RGI) Protocol

LiPing Chu
Department of Statistics
University of California, Riverside

S. James Press
Department of Statistics
University of California, Riverside

Judith M. Tanur
Department of Sociology
SUNY, Stony Brook

The Respondent-Generated Intervals protocol (RGI) has been used to have respondents recall the answer to a factual question by giving not only a point estimate but also bounds within which they feel it is almost certain that the true value of the quantity being reported upon falls. The RGI protocol is elaborated in this article with the goal of improving the accuracy of the estimators by introducing cueing mechanisms to direct confident (and thus presumably accurate) respondents to give shorter intervals and less confident (and thus presumably less accurate) respondents to give longer ones.

Key words: RGI, surveys, Bayes, hierarchical models, record-checks, anchoring

Introduction

The Respondent-Generated Intervals protocol (RGI) has been used to have respondents recall the answer to a factual question by giving not only a point estimate but also bounds within which they feel it is almost certain that the true value of the quantity being reported upon falls (Press, 2004). This paper reports on new thinking that aims to elaborate the RGI protocol with the goal of improving the accuracy of the estimators derived from the protocol.

There are two aspects to the new thinking. The first is a new analytical Bayesian procedure for estimating the population mean in an RGI survey; it is derived in the Appendix.

LiPing Chu is a doctoral candidate in applied statistics at the University of California, Riverside. Her dissertation centers on the robustness of Respondent Generated Intervals. Email: Liping13@netzero.com. Dr. S. James Press is a Distinguished Professor of Statistics, at the University of California, Riverside. Email: jpress@ucr.edu. Dr. Judith M. Tanur is a Distinguished Teaching Professor at the State University of New York, Stony Brook. Email: jtanur@notes.cc.sunysb.edu.

The second is a new type of anchoring questioning technique that cues and encourages confident (and presumably accurate) respondents to give short intervals and less confident (and presumably less accurate) respondents to give long intervals. The new analytical procedure is summarized briefly in the next section (and elaborated in the Appendix), followed by a section containing a discussion of a classroom survey experiment and how it incorporates the new questioning technique. The final section provides a discussion of the implications of these innovations.

Vague Prior Bayesian Point Estimator for the Population Mean

For a sample of n independent respondents in a survey, let y_i, a_i, b_i denote the basic usage quantity response, the lower bound response for where the true value to the question lies for that respondent, and the upper bound response for where the true value to the question lies for that respondent, respectively, of respondent i, $i = 1,\ldots,n$. Suppose that the y_i's are all independent and normally distributed. Suppose also that we adopt a vague prior distribution for the population mean, θ_0, to represent knowing little, a priori, about the value of the population mean. It is shown in Press (2004) using a hierarchical Bayesian model, that

in such a situation, the posterior distribution of θ_0 is given by:

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

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

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

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

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

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

The new estimating procedure used here substitutes for $(b_0 - a_0)$:

$$4\tau \doteq (b_0 - a_0) + \frac{2}{\sqrt{n}}(s_a + s_b)$$

to form what will be called the extended range estimator, and

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

to form what will be called the extended average estimator (see Appendix). Here $\bar{b}$ and $\bar{a}$ are the means of the upper bounds and of the lower bounds given by the respondents, respectively; and s_a and s_b are the sample standard deviations of the lower bounds and upper bounds, respectively.

Methodology

The Classroom Survey: Confidence and Question Wording

Because point estimates of respondents who give short intervals are weighted more heavily in the Bayesian RGI estimator than are point estimates of respondents who give longer intervals (see 2.3), it is advantageous to encourage respondents who are more accurate to give shorter intervals and respondents who are less accurate to give longer ones. It is known from earlier uses of the RGI procedure that, among respondents who do not receive any special guidance about the length of their intervals, there is a substantial correlation between interval length and accuracy (with less accurate respondents giving longer intervals; Press & Tanur, 2003). There is also a

correlation between confidence and interval length (with less confident respondents giving longer intervals; Press & Tanur, 2002). The aim is to increase the correlation between accuracy and interval length, by working through respondents' confidence and cueing them appropriately. We have developed a questioning protocol that aims to increase that correlation.

First, the respondent is requested to give his/her best guess about the quantity being investigated, and then is asked how confident s/he is of that answer on a scale from 0 (least confident) to 10 (most confident). Figure 1 shows the form of this confidence scale for a question used in our experiment involving recall of the respondent's grade on a classroom exam. Respondents who represent themselves as highly confident (confidence ratings 7.5 or 10) are directed to a question that encourages them to give a narrow bounding interval. Less confident respondents (confidence ratings of 5 or less) are directed to a question encouraging a wide bounding interval.

The design for this experimental application of the new protocol used three versions of the bounding questions (and each version was completed by a different group of respondents). Version 1, referred to as unanchored, simply asks the respondent to give a narrow, or a broad, interval; this version was administered to Group 1. See Figure 2 for the wording of Version 1 for the question about the classroom exam. Version 2, administered to Group 2, which is referred to as the narrow-wide anchored condition, not only encourages respondents to give narrow or wide intervals, but it also tells them that the narrow interval should be no more than a specified width and that the wide interval should be at least a specified width. See Figure 3 for the wording of Version 2 as used for the question about the classroom exam. Version 3 (referred to as the wide-wide anchored condition and administered to Group 3) is the same as Version 2, except that the suggested width of the wide interval was considerably wider (see Figure 4).

Figure 1. New Form for RGI Protocol.

1) What is your best guess as to what your score was on your first exam in this class? (Please don't answer if you've missed the first exam)._____.

2) How confident are you about your answer to Question 1? Please answer on the following confidence scale. (Place a check in the first column next to the answer you prefer.)

Confidence Scale

Place a check somewhere in this column	Numerical Score	Interpretation of confidence rating	Which question should I answer next?
	0	I have absolutely no idea what my exam score was	Go to Question 3b
	2.5	I am uncertain what my exam score was	Go to Question 3b
	5.0	I might be right and I might be wrong about what my exam score was	Go to Question 3b
	7.5	I think that I know what my exam score was	Go to Question 3a
	10.0	I am absolutely certain what my exam score was	Go to Question 3a

Figure 2: Unanchored Bounds Condition.

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. Please fill in:

The smallest my exam score could have been is %_____,

The largest my exam score could have been is %_____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score
Please fill in:
The smallest my exam score could have been is %_____,

The largest my exam score could have been is %_____.

Figure 3: Narrow Wide Anchor-Type Bounds Question.

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. For example, if your best guess about your exam score is 75%, give a narrow interval of no more than 4 points in length, such as (73%, 76%).
Please fill in:

The smallest my exam score could have been is %_____,

The largest my exam score could have been is %_____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score. For example, if your best guess about your exam score is 75%, give a wide interval of at least 20 points in length, such as (65%, 85%).
Please fill in:
The smallest my exam score could have been is_%_____,

The largest my exam score could have been is_%_____.

Figure 4: Wide Wide Anchor-Type Bounds Question**.**

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. For example, if your best guess about your exam score is 75%, give a narrow interval of no more than 4 points in length, such as (73%,76%).

Please fill in:

The smallest my exam score could have been is_____,

The largest my exam score could have been is_____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score. For example, if your best guess about your exam score is 75%, give a wide interval of at least 30 points in length, such as (60%, 90%).

Please fill in:
The smallest my exam score could have been is_____,

The largest my exam score could have been is_____.

Figure 5: Memory Evaluation Scale.

Does it ever happen that when you are sure you know something, it turns out that you are mistaken?
Please check one:

Never_____

} **Good Memory**

Seldom_____

Sometimes_____

} **Poor Memory**

Frequently_____

Ratings of Memory

Respondents were asked to evaluate their memory on the scale shown in Figure 5 (The designations "Good Memory" and "Bad Memory" as shown in Figure 5 did not appear in the questionnaire given to the respondents). If respondents are good judges of their own memory, then perhaps rather than asking confidence questions for each survey item we can use a procedure that simply classifies respondents into good memory and poor memory groups and encourage good memory respondents to give short intervals and poor memory respondents to give long ones. Such a procedure would impose considerably less respondent burden than does asking for confidence for each question.

The Survey

In the spring of 2003 we ran a small experimental record-check survey in an undergraduate, lower division, statistics class at the University of California at Riverside. In a randomized design three groups of students were each given a different version of the questionnaire and the students were asked to recall their midterm exam score, their score on their second homework assignment, and the amount they had paid at the beginning of the quarter as a registration fee. Because there were three versions of the questionnaire, and because participation was voluntary, sample sizes in the three groups were rather small, but sufficiently large for us to derive some preliminary results. (A similar experiment from a larger class was run several months later in the fall of 2003 – results will be available shortly.) With the students' permission we were able to compare their reported grades with those recorded in the professor's grade book; the registration fee was fixed by the university for all full-time students at $239.

Results

The first finding was that the manipulation worked. Table 1 shows that the mean length of intervals generated by respondents who were asked to give a wide interval were always wider than those from respondents asked to give a narrow interval. In every case in which a t-test was possible (that is, whenever both group sizes were greater than 1) this finding reached at least marginal statistical significance, in spite of the small sample sizes.

For both the homework question and the midterm question, the mean of the wide intervals for respondents given the wide-wide anchor was longer than the mean of the intervals for respondents given the narrow-wide anchor. This relationship did not hold for the question about registration fee, for which most respondents seem to have been very much lacking in knowledge about how much the actual fee was (which resulted in low confidence).

It is interesting to note that there seems to be a relationship between respondents' confidence and the salience of the question. A large majority of respondents were quite confident that they remembered their midterm grade correctly, a large majority lacked such confidence for the registration fee, and for the homework grade the respondents split about half and half.

Table 2 further checks the manipulation, asking whether there was indeed a correlation between respondents' confidence in the accuracy of their recall and their actual accuracy in reporting their usage quantities. The actual accuracy is measured as the absolute value of the differences between the reported usage quantity and recorded truth. Large values of these differences represent *inaccuracy*. If there is a relationship between accuracy and confidence, negative correlations would be expected, as indeed are indicated in Table 2. (We might have labeled the absolute value of the difference between truth and the usage quantity as inaccuracy, but calling it accuracy simplifies our discussion as long as the reader keeps in mind how the variable is measured and that we hope for negative correlations between it and interval length.)

Table 1: Manipulation Check.

Average Lengths of Intervals for Wide and Narrow Anchors.

		Narrow Interval	n	Wide Interval	n	p value** (1-tailed)
Midterm						
	Unanchored	6.7	19	14.6	8	0.006
	Narrow/Wide Anchor	9.0	18	**16.7**	3	0.069
	Wide/WideAnchor	8.5	19	**25.0**	3	0.026
RegFee						
	Unanchored	165.00	5	1280.40	19	0.004
	Narrow/Wide Anchor	0.00	1	763.90	13	*
	Wide/WideAnchor	20.00	1	608.33	18	*
Homework						
	Unanchored	2.9	12	6.6	10	0.033
	Narrow/Wide Anchor	2.8	10	**6.4**	10	0.030
	Wide/WideAnchor	2.7	9	**7.6**	8	0.020

*Narrow interval group n=1; no test of significance possible.
**p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers.

Table 2: Correlations between "*r*" Confidence and "Accuracy" (|usage-truth|).*

All respondents

		r	n	p value** (1-tailed)
Midterm				
	Unanchored	-0.110	27	0.305
	Narrow/Wide Anchor	-0.518	21	**0.008**
	Wide/WideAnchor	-0.443	23	**0.017**
RegFee				
	Unanchored	-0.111	27	0.291
	Narrow/Wide Anchor	-0.049	14	0.434
	Wide/WideAnchor	-0.375	21	**0.047**
Homework				
	Unanchored	-0.385	20	**0.047**
	Narrow/Wide Anchor	-0.355	20	0.062
	Wide/WideAnchor	-0.184	17	0.289

*We expect high levels of confidence to go with greater accuracy (small error, the absolute difference between the usage quantity offered by the respondent and truth); thus increasing confidence should go with decreasing error, Hence, if our hypothesis is correct, the correlations should be negative. They are.
**p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. p-values in bold-face are significant at the 5% level or smaller.

Although these correlations are hardly enormous, there is a relationship between accuracy and confidence in all cases. Each group of respondents contributed at least one low correlation – the unanchored group showing a low correlation for both the midterm question and the registration fee question, the narrow-wide anchor group showing a low correlation for the registration fee question, and the wide-wide anchor group showing a low correlation for the homework question.

Hence, the low correlations cannot be attributed either to a particular group of respondents or to the difficulty of a particular question. It is suspected, however, that the correlations coming from the registration fee question are influenced by the fact that very few respondents were confident about their answers to this question – see the *n*'s in Table 1. We speculate that respondents knew more about the total fees they paid than about the specific registration fee, about which they knew almost nothing, so they guessed wildly. There is also some evidence from student comments that if their parents paid their fees or if they received financial aid, they have little knowledge about the amount of any fees.

Table 3 examines the relationship between interval length and accuracy (measured as explained above, that is, as inaccuracy). If, as hoped, respondents who are less accurate give longer intervals, positive correlations would be expected. The correlations in Table 3 are all positive. There are two panels for Table 3 – the top panel includes all respondents who gave the 4 pieces of data requested – confidence rating, usage quantity, lower bound, and upper bound – and whose usage quantity properly fell within the bounds.

The bottom panel includes only what is called obedient respondents – those who followed the directions given in the anchoring instructions and gave a wide interval at least as wide as prescribed, or a narrow interval at least as narrow as prescribed. Two comments are in order for this table. First, it seems to have been successful in increasing the correlation between interval length and accuracy from the level obtained from respondents without any special instructions regarding interval length. Most of these correlations are larger than those reported in Press and Tanur (2002), where the median of 18 correlation coefficients (for 18 items) was 0.13; 6 of the 18 were negative; and the only correlations exceeding 0.40 were those relating to the frequencies of behaviors (a case where those who really had no occurrences of the requested behavior could easily remember that they had none, and could be quite confident about their recall).

Second, limiting ourselves to obedient respondents seems to be useful. (Note that, because the unanchored group was not given a suggested length of interval, the obedient vs. disobedient distinction does not pertain to this group and the data for this group in the lower panel of Table 3 simply repeat the data in the upper panel.) When we omit those respondents who were disobedient we find that the correlations never decrease substantially and two correlations that were originally small increase considerably.

Table 4 shows the results of the estimation process using the Bayesian estimators for the obedient respondents only. In Table 4 the estimator that is closest to the truth is presented in boldface. We see that although all the estimates were very close to one another, the extended average estimator is closest to truth for the midterm grades and for the registration fee. The sample mean seems to work best for the homework question, except for the unanchored condition where the extended range estimate is a tiny bit closer to truth.

Note that the median correlation between accuracy and interval length for the midterm question is 0.349; for the registration fee question its is 0.395; but for the homework question it is only 0.274. Hence we should not be surprised that the Bayesian estimator works better for the midterm and registration fee questions than it does for the homework question. The findings for the extended average estimator are also shown graphically in Figure 6. What is graphed is the absolute value of the difference between the RGI estimated value and average truth. G1 refers to the groups in the unanchored condition, G2 to groups in the narrow-wide anchor condition, and G3 refers to the groups in the wide-wide anchor condition.

Table 3: Manipulation Check

Correlations "r" between Interval Length and Accuracy (|usage-truth|).

All Respondents with Useable Data

		r	n	p value* (1-tailed)
Midterm				
	Unanchored	0.311	25	0.065
	Narrow/Wide Anchor	0.149	19	0.272
	Wide/WideAnchor	0.069	22	0.381
RegFee				
	Unanchored	0.395	24	**0.028**
	Narrow/Wide Anchor	0.671	12	**0.008**
	Wide/WideAnchor	0.286	19	0.128
Homework				
	Unanchored	0.011	20	0.482
	Narrow/Wide Anchor	0.273	18	0.138
	Wide/WideAnchor	0.320	17	0.105

Obedient Respondents Only

		r	n	p value* (1-tailed)
Midterm				
	Unanchored	0.311	25	0.065
	Narrow/Wide Anchor	0.624	11	**0.020**
	Wide/WideAnchor	0.349	14	0.110
RegFee				
	Unanchored	0.395	24	**0.028**
	Narrow/Wide Anchor	0.638	9	**0.032**
	Wide/WideAnchor	0.247	16	0.178
Homework				
	Unanchored	0.011	20	0.482
	Narrow/Wide Anchor	0.274	12	0.194
	Wide/WideAnchor	0.305	11	0.181

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. p-values in bold-face are significant at the 5% level or smaller.

Table 4: Point Estimate Results Using Vague Prior and Extended Average and Extended Range Procedures.

Obedient Respondents Only

		n	Average Truth	x-bar	Extended Average	Extended Range
Midterm						
	Unanchored	25	83.88	83.04	**83.79**	83.17
	Narrow/Wide Anchor	11	81.36	79.64	**79.94**	79.70
	Wide/WideAnchor	14	86.57	86.71	**86.68**	86.70
RegFee						
	Unanchored	24	$239.00	$1,366.46	**$1,202.25**	$1,328.28
	Narrow/Wide Anchor	9	$239.00	$1,090.78	**$974.77**	$1,047.50
	Wide/WideAnchor	16	$239.00	$1,190.88	**$1,122.65**	$1,176.97
Homework						
	Unanchored	20	16.91	18.00	18.06	**17.99**
	Narrow/Wide Anchor	12	16.72	**17.92**	18.04	18.01
	Wide/WideAnchor	11	16.69	**18.63**	18.81	18.65

Table 5: Average Confidence Scores by Respondents' Memory Rating.

All Respondents

		Good Memory	n	Poor Memory	n	p value* (1-tailed)
Midterm						
	Unanchored	**8.75**	6	**7.05**	22	**0.028**
	Narrow/Wide Anchor	**9.58**	6	**7.83**	15	**0.017**
	Wide/WideAnchor	8.21	7	9.22	16	
RegFee						
	Unanchored	**4.29**	7	**3.50**	20	0.257
	Narrow/Wide Anchor	**4.00**	5	**2.78**	9	0.266
	Wide/WideAnchor	**3.33**	6	**1.17**	15	0.074
Homework						
	Unanchored	**7.50**	6	**6.84**	19	0.276
	Narrow/Wide Anchor	4.58	6	6.25	14	
	Wide/WideAnchor	4.50	5	6.35	13	

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. Pairs of numbers in bold-face are consistent with out hypothesis that good-memory respondents have higher confidence than poor-memory respondents. P-values in bold-face are significant at the 5% level or better.

Figure 6: Bias (|estimate - truth|) for Extended Average Bayesian Estimate Compared with ABS Bias (Absolute Error) of Sample Mean.

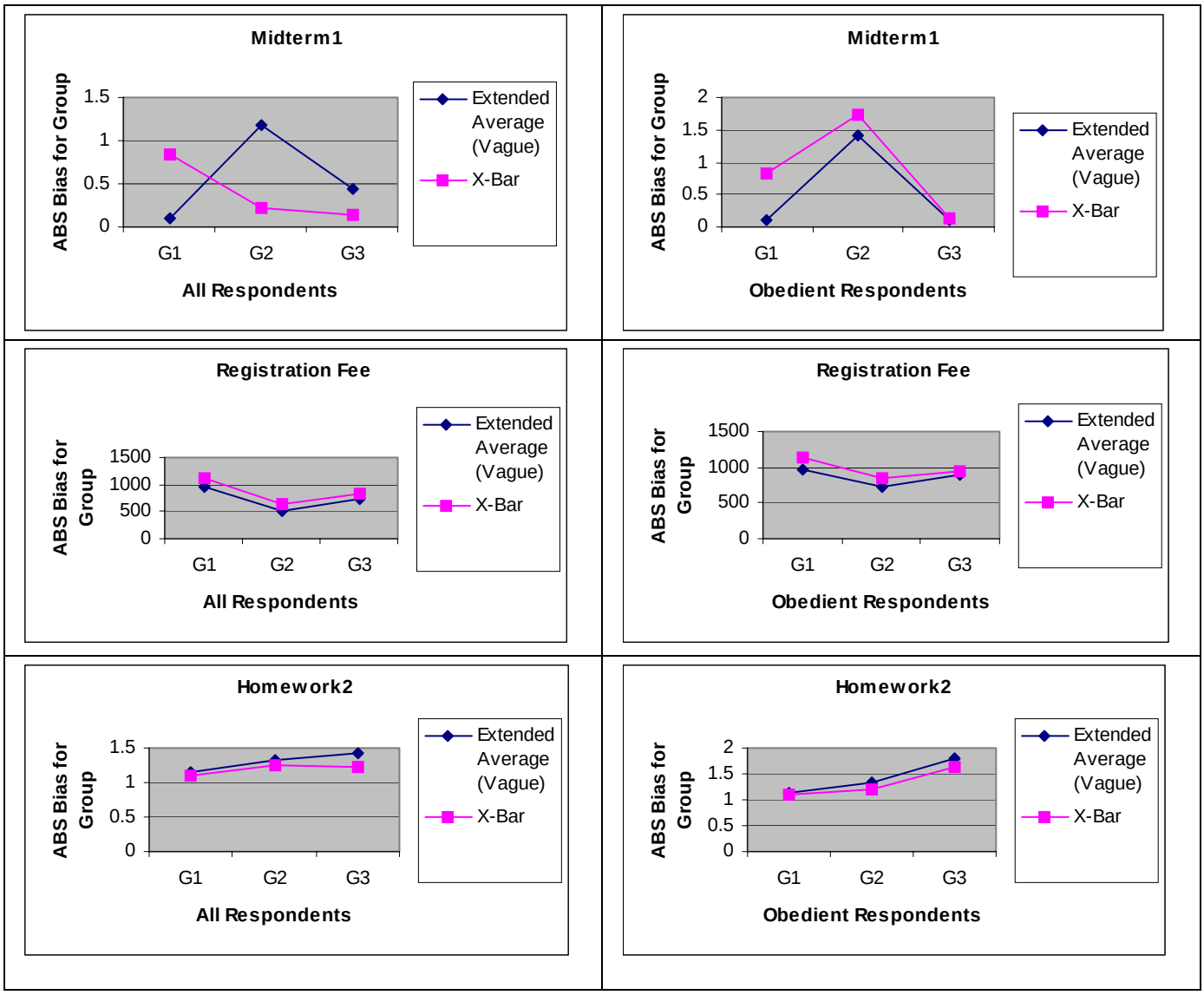

Table 6: Accuracy (|usage-truth|) by Respondents' Memory Rating.

All Respondents

		Good Memory	n	Poor Memory	n	p value* (1-tailed)
Midterm						
	Unanchored	**2.33**	6	**5.10**	21	0.134
	Narrow/Wide Anchor	7.33	6	6.00	16	
	Wide/WideAnchor	**0.57**	7	**0.69**	16	0.424
RegFee						
	Unanchored	**961**	7	**1245**	19	0.253
	Narrow/Wide Anchor	**685**	5	**1519**	10	0.112
	Wide/WideAnchor	1405	6	729	15	
Homework						
	Unanchored	**1.50**	5	**1.60**	15	0.453
	Narrow/Wide Anchor	3.00	6	2.13	15	
	Wide/WideAnchor	2.90	5	2.26	13	

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. Pairs of numbers in bold-face are consistent with out hypothesis that good-memory respondents are more accurate than poor-memory respondents.

Table 5 relates respondents' ratings of their memory to their confidence as rated on the confidence scales for the questions. Those respondents who rated their memory good (those who claimed never or seldom to be mistaken when sure they knew something) in many cases give higher average confidence ratings than respondents who say their memory is less good (those who claimed sometimes or frequently to be mistaken when they were sure they knew something). This finding holds true for all questions for the unanchored-type condition, for the midterm and the registration fee questions for the narrow-wide anchored-type condition, and only for the registration fee for the wide-wide anchored-type condition.

Confidence ratings were higher on average for the good memory group than in the poor memory group in 6 of the 9 comparisons. Two of these 6 wins reached statistical significance at conventional levels and another was marginally significant.

Table 6 shows the accuracy achieved by respondents at different levels of self-rated memory. Note that accuracy is again measured by the absolute value of the difference between a respondent's reported usage quantity and truth. Thus large values represent inaccuracy, and smaller values are more accurate. We see that on the average respondents who rated themselves to have good memories were closer to the truth in 5 of the 9 possible comparisons (shown in boldface). None of these wins reached statistical significance at conventional levels.

Conclusion

There was some success with these new directions. We seem to have affirmed the need to ask confidence questions separately for each usage quantity, for while respondents' estimates of their own memory seem to be good predictors of that confidence, those memory estimates do not relate nearly as well to actual accuracy as do the confidence ratings themselves. It was hoped to minimize respondent burden by asking a single memory question, but it seems the burden of asking separate confidence questions is a necessary one.

We have established that respondents directed to give wide intervals give wider ones on the average than do respondents directed to give narrower ones. There does not seem to be much effect of the length of the anchoring-type interval, but the results of a considerably larger sample size experiment is necessary to see if that lack of effect is real. The correlation between accuracy and interval length was improved through the use of the confidence scale. It would be useful to increase that correlation even more, as it is the *sine qua non* for the successful application of the RGI protocol.

Other methods will be used to ask for respondents' confidence, but it will be limited by any imperfections in respondents' understanding of their own accuracy. Respondents who are honestly confident but nevertheless inaccurate, and respondents who honestly lack confidence but are nevertheless accurate, will continue to haunt us. Even in this test, however, it was apparent that the manipulation of respondents' interval length, based on their confidence, results in the RGI Bayesian estimator showing less bias than the sample mean in a majority of the cases examined.

References

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

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

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

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

APPENDIX

Each of *n* respondents in a sample survey provides a triple of data: (y_i, a_i, b_i) representing respondent *i*'s usage quantity (the term "usage" was introduced originally to reflect estimated frequency of a behavior), her/his lower bound (for true value of the usage), and his/her upper bound (for true value of the usage); $i = 1,\ldots, n$. These quantities are jointly distributed. Suppose that marginally:

1) $(y_i \mid \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2);$

2) $(a_i \mid a_{i0}, \psi_{ai}^2) \sim N(a_{i0}, \psi_{ai}^2);$

3) $(b_i \mid b_{i0}, \psi_{bi}^2) \sim N(b_{i0}, \psi_{bi}^2),$

where θ_i denotes the true population value for the mean usage for respondent *i*; a_{i0}, b_{i0} denote the true population values for respondent *i*'s lower and upper bounds, respectively; and $(\sigma_i^2, \psi_{ai}^2, \psi_{bi}^2)$ denote the corresponding population variances, respectively.

Note that although (y_i, a_i, b_i) are observed quantities, $(\theta_i, \sigma_i^2, a_{i0}, b_{i0}, \psi_{ai}^2, \psi_{bi}^2)$ are unknown and unobservable. Now, assume that the θ_i's are exchangeable, and

4) $(\theta_i \mid \theta_0, \tau^2) \sim N(\theta_0, \tau^2).$

Assuming $(\sigma_1, \ldots, \sigma_n, \tau)$ are known, it has already been shown, adopting a vague prior on θ_0, gives as the posterior distribution for θ_0 (see Press, 2004):

5) $(\theta_0 \mid \underline{y}, \sigma_1, \ldots, \sigma_n, \tau) \sim N(\sum_1^n \lambda_i y_i, \omega^2),$

$$0 \le \lambda_i \le 1, \qquad \sum_1^n \lambda_i = 1.$$

The λ_i's and ω^2 are proportions of total precision. The development for a normal (rather than a vague) prior distribution on the population mean is simple and unchanged by the sequel.

Assessment of the variances

A) Assessment of the σ_i's

Take $4\sigma_i \doteq b_i - a_i, \quad i = 1, \ldots, n,$ as our assessment for the σ_i's.

B) Assessment of τ.

Assume there are approximate bounds for all subjects in the population that are approximately 2 standard deviations on either side of the mean. 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:

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

Then,

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

Next note that the true population mean value for respondent *i* must be between its bounds,

7) $a^* \le \theta_0 \le b^*.$

Case 1: Extended Average Estimator

For 95% credibility on a^*:

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

for 95% credibility on b^*:

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

From (7), (8) and (9):

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

or:

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

From (4) and 95% credibility,

11)
$$k\tau \equiv 4\tau = \left(\bar{b} + 2\frac{\psi_b}{\sqrt{n}}\right) - \left(\bar{a} - 2\frac{\psi_a}{\sqrt{n}}\right)$$
$$= \left(\bar{b} - \bar{a}\right) + \frac{2}{\sqrt{n}}\left(\psi_a + \psi_b\right).$$

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

(12)
$$s_a^2 \equiv \hat{\psi}_a^2 \equiv \frac{1}{n}\sum_1^n (a_i - \bar{a})^2;$$
$$s_b^2 \equiv \hat{\psi}_b^2 \equiv \frac{1}{n}\sum_1^n (b_i - \bar{b})^2.$$

Then, the assessment procedure for τ becomes:

13) $$4\tau \doteq \left(\bar{b} - \bar{a}\right) + \frac{2}{\sqrt{n}}\left(s_a + s_b\right).$$

There is a Minitab 13 macro for computing the Bayesian RGI extended average estimator (see Remark c).

Case 2: Extended Range Estimator

From (10), since $a_0 < \bar{a}$, and $\bar{b} < b_0$, consider for an alternative assessment procedure,

10*) $$a_0 - 2\frac{\psi_a}{\sqrt{n}} \le \theta_0 \le b_0 + 2\frac{\psi_b}{\sqrt{n}}$$

Then, (11) becomes:

(11*)
$$k\tau \equiv 4\tau = \left(b_0 + 2\frac{\psi_b}{\sqrt{n}}\right) - \left(a_0 - 2\frac{\psi_a}{\sqrt{n}}\right)$$
$$= \left(b_0 - a_0\right) + \frac{2}{\sqrt{n}}\left(\psi_a + \psi_b\right).$$

Using (12) gives:

12*) $$4\tau \doteq \left(b_0 - a_0\right) + \frac{2}{\sqrt{n}}\left(s_a + s_b\right).$$

Remarks:

a) Note that these assessments give larger values of τ than our earlier assessments, $\left(\bar{b} - \bar{a}\right)$, and $(b_0 - a_0)$ the assessments called average, and range, The credibility intervals for the population mean will accordingly be larger.

b) The second term in (13), and in (12*) disappear for large sample sizes, leaving just the average or range of the bounds, but for smaller sample sizes, the 2nd term can have a substantial effect.

c) Minitab 13 macros for computing the Bayesian RGI extended average and extended range estimators are available (for information about these macros, contact Diane Miller at diane.m.miller@nge.com).